POLITECNICO DI MILANO

DEPARTMENT OF MATHEMATICS "F. BRIOSCHI"

PHD COURSE IN

MATHEMATICAL MODELS AND METHODS IN ENGINEERING

XXIV CYCLE

STATISTICAL METHODS FOR CLASSIFICATION IN CARDIOVASCULAR HEALTHCARE

Supervisor: Prof. Anna Maria Paganoni
Tutor: Prof. Anna Maria Paganoni

Ph.D. Coordinator: Prof. Paolo Biscari

Candidate: Francesca Ieva, matr. 738628

ACADEMIC YEAR 2011-2012

To my Mom

Acknowledgments

This work is within the Strategic Program “Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction” and PROMETEO (PROgetto sull’area Milanese Elettrocardiogrammi Teletrasferiti dall’Extra Ospedaliero).

The author wishes to thank: **Regione Lombardia** - Healthcare division, **AREU** - Azienda Regionale Emergenza Urgenza, **118 Dispatch Centre** of Milan, **Lombardia Informatica S.p.A.**, **Mortara Inc.** and **Mortara Rangoni Europe**, the Cardiological Societies **ANMCO** - Associazione Nazionale Medici Cardiologi Ospedalieri, **SIC** - Società Italiana di Cardiologia and **GISE** - Società Italiana di Cardiologia Invasiva, as well as all the physicians engaged in the **Working Group for Cardiac Emergency**, and last but not least **Politecnico di Milano** - Department of Mathematics.

In particular, within these institutions, a special acknowledgment is due to: Dr. **Maurizio Bersani** (Regione Lombardia - Healthcare Division), Dr. **Maurizio Marzegalli** (A.O. S.Carlo Borromeo), Dr. **Niccolò Grieco** (AAT 118 Milano - A.O. Niguarda Ca’ Granda), Dr. **Giovanni Sesana** (AAT 118 Milano - A.O. Niguarda Ca’ Granda), Dr. **Pietro Barbieri** (A.O. Melegnano), Dr.ssa **Elena Corrada** (Istituto Clinico Humanitas), Dr.ssa **Cecilia Del Vecchio** (A.O. S.Carlo Borromeo), Dr. **Luigi Oltrona Visconti** (Fondazione IRCCS Policlinico San Matteo di Pavia), Ing. **Fabrizio Pizzo** (Lombardia Informatica S.p.A.), Ing. **Morena Valzano** (Lombardia Informatica S.p.A.), Dr.ssa **Gabriella Borghi** (Cefriel), Ing. **Maurizio Fumagalli** (Mortara Instrument Inc.), Ing. **Johan DeBie** (Mortara Rangoni Europe).

Contents

Introduction **12**

I Regione Lombardia cardiovascular healthcare system: data sources, hospital network and Strategic Program **16**

1 A patient-focused documentation for a new definition of epidemiology **18**
1.1 Outcomes and performance measurements of public healthcare systems 18
1.1.1 Performance measurement in public health 21
1.1.2 Why do performance measurements? 23
1.2 Electronic Health Record (Fascicolo Sanitario Elettronico) 24
1.3 Focus on Acute Coronary Syndromes 27

2 Pre-hospital organization and systems of care **31**
2.1 The idea of the *Cardiological Network* 31
2.2 The Strategic Program of Regione Lombardia 34
2.2.1 P1 - Part of Strategic Program of Regione Lombardia 35
2.2.2 Cardiovascular process indicators: time to treatment and time to intervenction. 38

3 Data sources **42**
3.1 The administrative datawarehouse of Regione Lombardia 42
3.1.1 The star scheme: an overview of complexity 42
3.1.2 Data mining of administrative databanks 43
3.2 Clinical surveys 44
3.2.1 Past cardiological surveys of Regione Lombardia 44
3.2.2 The MOMI^2 experience on Milan area 45
3.2.3 The STEMI Archive 47
3.2.4 The PROMETEO database 49
3.3 More complex data 53
3.3.1 Data Mart of Regione Lombardia datawarehouse 54
3.3.2 Integrated system: examples of complex longitudinal data 55
3.4 Data mining of integrated databases 58

II Statistical models and methods for healthcare data **60**

4 Statistical models for healthcare: the frequentist approach **62**
4.1 Motivations 62

4.2 Linear parametric mixed effects models . . . 62
4.2.1 Single and multi level of grouping . . . 63
4.2.2 Estimation in LME models . . . 64
4.2.3 Classification and inference using LME models . . . 69
4.3 Generalized linear parametric mixed effects models . . . 70
4.3.1 Model formulation for GLME models . . . 70
4.3.2 Estimation for GLME models . . . 71
4.3.3 Inference for GLME models . . . 75
4.4 Linear nonparametric mixed effects models . . . 76
4.4.1 From parametric to nonparametric GLME models . . . 77
4.5 Nonlinear parametric mixed effects models . . . 78
4.5.1 Theory and computational methods of NLME models . . . 79
4.6 Nonlinear nonparametric mixed effect models . . . 81
4.6.1 NLNPEM: unsupervised classification in nonlinear nonparametric framework using random effects . . . 82
4.7 Problems due to unbalanced share . . . 85
4.7.1 Prediction probabilities in unequal sample size . . . 86
4.7.2 Undesirable effects for unbalanced samples . . . 88

5 Statistical models for healthcare: the Bayesian approach 89
5.1 Motivations . . . 89
5.2 Parametric Models . . . 91
5.2.1 Hierarchical linear mixed effects models . . . 91
5.2.2 Hierarchical generalized linear mixed effects models . . . 94
5.3 Semiparametric models . . . 96
5.3.1 Dirichlet Process for clustering . . . 97
5.3.2 Dependent Dirichlet Process for classification . . . 100
5.4 Optimal decision rules for hospital ranking and classification . . . 101
5.5 Problems due to unbalanced share: a Bayesian solution . . . 103

6 Statistical models for healthcare: more complex data 105
6.1 Unsupervised classification of multivariate functional data . . . 106
6.1.1 An overview of smoothing and registration techniques for functional data . 106
6.2 Depth measure for multivariate functional data and outlier detection . . . 107
6.2.1 Band depth and inference for multivariate functional data . . . 108
6.3 Generalized linear models with functional predictors . . . 111
6.3.1 Model for recurrent events . . . 111
6.3.2 Cumulative hazard smoothing and reconstruction . . . 112
6.3.3 Functional principal component analysis . . . 113
6.3.4 Generalized linear models with functional covariates . . . 115

III Data Analysis and Applications 116

7 Statistical analysis of STEMI Archive data 118
7.1 Descriptive analysis of data . . . 118
7.2 Frequentist approach to outcomes modelling . . . 131
7.2.1 In-hospital survival . . . 132

7.2.2 Long term survival . . . 134
7.2.3 MACE . . . 137
7.3 Frequentist approach to hospitals clustering . . . 137
7.3.1 Stadewide Survival Rate (SSR) . . . 138
7.3.2 Analysis of random effect estimates of a parametric GLME model . . . 139
7.3.3 Analysis of random effect estimates of a nonparametric GLME model . . . 142
7.3.4 Comparison of different methods . . . 145
7.4 Bayesian Hierarchical Models for Hospital Clustering . . . 146
7.4.1 Parametric and semiparametric models . . . 147
7.4.2 Posterior inferences and prediction . . . 149
7.4.3 Model fit and patients classfication . . . 153

8 Statistical analysis of other clinical surveys 156
8.1 Nonlinear parametric models for an epidemiologic enquire . . . 156
8.2 Nonlinear nonparametric models for an epidemiologic enquire . . . 161
8.2.1 Linear growth model . . . 161
8.2.2 Exponential growth model . . . 165
8.2.3 Logistic growth model . . . 166
8.2.4 Application to NON STEMI data . . . 169
8.2.5 Comparison of results . . . 170
8.3 Bayesian decision rules for provider profiling in cardiovascular context . . . 172
8.3.1 Statistical support to decision-making in health-care policy . . . 172
8.3.2 Application to MOMI^2 data . . . 174

9 Statistical analysis of ECG signals 177
9.1 Semiautomatic diagnosis for Bundle Brunch Block . . . 177
9.1.1 Data smoothing and registration . . . 177
9.1.2 Data analysis . . . 181
9.1.3 Results and discussion . . . 183
9.2 Depth measures for multivariate functional data . . . 187

Conclusions 193

References 196

List of Figures

1.1 Results chain of PM&E flow. 20
1.2 Pre-hospital use of antithrombotic and antiplatelet agents according to the different reperfusion strategies in STEMI. 23
1.3 Information collected in an Electronic Health Record, i.e. a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. 25
1.4 Administrative and clinical utilities of the Electronic Health Record. 26
1.5 Diagram of an Acute Myocardial Infarction (2) of the apex of the anterior wall of the heart (an apical infarct) after occlusion (1) of a branch of the Left Anterior Descendent (LAD) Coronary Artery. 27
1.6 An Acute Myocardial Infarction occurs when an atherosclerotic plaque slowly builds up in the inner lining of a coronary artery and then suddenly ruptures, causing catastrophic thrombus formation, totally occluding the artery and preventing blood flow downstream. . 28
1.7 Drawing of the heart showing anterior left ventricle wall infarction. 28

2.1 All steps of the integrated process of care provided by an efficient network. 32
2.2 Workflow for measuring performances and outcomes and taking decisions in cardiovascular healthcare process. 37
2.3 Extent of Myocardial Salvage along time. *Gersh (2005), JAMA* ***293****, 979–986*. 38
2.4 Pre-hospital times in STEMI. 39
2.5 Components of total ischaemic time in STEMI. 40
2.6 Time sequences stratified according to different ways of admission and/or pattern of care. . 40

3.1 Star Scheme structure of Public Health Database (PHD) of Regione Lombardia. Numbers are referred to a two-years time period between 2003 and 2005 43
3.2 Example of final result of filled sheet of STEMI Archive. 47
3.3 An example of file *Rhythm*. 51
3.4 An example of file *Median*. 51
3.5 Einthoven limb leads 52
3.6 Scheme of the stylized shape of a physiological single beat, recorded on ECG graph paper. Main relevant points, segments and waves are highlighted. 53
3.7 Conduction system of the heart: 1. Sinoatrial node; 2. Atrioventricular node; 3. Bundle of His; 4. Left bundle branch; 5. Left posterior fascicle; 6. Left-anterior fascicle; 7. Left ventricle; 8. Ventricular septum; 9. Right ventricle; 10. Right bundle branch. 54
3.8 The source DWRETEIMA aggregates all sources of data requested by STEMI Archive in the single record connected to each patient admission. 55
3.9 Example of subfolder collecting way of admission, symptoms and times. 55
3.10 Sketch of integration between STEMI Archive and Public Health Database. 57

3.11 An Overview of the Steps That Compose the KDD Process. 59

5.1 Graphical representation of the hierarchical model in 5.2. 92

7.1 Flanked Boxplots of patients' age, stratified by gender. 119
7.2 Declared symptoms stratification. 120
7.3 Boxplots of distributions of times from admittance to the transferring hospital to admittance to the receiving hospital (left panel) and from admittance to the receiving hospital to Balloon (right panel), for the 199 transferred patients. 123
7.4 Boxplot and histogram of OB time for patients with OB less than 24 hours. 124
7.5 Boxplot and histogram of OD time for patients with OD less than 6 hours. 125
7.6 Boxplot and histogram of DB time for patients with DB less than 6 hours. 125
7.7 Flanked boxplots of DB time stratified according to ECG tele-transmission. 126
7.8 Boxplot and histogram of EB time for patients with EB less than 6 hours. 127
7.9 Benchmark of DB time against Expoure for each hospital of STEMI Archive. The yellow line is the threshold of 25 cases inserted, the red one is the threshold of 50% of patients treated according to guidelines, and the green line the threshold of 50% of patients treated according to guidelines. 128
7.10 Flanked Boxplots of DB time for each hospital of STEMI Archive (only for patients whose DB time is less than 6 hours). 128
7.11 Benchmark of median DB times against median times of first ECG for each hospital of STEMI Archive. The diameter of each circle is proportional to the number of cases inserted in the registry by each hospital. Red lines indicate the thresholds of acceptability (80 minutes and 10 minutes respectively) suggested by international guidelines [151]. The green square shows the overall median of all hospitals. 129
7.12 Number of patients and in-hospital mortality (%) for each hospital of STEMI Archive. . . 130
7.13 Flanked boxplots of OB time in logarithmic scale for patients with positive (left) and negative (right) outcome of ST resolution. 133
7.14 Estimated survival surfaces in different case-mix scenarios, obtained fitting a GLM model for survival outcome. Green points indicate the survival probability for a patient aged 75 and with 50% of ejection fraction at the entrance. 135
7.15 Flanked boxplots of FE for patients presenting no MACE (left) and at least one MACE (right). 137
7.16 Silhouette plots for the choices of the number of cluster, with $k = 2$ (left) $k = 3$ (right) respectively. 140
7.17 Clustering of estimated random effects of the GLMM model in (7.3) into 2 (left panel) and 3 (right panel) groups respectively. 141
7.18 Survival surfaces in different case-mix scenarios (best case, first row vs worst case, second row), in the hospital centroid of group "A" (left panel), "C" (central panels) and "B" (right panels) respectively. Green points indicate the survival probability for a reference patient aged 75 and with 50% of ejection fraction at Admittance. 142
7.19 Estimted survival surfaces in different case-mix scenarios (best case, first row vs worst case, secondo row), in a hospital belonging to group "A" (left panel) and "B" (right panels) respectively. Green points indicate the survival probability for a reference patient aged 75 and with 50% of ejection fraction at admittance. 144
7.20 Posterior 95% CIs of hospital random intercepts with at least ten patiens. The estimates are in increasing order of number of patients. The last estimates represent new random intercepts. 151

7.21 Posterior 95% CIs of hospital random intercepts plus *Milano* effect with at least ten patiens. The hospital effect in Milano are depicted in blue dashed, those ooutside Milano in red solid lines. The estimates are in increasing order of number of patients per hospital. The last two intervals represent new random intercepts for a hospital in and outside Milano, respectively. 152
7.22 ROC curves for the three models. 153
7.23 90% posterior predictive CIs of all the patients (ordered by increasing median) under the DDP model. The positive outcomes are in blue and the negative ones in red. 154
7.24 90% posterior predictive CIs of all the patients from hospital 19, under the DDP model. The CIs corresponding to alive patients are in blue solid line, while those corresponding to dead patients are in red dasehd. There are six unclassified patients and only one misclassified. . 155

8.1 Left panel: Number of AMI without ST-elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. Right panel: Standardized number of AMI without ST-elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. For each hospital the yearly number of diagnoses has been divided by the hospital total number of diagnoses in the time period $2000-2007$. . . 157
8.2 Estimated growth curves through model (8.1) together with the original data. 158
8.3 Silhouette plot of PAM procedure on the estimated inflection points with $k=3$ clusters. . . 160
8.4 Estimated logistic growth curves for different medical institutions. 160
8.5 Simulated data (left panels), `npmlreg` (central panels) and NLNPEM classification (right panels) in lin2I, lin3S, lin9SI, lin10I and lin10S datasets respectively. Different colors are used to represent real groups (left panels), groups identified by `npmlreg` and NLNPEM methods (central and right panels respectively). 164
8.6 NLNPEM classification in exp2A, exp3A and exp10A datasets respectively with exponential model. 166
8.7 NLNPEM classification in logis2A, logis2I, logis4AI, logis3A, logis3I, logis9AI, logis10A and logis10I datasets respectively with logistic growth model. 168
8.8 Standardized number of AMI without ST-segment elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. The year has been centered and normalization has been carried out standardizing the yearly number of diagnoses for each hospital by total number of diagnoses in the time window $2000-2007$. Real data are colored according to the NLNPEM clusters and NLNPEM fitted models are superimposed. 170
8.9 Number of hospitals labelled as "unacceptable" as a function of k, under the Squared Loss function (solid black) and the LINEX Loss function (dotted blue). The threshold parameter β_t is 3.6635. 176

9.1 Raw data of the eight leads (black points) and wavelet functional estimates (blue) for a normal subject. 179
9.2 First central finite difference of the eight leads (gray) and wavelet estimates of the first derivatives (blue) for a normal subject. 179
9.3 Original I leads for the 198 patients (left) and registered ones (right). Vertical lines indicate position of mean landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset} 181
9.4 Silhouette plots of the clustering result obtained via multivariate functional k-means procedure, setting $k=2,3,4,5$ and with distance given by (9.1); data are ordered according to an increasing value of silhouette within each cluster, and are coloured according to the cluster assignment. 183

9.5 Smoothed and registered ECG traces (QT-segment): the whole dataset is coloured according to the final cluster assignments of multivariate functional 3-mean clustering, with distance given by (9.1); the superimposed black lines are the three final cluster centroids (functional means). Each panel correspond to a different lead of the ECG traces. 185
9.6 Raw signals of the 100 physiological patients. 188
9.7 Raw signals of the 50 pathological patients. 189
9.8 Functional boxplots of each component (*lead*) of the 100 physiological ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4). . . 190
9.9 Functional boxplots of each component (*lead*) of the 50 pathological (Left Bundle Brunch Block) ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4). 191
9.10 Functional boxplots of each component (*lead*) of the 150 physiological (100) and pathological (50 Left Bundle Brunch Block) ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4). 192

List of Tables

7.1 Stratified overall age. . . . 119
7.2 Stratified way of admission, without patients transferred from one hospital to another one. . 120
7.3 Number of dead patients within each Killip class. . . . 121
7.4 Reperfusion therapy of patients of STEMI Archive . . . 122
7.5 Causes of missed treatment. . . . 123
7.6 Summary indexes concerning distributions of the main process indicators. . . . 124
7.7 Estimated survival probability of a reference patient ($Age = 75$ years, $FE = 50\%$ in different case-mix scenarios. . . . 134
7.8 Providers' clustering according to SSR criterion. . . . 138
7.9 Providers clustering according to GLME model random effect estimates criterion. . . . 141
7.10 Providers clustering according to `npmlreg` random effect estimates criterion. . . . 144
7.11 Providers clustering provided by the three different criteria described in Section 7.3. . . . 145
7.12 Posterior 95% CIs of the fixed effects . . . 150
7.13 Posterior 95% CIs of the random effects' covariance matrices elements. . . . 151
7.14 Predictive tables using point estimates and cut-off point equal to $\bar{p} = 0.97$. . . . 153
7.15 Predictive tables using 90% CIs and cut-off point equal to 0.5. . . . 154

8.1 Fixed effects estimates and Anova table for model (8.2). . . . 159
8.2 Fixed effects estimates for model (8.3). . . . 161
8.3 Normalized Wasserstein distances and $-2\log L$ index for `npmlreg` and NLNPEM algorithm respectively in the simulated linear cases. . . . 165
8.4 Normalized Wasserstein distances for NLNPEM algorithm in the simulated exponential cases. 166
8.5 Normalized Wasserstein distances for NLNPEM algorithm in the simulated logistic cases. . 169
8.6 Estimates carried out by `npmlreg` and NLNPEM method on lin2I dataset, where intercept is considered as random, with 2 balanced groups. . . . 170
8.7 Estimates carried out by `npmlreg` and NLNPEM method on lin3S dataset, where slope is considered as random, with 3 unbalanced groups. . . . 171
8.8 Estimates carried out by `npmlreg` and NLNPEM method on lin10I dataset, where intercept is considered as random, with 10 balanced groups. . . . 171
8.9 Providers labelled as "unacceptable", for different loss functions and different values of the threshold. . . . 175

9.1 Landmarks obtained at the end of the registration procedure, as the mean of landmarks of all the curves, and used to select the portion of smoothed and registered ECG curves relevant to our analysis (first line of the table); in the second line, landmarks standard deviations. Landmarks values are referred to a registered time in ms. . . . 180

9.2 Confusion matrix related to patients disease classification. Results are obtained performing 3-means clustering algorithm on interval lengths. 180
9.3 Confusion matrices related to patients disease classification. Results are obtained by application of multivariate functional 3-means clustering algorithm to smoothed and registered QT-segment of ECG curves, with different choices of the distance between ECGs: H^1 norm (eq. (9.1), first table), H^1 semi-norm (eq.(9.2), second table) and L^2 norm (eq. (9.3), third table). In the first table, cluster 1,2,3 respectively correspond to orange, green and red in Figure 9.5. 184
9.4 Mean misclassification cost (first row) and standard deviation (second row) computed over 20 repetitions of the cross-validation procedure via equation (9.4). 186

Introduction

Since a strong attention has been devoted in the last decades to healthcare management for social, medical and economical reasons, over recent years also the analysis, development and improvement of suitable tools apt at measuring the quality of care has become a research field of extreme importance. Within this context, since patient outcomes enable researchers (at least in part) to assess the quality of care, there has been a widespread diffusion of techniques for monitoring and evaluating the underlying processes generating such outcomes (see [171], [207], [208], [212] and in particular the bibliographic overview reported in [155]). This task is nowadays unavoidable in order to get a sensible improvement of healthcare services quality, as well as to contain economical costs. For this reasons, performance measurements have to be carried out both at hospital and at physicians levels. In order to achieve a suitable strategy for assessing healthcare performances, it is necessary to identify suitable operating protocols, as well as the functional competence of institutions, and then to monitor them over time in order to understand the behavioural effects of rules on results (as shown, for example, in [170] and [181]). In this way, a continuous collection of data proved to be mandatory, entailing a very strong burden as long as the number of players and the number of indexes involved increases. Hence, in view of setting up a process for monitoring and evaluating an healthcare system, it is necessary to answer several questions concerning (i) the definition of proper outcomes to be measured in order to fulfill the desired targets; (ii) the data to be collected and how they should be collected; (iii) how data will be checked and audited; (iv) how data should be analysed (and how frequently) and adjusted within a given context.

With this respect, a dataset (clinical registry and/or administrative database) describes the process underlying itself. The more complex, structured and detailed is the dataset, the more difficult is the analysis required for its exploitation (see [13] and [83]). In fact, not always the availability of more information leads to more accurate predictions, since confounding effects grow up with complexity of the problem [54]. In other words, monitoring healthcare systems through data collections asks for a careful design of experiment and a high quality of collected data, for shared standards of collection as well as for strict controls on the filling compliance and the reliability of the data. Moreover, it calls for suitable statistical methods for analysis, modelling and predictions. Only in this way it is possible to derive models which are capable of realistic interpretations of the process underlying the dataset. The statistical analysis of data provides an invaluable insight into the behaviour of healthcare processes [30]. Statistical techniques, when applied to measurement data, can be used firstly to highlight areas that would benefit from further investigation [133], then to model processes relating patterns of care, patients case-mix, hospital influences and outcomes of interest, and finally to make predictions (see [62] and [206]). Statistics enables the researchers to identify variation within the process under observation. Understanding, modelling and then quantifying this variation are the first steps towards quality improvement.

An important goal of Regione Lombardia (healthcare division) is the use of performance measures for monitoring cardiological and cardiovascular healthcare offer, as well as to assess institutions

within the regional healthcare service in order to provide evidence for initiatives aimed at enhancing professional accountability in the public sector. Specifically, a Strategic Program, named *"Sviluppo di nuove strategie conoscitive, diagnostiche, terapeutiche e organizzative in pazienti con sindromi coronariche acute"* (`www.salute.gov.it/ricercaSanitaria` [35]), has started in 2008 with, among others, the following main goals:

- To point out a comprehensive clinical and epidemiological picture of how Acute Myocardial Infarction (AMI) is treated in Regione Lombardia.
- To assess the effectiveness of patterns of care for AMI patients, in order to invest in innovations starting from real epidemiological evidence and needs.
- To exploit administrative databanks for addressing clinical and epidemiological enquires.
- To highlight critical situations in healthcare delivery and then to improve hospital performances.
- To provide people in charge with healthcare government with decisional support based on statistical evidence and real time data.

In order to address these issues, suitable methods to collect, analyse and model data are needed. The results of statistical analyses carried out on data arising both from clinical registries and administrative databanks may influence funding and policy decisions, and are used to generate feedback for providers (see [36] and [112]). On the other hand, for reports on the performance of health care providers to be effective, profiling must be done using the best statistical methods. The providers' profiling based on current data collections is a new way for improving quality of healthcare offer. To this aim and to set an efficient network among providers, it is necessary a shared information-technology systems of data collection and advanced statistical methods able to classify providers, to quantify their effect on outcomes of interest at patients' level, to analyse complex data arising from biomedical context and to make reliable predictions.

Statistics is then of paramount importance in more than one step of the cardiovascular healthcare process, especially in supporting this new concept of "real-time" epidemiology based on observational clinical registries and administrative databanks. In fact, as shown and explained in the thesis, the statistician plays a central role during the design of experiment, carries out the monitoring of data collection, evaluates the process and produces a feedback for involved players, elaborates models necessary for providers' profiling, classification and outcomes prediction. The decisional support provided by statisticians is evidence based and it is based on real epidemiological evidence and needs, involving low cost data sources, i.e., real-time and sustainable from economic perspective.

In this thesis a general approach to model fitting aimed at clustering is considered, focusing on grouped (longitudinal) data arising from healthcare context, where examples of grouping factors are hospitals (and diseases) with respect to patients or patients themselves with respect to their own measurements performed over time. The main goal of the thesis is then to present a number of statistical techniques for the analysis of such data, in order to provide methods for supporting decisions of people in charge with healthcare government. Clinically speaking, we will focus specifically on problems related to the improvement and optimization of pattern of care for patients affected by Acute Coronary Syndromes. On the other hand, the main statistical topic we will deal with is clustering carried out starting from random effects models estimation.

In fact, mixed effects models are used in a wide variety of biostatistical contexts, and can be analysed both from a classical and Bayesian viewpoint. We can distinguish two types of applications: those in which the random effects are nuisance parameters, and are not of direct interest, and

those in which the individual effects are of paramount interest. Although mixed effects models are used in many applications in medical statistics, for our scopes the most interesting and new one is the problem of hospital comparisons using routine performance data. Among other benefits, this approach provides a diagnostic criterion to detect clusters of providers with unusual results.

Thesis outline

The outline of the thesis is the following:

- In **Part I**, the funding project of the Ph.D. Scholarship and the clinical context it concerns are presented, together with data arising from the whole process under investigation. The idea is to carry out an overview of motivating problems the project wants to address and to present data used to carry out analyses.

  In particular, in **Chapter 1** motivations for performing Monitoring and Evaluation of healthcare systems (Section 1.1) are provided, together with the Information Technology devices (Section 1.2) adopted to do it, in order to set the context and motivating the new concept of epidemiology and healthcare assessment arising from the application of these methods. Finally (Section 1.3), a brief overview of clinical diseases of interest is proposed, in order to make the reader familiar with some clinical concepts and terms he/she will find again later on. In **Chapter 2**, the idea of Cardiological Network is introduced and aims and scopes of the Strategic Program are detailed. In particular, the healthcare process of interest, i.e., the pattern of care which patients affected by Acute Coronary Syndromes undergo, is explained together with the process indicators necessary to monitor it and then make it more efficient and effetive. Finally, **Chapter 3** describes all data sources arising from each step of the healthcare proces of interest, which will be analysed in Part III using statistical methods proposed in Part II, in order to give answers to healthcare problems set in this Part.

- In **Part II**, statistical methods for the analysis of data presented in Part I, Chapter 3 are presented. The aim is to model hierarchical structured data, longitudinal and grouped data, and multivariate functional data, in order to carry out both assessment of the effect of grouping factors and reliable prediction and/or classification at patients' level.

  In particular, **Chapter 4** provides an overview of the frequentist approach to Mixed Effect Models, in the Linear, Generalized Linear and Nonlinear case. Both parametric and nonparametric estimations of random effect distribution are considered, with the aim of classifying random effect estimates in order to investigate the presence of upper level clusters among grouping factors. **Chapter 5** refers to a similar approach to hierarchical models from a Bayesian perspective, adopting Dirichlet and Dependent Dirichlet Process for modelling random components and carrying out clustering of random effects, taking advantage by the semiparametric setting. Results are set in the framework of Bayesian Decision Theory. Moreover, the problem of strongly unbalanced sample size is faced. Finally, in **Chapter 6**, statistical methods for dealing with multivariate functional data arising from medical diagnostic devices and longitudinal event-dependent data arising from the integration of clinical registries and administrative database are proposed.

- In **Part III**, data described in Chapter 3 are analysed adopting methods proposed in Part II, in order to address different clinical and management problems brought out in Part I.

We present firstly the analyses carried out on STEMI (ST segment Elevation Myocardial Infarction) Archive data (**Chapter 7**), from the descriptive analysis and data mining of the clinical registry up to the application of both frequentist and Bayesian mixed effect models for clustering and prediction. Then in **Chapter 8** further analyses on different data sources are presented, consisting of nonlinear parametric and nonparametric mixed effects models applied to administrative data for the identification of clusters of hospitals, as well as Bayesian decision rules for providers' profiling. Finally, in **Chapter 9** clustering of multivariate functional data (Electrocardiograms) and nonparametric techniques are discussed and implemented, aimed at semi-automatic diagnosis of a specific type of infarction and at multivariate functional outlier detection and inference respectively.

All the analysis are carried out using `R` software, version `2.13.0` [124].

Part I

Regione Lombardia cardiovascular healthcare system: data sources, hospital network and Strategic Program

In order to improve something, you must be able to change it.
In order to change it, you must be able to understand it.
In order to understand it, you must be able to measure it.

Key words: Process Monitoring and Evaluation; Provider Profiling; e-Health; Electronic Health Record; Clinical Registries; Administrative Databanks; Record Linkage; Cardiovascular Syndromes; ST-Elevation Acute Myocardial Infarction; Electrocardiography.

Chapter 1

A patient-focused documentation for a new definition of epidemiology

In this chapter, some concepts like monitoring and evaluation of healthcare service are introduced in order to set the context the Strategic Program moves into. Moreover, the Electronic Health Record and its utilization as an instrument for the definition of a new epidemiology are pointed out.

1.1 Outcomes and performance measurements of public healthcare systems

Public programs of any type and size across the nation are shifting from seeing themselves as accountable for creating and carrying out activities to being accountable for achieving results, meeting goals and improving the quality of services. Such transformation implies capability of changing the way you work, the way you assess your work, and the way you inform others of your progress. And it can be difficult redefining roles and responsibilities, creating new collaborations, overcoming resistance to change [191]. In public health, such struggling to understand a program's role and striving to fairly evaluate how well it's carrying out that role has its realization in the so called *Performance Monitoring* (PM) [178], which records, analyses and publishes data in order to give the public a better idea of how government policies change the public services and to improve their effectiveness. The developing countries can begin to address the challenges of working within results based orientation and thus moving towards a result-based management approach to public sector management by documenting their performance with credible information that goes beyond the traditional reporting on inputs, activities, and outputs to now include outcomes and impacts.

In what follows, definition and analysis of the main concepts concerning performance measurements [183], monitoring and evaluation in healthcare are proposed. They will be recalled and adapted to the context of providers' evaluation within Regione Lombardia policy for cardiovascular healthcare in the next chapter. In fact, the idea of the project this thesis deals with is to develop specific indicators, starting from clinical and administrative databases, in order to perform such evaluation through suitable performances indexes and so doing providing decisional support to people in charge with healthcare government.

Performance measurement analyzes the success of a work group, program, or organization's efforts by comparing data on what actually happened to what was planned or intended. On the other hand, **performance management** uses performance information to manage organizational capacity

and processes, reviews programs, assesses and revises goals and objectives, monitors progresses against targets; conducts employee evaluations. Performance measurement is needed as a management tool to clarify goals, document the contribution toward achieving those goals, and the benefits received from the investment in each program. Therefore, performance measurement (*management for results*) seeks to assess, verify and demonstrate results, while performance management (*management by results*) focuses more on experimentation, innovation, process, learning and responsiveness. Thus, performance management helps set agreed-upon performance goals, allocate and prioritize resources to meet those goals, and report on the success in meeting those goals.

Monitoring is defined as a continuing function that aims primarily to provide the management and main stakeholders of an ongoing intervention with early indications of progress, or lack thereof, in the achievement of results. An ongoing intervention might be a project or other kind of support to an outcome. It provides managers and stakeholders with regular feedback on program performance.

Evaluation on the other hand, provides a judgment based on assessments of relevance, appropriateness, effectiveness, efficiency, impact and sustainability of development efforts. It involves a rigorous and systematic process in the design, analysis and interpretation of information to answer specific questions. It highlights both intended and unintended results, and provides strategic lessons to guide decision-makers and inform stakeholders [199]. Though monitoring can provide critical inputs to evaluation by way of systematic collection of data and information, yet an evaluation system serves a complementary but distinct function from that of a monitoring system within a performance management framework.

The **Performance-based Monitoring and Evaluation (PM&E)** combines both, the traditional approach of monitoring implementation with the assessment of performance and results. It is this linking of both implementation progress with progress in achieving the desired objectives or goals of government policies and programs that make PM&E most useful as a tool for public management. Implementing a PM&E system allows the organization to modify and make adjustments to the implementation processes for achievement of desired results and outcomes. However, introducing PM&E for Result-based Management will often require interventions that address a wide range of possible “determinants of performance”. These determinants are technical, organizational and behavioural. The sustainable PM&E system is more likely to emerge from a cohesive strategy harmonizing these three determinants. The result is a change, which can be an improvement, an increase, a strengthening, a reduction, or a transformation in attitudes and behaviours of a given group.

Outcomes are the changes necessary to achieve the project purpose. The long-term socioeconomic results to which projects contribute are Impacts, and are necessary to achieve the project goal. Hierarchy of Results must be chained in a continuum for Performance Management and Impact. Figure 1.1 shows the flow chart of the chain, where activities lead to outputs, outputs lead to outcomes, outcomes fulfill the purpose and lead to impact, impact will lead to the goal.

Performance indicators are variables to measure changes towards progress of results and should be identified for each output and outcome. Performance indicators should answer the question, “What will be observed if the result is achieved?”. If a result is to improve or increase knowledge or capacity, baseline data may need to be collected early on during project implementation to allow measurement of what has changed. It is essential that one to three performance indicators be identified for each result, particularly for outcomes and outputs. These indicators may be qualitative and quantitative and must serve to measure the achievement of results. The performance indicators

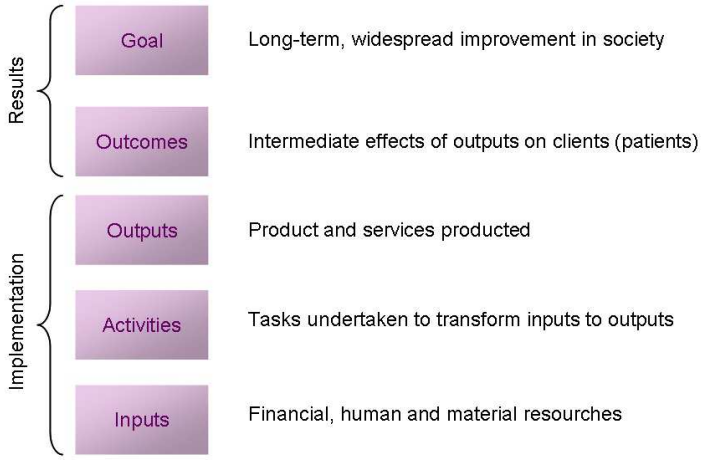


Figure 1.1: Results chain of PM&E flow.

should be objectively verifiable and *CREAM* (Clear Relevant Economic Adequate Monitorable) for providing meaningful measurement and being useful [183]. The ultimate aim of performance information system is to promote the utilization of data and performance information emerging from it for decision making from the operational to the policy making level. In the process towards selection of performance indicators, a step-wise approach is required starting from analysis of the management functions at each level, identifying their information needs according to those functions and deciding on performance indicators that provide those information needs.

An Outcome Performance Indicator must be **CREAM**:

Clear: Precise and unambiguous
Relevant: Appropriate for measure
Economic: Available at reasonable cost
Adequate: Sufficient to assess performance
Monitorable: Amenable to validation

On the one hand, rational selection of indicators will help to address the classic problem of too many indicators. On the other hand, use of hierarchy of indicators at different levels will help to focus on the national strategic outcomes through bottom-up filtration until monitoring and evaluation information needs are met at the provincial, regional and national levels.

In what follows (see Chapters 2 and 7), our attention will be focused in pointing out suitable performance indicators within the Strategic Program aims and scopes, easy to be measured and able to quantify the providers' efficiency in offering and managing suitable patterns of care for patients affected by Acute Coronary Syndromes (ACS, see Section 1.3). Furthermore, it is important to build a review mechanism into the system. The various sets of indicators for all levels should be reviewed and challenged in terms of their effectiveness, cost implications, data quality and source of data collection. In order to do this, it is necessary to develop linkages between performance indicators with the following aims:

i. to whole results chain $\rightarrow$ The performance indicators must logically link outputs, immediate outcomes, intermediate outcomes, and end outcomes.

ii. to planning and budgeting $\rightarrow$ The performance indicators should link planning, budgeting and accountability processes, otherwise their role will be limited to the implementation and operational management targeting lower level results.

Once the indicators are finalized and consensus has been reached, the next step involves preparation of data collection and reporting tools and instruments. The templates for data cross check to ensure data quality of performance information should also be part of the information systems.

A performance monitoring plan is a critical tool for planning, managing, and documenting data collection (in Section 1.2 we will see how collection of information can be performed through the use of Electronic Health Records). It contributes to the effectiveness of the performance monitoring system by assuring that comparable data will be collected on a regular and timely basis. This is essential to a credible and useful performance-based management approach. It involves the regular collection of information on actual results and demonstrates whether a project, program, or policy is achieving its stated goals. That's why the first need addressed within the Strategic Program has been the planning and the activation of a new registry for collecting data which is common to all providers on the territory of Regione Lombardia and which is designed to be easily handled and analysed by researchers asked to provide monitoring and evaluation of providers' performances during the time.

1.1.1 Performance measurement in public health

Assessing service delivery at the local level of government and measuring public health with the intent of gathering information to improve public health practice is not a new enterprise in clinical context (see [178], [191] and [183] among others). Anyway, linking the measures, or indicators, to program mission, setting performance targets and regularly reporting on the achievement of target levels of performance are new features in the performance measurement movement sweeping across the public health service. This is also what Strategic Program aims to achieve within the cardiovascular healthcare planning of Regione Lombardia.

Within the healthcare context, performance measurement is a simple concept with no simple nor unique definition. Essentially, performance measurement analyses the success of a work group, program, or organization's efforts by comparing data on what actually happened to what was planned or intended. Two simple but quite effective definitions of performance measurement are the following:

↪ *Performance measurement is the selection and use of quantitative measures of capacities, processes, and outcomes to develop information about critical aspects of activities, including their effect on the public.*

↪ *Performance measurement is the regular collection and reporting of data to track work produced and results achieved.*

To understand the first definition, you need to know what is meant by capacity, process, and outcome. These are three key components of public health practice. On the other hand, the second definition underlines other critical aspects of measurement and evaluation process, i.e. the regularity and the continuity of monitoring. Moreover, these definitions enable us to highlight two fundamental aspects connected with the role of a statistician in the PM&E process: how to exploit the results of measurements in order to develop mechanisms for acquiring information about critical aspects of the process, and how to design surveys for data collection so that such information can be pointed out as well as possible.

Back to the first definition of performance measurement, *Capacity* means the ability of a work group, program, or organization to carry out the essential public health services, and in particular, to provide specific services; for example, disease surveillance, community education, or clinical screening. This ability is made possible by specific program resources as well as by maintenance of the basic infrastructure of the public health system. *Process* means the things that are done by

defined individuals/groups - or to, for, or with individuals/groups - as part of the provision of public health services. *Process* means all the things we do in public health practice. Finally, *Outcome*, as we said before, means a change in the health of a defined population that is related to a public health intervention.

Public health goals are broad-based community goals, not a specific goal for a specific organization, then it is difficult to get clear causal connections. Consequently, a good starting point when thinking about implementing performance measurement in public health is to understand those things that are unique or different about public health practice. Public health offers services to the whole population in all of its diversity. These services can be resumed in the following list:

1. Monitor health status to identify and solve community health problems (i.e., community health profile, vital statistics, and health status)
2. Diagnose and investigate health problems and health hazards in the community (i.e., epidemiologic surveillance systems, laboratory support)
3. Inform, educate, and empower people about health issues (i.e., health promotion and social marketing)
4. Mobilize community partnerships and action to identify and solve health problems (i.e., convening and facilitating community groups to promote health)
5. Develop policies and plans that support individual and community health efforts (i.e., leadership development and health systems planning)
6. Enforce laws and regulations that protect health and ensure safety (i.e., enforcement of sanitary codes to ensure safety of environment).
7. Link people to needed personal health services (i.e., services that increase access to health care).
8. Assure competent public and personal health care workforce (i.e., education and training for all public health care providers).
9. Evaluate effectiveness, accessibility, and quality of personal and population based health services (i.e., continuous evaluation of public health programs).
10. Research for new insights and innovative solutions to health problems (i.e., links with academic institutions and capacity for epidemiologic and economic analyses).

Public health practitioners can use these broad service categories for developing performance measures of capacity (the capacity to conduct each service), process (the processes used to conduct each service), and outcomes (the results of each service). In summary, public health offers a huge array of services to a huge number of people. Even if it hopes to influence people's lives and well-being, it is not solely responsible for either.

Our goals within the Strategic Program and this thesis will focus on points 1, for what concerns the population of patients affected by Acute Coronary Syndromes, and 5, 6, 9 and 10, in terms of providing decisional support through the statistical analysis and modelling of data arising from clinical registries and administrative databanks.

1.1.2 Why do performance measurements?

In order to improve something you have to be able to change it. In order to change it you have to be able to understand it. In order to understand it you have to be able to measure it.

Performance measurement compels anyone who wants to implement it to reassess work, group, programs and/or organization goals and objectives. Goals describe where the direction to be pursued should lead. Objectives define specific results that will show movement toward goals. Thinking about how to measure performances might inspire who wants to carry it out to set new long-term goals, new long-term and short-term objectives, and new or revised approaches to work for reaching them. Rethinking goals and objectives might result in developing a new strategic plan for many efforts. Implementing performance measurement gives the opportunity to create working arrangements with other groups, programs, agencies, organizations and stakeholders. This collaborative cross-fertilizing can make for a stronger approach to meeting goals. A strictly program-specific approach might lead to duplication of data collection efforts or missed opportunities to adopt measures that can be used by more than one program. Implementing performance measurement also provides an opportunity to assess more pragmatic accountability issues, such as evaluating and defining roles and responsibilities, and levels and lines of authority. Moreover, it enhances the definition of operating integrated protocols to be applied in emergency intervention. Since we will focus in the following chapters on Acute Coronary Syndromes, in Figure 1.2 an example of protocol for pre-hospital use of specific drugs for a particular type of acute cardiovasular event is shown (i.e., the protocol for pre-hospital use of antithrombotic and antiplatelet agents according to different reperfusion strategies in STEMI, see also [213] for further details).

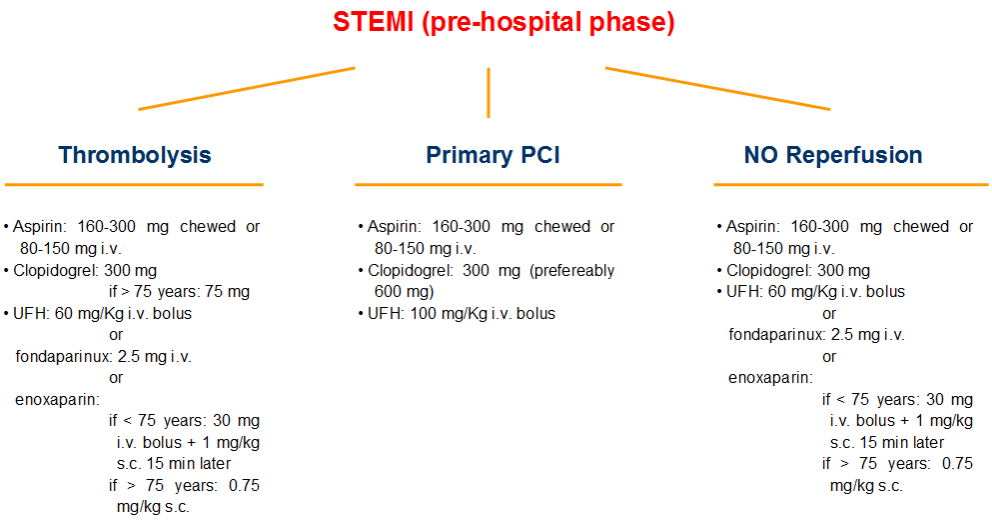


Figure 1.2: Pre-hospital use of antithrombotic and antiplatelet agents according to the different reperfusion strategies in STEMI.

Within the cardiovascular policy of Regione Lombardia, all this means to deal with the organization, rationalization and long-term planning of the hospitals network that provides healthcare services to the patients affected by Acute Coronary Syndromes. If the main and most important goal is to provide a good treatment for all patients who need it, this assumes several different declensions once it must be implemented in practice. Moreover, any practical goal in this context must be subordi-

nated to times constraints, i.e. you are asked not only to do the best, but to do it quickly. Finally, the treatment of a patient affected by Acute Coronary Syndromes is a complex process involving more than one player (and institution) within the healthcare context: from General Practitioners to hospital physicians, passing through the network of 118 emergency rescue. As a consequence, not only an hard work on hospital organization is expected, but also a joint plan taking into account the prehospital care phase. Measuring such a complex process, avoiding redundances and waste of time and money, is the real challenge of the Strategic Program presented in Chapter 2.

Often, several issues of those mentioned up to now can be addressed with already existing resources. For example, a question should always be taken into account when there is the need of identifying performance measurement is: what measures can be implemented with existing databases, research methods, and personnel, rather than new or complicated data collection schemes? In most cases, the greatest part of the key measures can be derived and reported from existing systems and processes. When it comes to identifying the data sources at hand, often it is easy to notice that you are surrounded with sources that can be used for performance measurement. These include, for example, data and information collected from stakeholders, including recipients of services, for example through surveys and case studies. Ultimately, it would be desirable that the performance measurement process reveals improvement on past performance or, if an attainment level of performance has been achieved, at least the steady maintenance of that level. That means that an ongoing assessment of the capacities of the staff is required.

Anyway, having a lot of data does not necessarily mean having a lot of meaningful performance measurement information. The philosophy of *"Let's collect everything and we'll figure it all out in the morning"* (see [184]) is a very expensive and often useless philosophy. We need a different model, i.e., a model that derives meaningful information from what the stakeholders want to know about performance. If there is a good question about performance, a good measure can be provided for it. If there are 50 good questions, the answer is probably a meaningful, focused database. Matching performance measurement data and information demands advanced statistical methods and modelling skills. That's why suitable informatic tools for collecting data and models for analyzing them are requested to reach these goals.

1.2 Electronic Health Record (Fascicolo Sanitario Elettronico)

When you organize to develop a performance measurement process, you are asking a lot of your internal stakeholders: the people whose performance will be measured. You are asking them to understand, accept, and promote the concepts and values behind performance measurement. You are asking them to think about how and why they conduct their work tasks and to rethink the goals and objectives of their work group. You are asking them to develop ways to measure their own performance and that of others. And you are asking them to report on the results of their performance measurement. You are asking them to generate change. That's a lot to ask. Consequently, one of the key components in developing an effective performance measurement process is providing those involved with the assistance they need to understand and implement the process, as well as the training they need to improve their performance.

(P. Lichiello, Guidebook for Performance Measurement) [184]

The role of Information Technology (IT) in the PM&E process of healthcare is crucial, as well as the statistical analysis of data arising from digital data collection procedures. In fact, the complexity of healthcare request is growing up along time: citizens are more and more conscious of their needs

and aware on answers that they could pretend by the clinical world. Moreover, the prevalence of chronic-degenerative diseases like the cardiovascular ones asks for continuing patterns of care involving many different players of the health care context. From such a context a new vision of the patterns of care comes out: it is no more a sequence of independent events, but becomes a specific path for each patient. This calls for the presence of networks and informatic systems that allow physicians to seek and share information needed for diagnosis and treatment of patients. Providing a common framework for collecting data enables to improve quality of data entry, to ease communication of results as well as communication among providers and physicians involved.

The study of IT solutions aimed at doing what mentioned above is called *e-Health*. This is a paradigm based on the centrality of the patient and aimed at connecting clinical and administrative needs. In this context can be set the *Sistema Informativo Socio Sanitario (SISS)* of Regione Lombardia, i.e. the Italian platform for supporting Electronic Health Record (see `http://www.siss.regione.lombardia.it/` for details) and the related Electronic Health Record (EHR), which is the main tool for defining a new epidemiology. The Electronic Health Record is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting, as shown in Figure 1.3. Among the information provided by such an instrument there are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports.

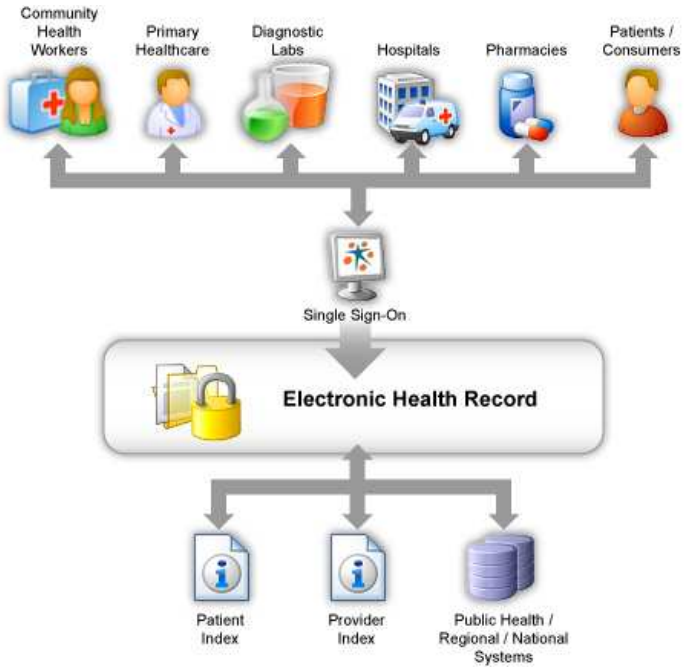


Figure 1.3: Information collected in an Electronic Health Record, i.e. a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting.

The EHR automates and streamlines the clinician's workflow. Through the linkage of information contained in different datawarehouses (like those described in Sections 3.1 and 3.2) the EHR has the ability to generate a complete record of a clinical patient encounter, as well as supporting other care-related activities directly or indirectly via interface-including evidence-based decision support, quality management, and outcomes reporting.

An EHR enables an administrator to obtain data for billing, a physician to see trends in the effectiveness of treatments, a nurse to report an adverse reaction, and a researcher to analyse the efficacy of medications in patients with co-morbidities. If each of these professionals works from a data silo, each will have an incomplete picture of the patient's condition. An EHR integrates data to serve different needs, as highlighted by the picture in Figure 1.4. The goal is to collect data once, then use it multiple times. EHRs are used in complex clinical environments. The data presented, the format, the level of detail, and the order of presentation may be remarkably different, depending on the service venue and the role of the user.

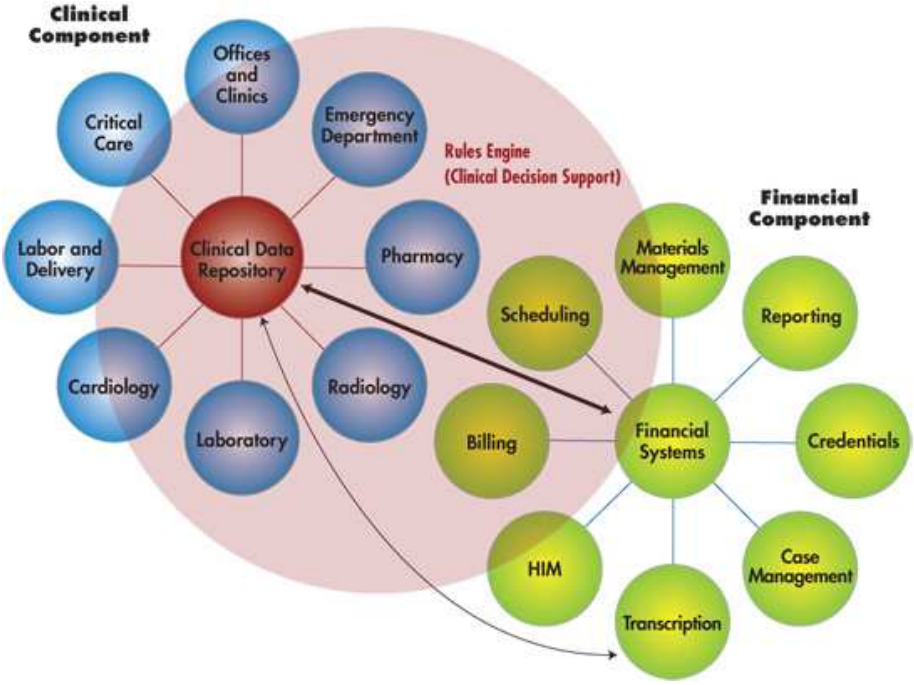


Figure 1.4: Administrative and clinical utilities of the Electronic Health Record.

If a clinician has integrated access to the semantic content of the data, then the system will be able to show, for example, all cases in which patients were diagnosed with cardiovascular diseases. In this sense, availability of data in such format enables to select sub population of interest for any clinical study. Retrospective surveys obtained in this way match all the problems of observational studies in terms of biases; on the other hand, they have the advantage of providing real time information without additional costs.

In the next section, details on pathologies we are interested in will be given, in order to understand which population we focus on when we perform queries in EHR contents.

1.3 Focus on Acute Coronary Syndromes

In this section, a brief overview of the pathology our studies deal with is presented, and the main clinical terms and concepts used in the following chapters are provided, in order to introduce the reader to the contents of the clinical registry described in Section 3.2.3.

Acute Coronary Syndromes (ACS) is a unifying term representing a common end result, i.e., the acute myocardial ischaemia. Acute ischaemia is usually, but not always, caused by atherosclerotic plaque rupture, fissuring, erosion, or a combination with superimposed intracoronary thrombosis (like in Figure 1.5), and is associated with an increased risk of cardiac death and necrosis. It encompasses Acute Myocardial Infarction (AMI), resulting in ST segment Elevation Myocardial Infarction (STEMI) or non-ST segment Elevation Myocardial Infarction (NON-STEMI), and unstable angina (ST segment is a subsegment of the ECG trace, see Paragraph 3.2.4). Recognizing a patient with ACS is important because the diagnosis triggers both triage and management. Acute Myocardial Infarction (AMI), commonly known as heart attack, is a common presentation of ischaemic heart disease. Worldwide more than 7 million people per year experience an event of Infarction (see [217]). This makes the AMI the leading cause of death for both men and women worldwide. As mentioned before, it results from the interruption of blood supply to a part of the heart, causing heart cells to die. This is most commonly due to occlusion of a coronary artery following the rupture of a vulnerable atherosclerotic plaque, which is an unstable collection of lipids (cholesterol and fatty acids) and white blood cells in the wall of an artery. The resulting ischaemia (restriction in blood supply) and ensuing oxygen shortage, if left untreated for a sufficient period of time, can cause damage or death (infarction) of heart muscle tissue (myocardium), as can be observed in Figures 1.6 and 1.7. If impaired blood flow to the heart lasts long enough, it triggers a process called the ischaemic cascade; the heart cells in the territory of the occluded coronary artery die (chiefly through necrosis) and do not grow back.

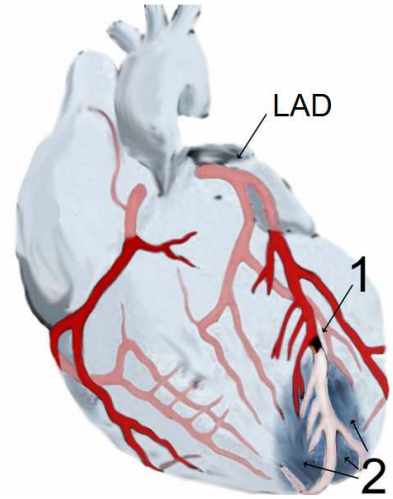


Figure 1.5: Diagram of an Acute Myocardial Infarction (2) of the apex of the anterior wall of the heart (an apical infarct) after occlusion (1) of a branch of the Left Anterior Descendent (LAD) Coronary Artery.

Indeed, the most common triggering event for AMI is the disruption of an atherosclerotic plaque in an epicardial coronary artery, which leads to a clotting cascade, sometimes resulting in total occlusion of the artery. Blood stream column irregularities visible on angiography reflect artery lumen

narrowing (as shown in Figures 1.5 - (1) and 1.6) as a result of decades of advancing atherosclerosis. Atherosclerosis is the gradual buildup of cholesterol and fibrous tissue in plaques in the wall of arteries (in this case, the coronary arteries), typically over decades. Plaques can become unstable, rupture, and additionally promote a thrombus (blood clot) that occludes the artery; this can occur in minutes. When a severe enough plaque rupture occurs in the coronary vasculature, it leads to Acute Myocardial Infarction (necrosis of downstream myocardium). As a result, the patient's heart will be permanently damaged.

Injured heart tissue conducts electrical impulses more slowly than normal heart tissue. The difference in conduction velocity between injured and uninjured tissue can trigger re-entry or a feedback loop that is believed to be the cause of many lethal arrhythmias (for example, ventricular fibrillation, ventricular tachycardia). Cardiac output and blood pressure may fall to dangerous levels, which can lead to further coronary ischemia and extension of the infarction.

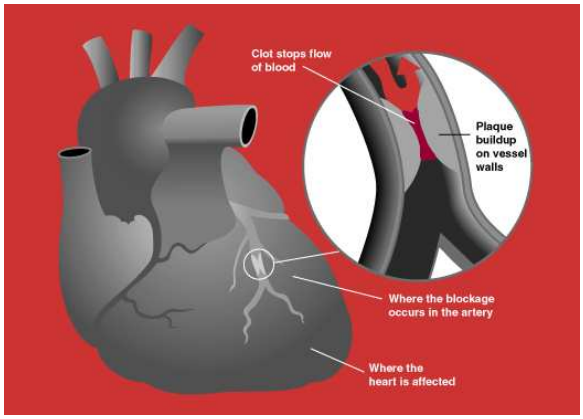


Figure 1.6: An Acute Myocardial Infarction occurs when an atherosclerotic plaque slowly builds up in the inner lining of a coronary artery and then suddenly ruptures, causing catastrophic thrombus formation, totally occluding the artery and preventing blood flow downstream.

Figure 1.7: Drawing of the heart showing anterior left ventricle wall infarction.

The diagnosis of Acute Myocardial Infarction can be made after assessing patient's complaints and physical status. Among the diagnostic tests available to detect heart muscle damage are electrocardiogram (ECG), echocardiography, cardiac MRI and various blood tests. The most often used blood markers are the Creatine Kinase-MB (CK-MB) fraction and the troponin levels.

The World Health Organization (WHO) criteria formulated in 1979 have classically been used to diagnose AMI. A patient is diagnosed with Acute Myocardial Infarction if two (probable) or three (definite) of the following criteria are satisfied: (i) Clinical history of ischaemic type and chest pain lasting for more than 20 minutes; (ii) Changes in serial ECG tracings; (iii) Rise and fall of serum cardiac biomarkers such as CK-MB fraction and troponin. The WHO criteria were refined in 2000 to give more prominence to cardiac biomarkers. According to the new guidelines, a cardiac troponin rise accompanied by either typical symptoms, pathological ECG traces, ST elevation or depression or coronary intervention are diagnostic of AMI.

In general, five main types of AMI can be identified, according to the causes originating them (see the 2007 consensus document [210] for further details):

- Type 1 - Spontaneous Acute Myocardial Infarction related to ischaemia due to primary coronary event such as plaque erosion and/or rupture, fissuring or dissection;
- Type 2 - Acute Myocardial Infarction secondary to ischaemia due to either increased oxygen demand or decreased supply, for example: coronary artery spasm, coronary embolism, anaemia, arrhythmias, hypertension, or hypotension;
- Type 3 - Sudden unexpected cardiac death, including cardiac arrest, often with symptoms suggestive of myocardial ischaemia, accompanied by presumably new ST segment elevation, or new Left Bundle Branch Block (LBBB), or evidence of fresh thrombus in a coronary artery by angiography and/or at autopsy, but death occurring before blood samples could be obtained, or at a time before the appearance of cardiac biomarkers in the blood;
- Type 4 - Associated with coronary angioplasty or stents: *Type 4a* - Acute Myocardial Infarction associated with Percutaneous Coronary Intervention (PCI); *Type 4b* - Acute Myocardial Infarction associated with stent thrombosis as documented by angiography or at autopsy;
- Type 5 - Acute Myocardial Infarction associated with Coronary Artery Bypass Graft (CABG).

Classical symptoms of Acute Myocardial Infarction include sudden chest pain (typically radiating to the left arm or left side of the neck), shortness of breath, nausea, vomiting, palpitations, sweating, and anxiety (often described as a sense of impending doom). Women may experience fewer typical symptoms than men. Approximately one quarter of all Acute Myocardial Infarctions are "silent", that is without chest pain or other symptoms. Important risk factors are previous cardiovascular events, older age, tobacco smoking, high blood levels of certain lipids (triglycerides, low-density lipoprotein) and low levels of High Density Lipoprotein (HDL), diabetes, high blood pressure, obesity, Chronic Kidney Disease (CKD), heart failure, excessive alcohol consumption and chronic high stress levels. To be stressed is that an AMI is a medical emergency which requires immediate medical attention. Treatment attempts to salvage as much myocardium as possible and to prevent further complications, thus the common phrase "time is muscle".

Focusing on STEMI, the most part of cases are treated with thrombolysis or Percutaneous Coronary Intervention (PCI). The former consists in a pharmacological treatment which causes a breakdown of the blood clots, whereas in the latter a balloon driven by a wire, called catheter, is inserted into

the narrowed or obstructed vessels and then inflated to a fixed size. The balloon crushes the fatty deposit, so that the vessel can be opened up, the blood flow improved, and finally the balloon is deflated and withdrawn. Where the culprit lesion occurred, once the stenotic plaque is removed, a medicated stent (i.e. a metal biocompatible device) is located. NON-STEMI should be managed with medication, although PCI is often performed during hospital admission.

In what follows, we will be concerned mainly with Acute Coronary Syndromes in general and their correlated issue, but we will focus on STEMI in particular. Specifically, the next chapter will present organizational and informatics setting adopted by Regione Lombardia for prompting an optimal pattern of care for patients affected by this disease. Moreover, in Part III, the statistical analyses carried out on data coming from STEMI Archive (see Section 3.2) and all other data on STEMI patients (see Section 3.1 and Paragraph 3.2.4) will be presented.

Chapter 2

Pre-hospital organization and systems of care

In this chapter the project funding our research activity is described. It is called *Strategic Program for Acute Coronary Syndromes* (briefly Strategic Program - SP) and among its aims there is the establishment of an effective cardiological emergency network among the hospitals of the Regional district, in the sense explained in the following sections.

2.1 The idea of the *Cardiological Network*

During the last years a growing attention has been reserved by Regione Lombardia to the development of a network of medical institutions connected by an efficient emergency medical service. A hospitals network is a network or group of providers that work together to deliver a broad spectrum of services to their community. Of course a central coordination of the activity of the Emergency Room (ER) of these hospital is needed, so that an efficient management of patients can be reached. The main features of a successful network include a clear definition of the geographical areas of interest, reduction of delays, and close cooperation amongst care givers and institutions.

Focusing on cardiovascular topics, the main aims of a Cardiological Network are to drive the development and to facilitate the delivery of more patient-centred, sustainable and effective clinical services for all patients affected by cardiovascular syndromes. It endeavours to forge effective and productive working relationships between a broader range of people and organizations, in order to promote a more direct and holistic focus on priorities for patients and carers. Concerning cardiovascular diseases, the sooner the interventions are, the higher the probability of good prognoses, so cardiological networks of care must provide optimal services to the patients as quickly as possible. This issue never ends, since emergency networks ask for continue improvements as well as constant monitoring of performances during the time, in order not to lay down on reached goals. A network for acute cardiovascular syndromes management should be developed at a national/regional level, with continuous outcome and quality metrics, to ensure that the reperfusion strategies will continue to be effective after their implementation [202].

The need of a “network structured” emergency within cardiovascular healthcare delivery comes out of the fact that over recent years, there has been a growing complexity of cardiovascular patient conditions, due also to the ageing of affected population. So a multidisciplinary and delocalized management of cardiovascular patients is requested, as can be evinced by the picture reported in Figure 2.1. For the multidisciplinary management of cardiological patients in emergency setting, an

integrated organization of assistance plans is mandatory. This must be realized through the development of a network efficient and complementary services among providers, irrespective of their location. In such a model, the attention is shifted from single health care service to the whole welfare path, in order to make it integrated and homogeneous, irrespective of where single treatments are delivered.

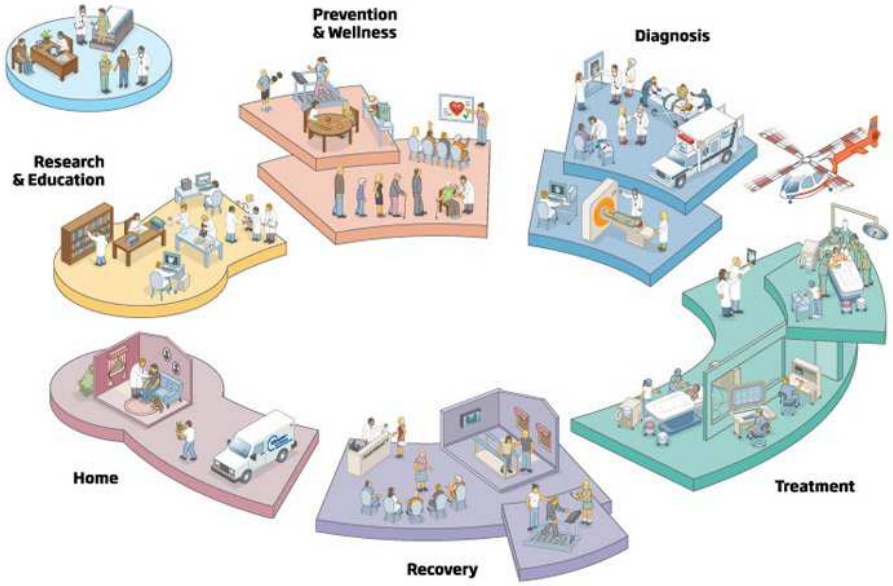


Figure 2.1: All steps of the integrated process of care provided by an efficient network.

The integrated network for emergency cardiovascular delivery classifies providers according to increasing complexity of the services they can offer. The Italian model for pre-hospital patients management is the "Hub-Spoke" model, where hubs are reference centres where advance rescue resources are concentrated and always available, supported by the spoke system of services, i.e., centres non-PCI capable that must provide primary treatments but then transfer patients to the nearest or most suitable hub. For further details on definition of hub and spoke characteristics see [195]. In fact, a broad evidence exists (see [151], [156], [157], [158], [160], [182], [189], [197], [211] and [213]) to testify the benefit of early intervention in the treatment of STEMI. Networks are based on the idea that patients, like those affected by STEMI or Acute Coronary Syndromes in general, should be:

- rescued by advanced units of 118 Dispatch Centre (the National free toll number for emergencies), which are equipped by tele ECG machinery and defibrillators;
- delivered in a *hub* (i.e., a hospital with 24 hours operating coronary care unit, 7 days per week). If patient are self presented or delivered by a basic rescue unit in a *spoke* (i.e., a hospital without always operating coronary care unit), they should be transferred in a hub following local specific protocols that take care of total ischaemic time, of possibility of performing a primary angioplasty within 90-120 minutes and so on.

In literature, the reference guidelines for the definition of such terms are those proposed by the American Heart Association (AHA) and the American College of Cardiology (ACC) (see [188]). Many examples suggest that the presence of a net connecting the territory to the hospitals, by a

centralized coordination of the emergency resources, helps to promote and realize the best utilization of the different reperfusion strategies, reducing transport and decisional delays connected with logistics and therapies, and increasing the number of patients undergoing primary PCI before 90 minutes since the arrival at ER (see for example [151], [156] and [214]).

In Italy, the reference document for the Cardiological Network structuring is [169], where the model for the cardiological network is established according to criteria approved by the main Italian Cardiological societies, namely ANMCO (Associazione Nazionale Medici Cardiologi Ospedalieri - `http://www.anmco.it/`), SIC (Società Italiana di Cardiologia - `http://www.sicardiologia.it/`) and GISE (Società Italiana di Cardiologia Invasiva - `http://www.gise.it/`), by the emergency dispatch of 118 and by people in charge with cardiovascular healthcare of each Region. In this document [169], some issues are set, like who are the main players of the network (physicians, healthcare district and government, hospitals, 118, etc.), who have to manage and administrate it, which are the basic structural standards to be satisfied for hemodynamics and cath-labs, the role of ERs and especially the main goals of the care process. Concerning patients affected by Acute Coronary Syndromes, the main issue to be accomplished by the cardiological network are:

- to increase the number of patients arriving alive to the hospital;
- to increase the number of patients undergone any reperfusive treatment;
- to manage appropriately in the UTIC (Unità di Terapia Intensiva Cardiologica) the patients arriving at ER with Acute Coronary Syndromes;
- to start the reperfusive treatment as soon as possible;
- to make reperfusive treatments available for all patients who need them, regardless of the place where the diagnosis is established;
- to ensure the intervention to high risk patients.

The pre-hospital phase is the most critical especially in STEMI patients' management, because the myocardial savage and the number of lives saved is inversely proportional to the delay in treatment. Implementation of STEMI systems of care has a pivotal role in modern STEMI treatment: they are based on the network among medical cardiology institutions, connected by an effective emergency medical service. Since a STEMI can happen anywhere and anytime, and since very rapid diagnosis and treatment are mandatory, networks have a key role in providing an equitable access to the most effective care to the vast majority of STEMI patients.

In Regione Lombardia, the coordination of a local network is established on a territorial base that corresponds to the 118 Dispatch Centre one. All Dispatch Centres depend on *Azienda Regionale Emergenza Urgenza* (AREU), which coordinates rules and actions of each Dispatch Centre and manage the network of ERs where 118 rescue units deliver patients. The guidelines for the establishment of a Cardiological Network for STEMI are contained in the law called *Decreto Regionale N° 10446 - "Determinazioni in merito alla rete per il trattamento dei pazienti con Infarto Miocardico Acuto con tratto ST elevato (STEMI)"* [36]. Anyway, as we said before, the issue of continue improvement of network efficiency never ends, since emergency networks ask for continue and constant monitoring along time. In this sense, clinical registries focused on STEMI could enable comparisons among strategies adopted in different countries for managing pre-hospital phase of STEMI, may lead changes in priorities and disease management and provide information for

monitoring and evaluating the performances of the network. Then, starting from the advantages carried on by the presence of an efficient network, it is possible to integrate it with the knowledge that comes out from clinical surveys and administrative data, in order to take a snapshot of the network activity and to evaluate its effects on high level output. The synergy between these aspects (i.e., the advantages carried on by the Cardiological Network and the monitoring of performances allowed by the use of clinical surveys and administrative data) is the main innovative purpose of the Strategic Program, the scientific enterprise that Regione Lombardia embarked on in the last years and that we have been involved in.

2.2 The Strategic Program of Regione Lombardia

The appropriate treatment of STEMI in a timely manner is instrumental in mortality reduction. In the previous section we saw how systems of care based on networks of medical institutions connected by an efficient emergency medical service are pivotal. To make these networks to work at best, a continuous monitoring and evaluating process is absolutely mandatory. Clinical registries of pathologies are the best candidates to reach this goal in a cost saving and effective way. In this Section the Strategic Program (SP) of Regione Lombardia [35] is presented. This project started in 2008, funded by Regione Lombardia, the Italian Ministry of Health and by the Regional district for healthcare, namely the "Direzione Generale Sanità - Regione Lombardia", aimed at stating the programmatic lines for development of a cardiovascular healthcare management based on criteria met in the previous sections.

The SP is structured into five sub-projects, two focused on health organization (P1 - Regione Lombardia and P4 - Regione Emilia Romagna), and three (P2 - IRCCS Centro cardiologico Monzino, P3 - IRCCS Istituto Auxologico Italiano and P5 - IRCCS Policlinico San Donato) concerned with new biomolecular and imaging strategies aimed at the identification of patients with ACS at the highest risk. Specifically, P1 and P4 are aimed at checking feasibility at regional level of

P1 a clinical registry of Acute Coronary Syndromes (ACS) - Regione Lombardia, Direzione Generale Sanità, Scientific Director Dr.Maurizio Marzegalli;

P4 a clinical registry of adverse events occurring after coronary angioplasty where stent with or without medication are employed - Regione Emilia-Romagna, Scientific Director Dr. Antonio Marzocchi.

On the other hand, projects P2, P3 and P5, focus on the identification of high-risk patient profiles in terms of:

P2 thrombotic complications due to reduced renal function and prothrombotic activation of blood elements - IRCCS Centro Cardiologico Monzino, Scientific Director Dr. Gian Carlo Silvio Marenzi;

P3 ventricular fibrillation and sudden death after Acute Myocardial Infarction, due to the prevalence of genetic polymorphisms relevant for the disease - Istituto Auxologico Italiano, Scientific Director Dr. Peter J Schwartz;

P5 poor revascularisation consequent to reduced ventricular viability or function as identified with new imaging techniques, also in relation to newly identified biomarkers - IRCCS Policlinico S. Donato, Scientific Director Dr. Lorenzo Menicanti.

These projects will provide information to build up an integrated model for ACS, which will include new organisational, preventive and therapeutical strategies, which will take specific account of patients' risk. The Coordinator of the SP, Dr. Maurizio Marzegalli, is the chair of Project P1. He met every year the Steering Committee composed by the scientific coordinators of other Projects. A Technical Committee, composed by one representative of each Unit, had meetings every 6 months to assess the accomplishment of deliverables and referred each time to the Steering Committee.

2.2.1 P1 - Part of Strategic Program of Regione Lombardia

We will focus now on the project P1 (we will refer to it in the following as SP-P1), whose title is "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction", and where our scientific activity moved into during these three years. Three are the sub-units whose work refers to it: UO 01 - Regione Lombardia, Scientific Director Dr. Luca Merlino; UO 02 - Politecnico Milano, Scientific Director prof. Piercesare Secchi; UO 03 - Università dell'Insubria, Scientific Director prof. Marco Ferrario.

The major goal of the SP-P1 is the identification of new diagnostic, therapeutic and organizational strategies to be applied to patients affected by Acute Coronary Syndromes, in order to increase the occurrence of positive clinical outcomes. The secondary objectives can be summarized in the work of each participating Unit, i.e., the management of regional database (UO 01), the integration of databases, the development of suitable new ones and the statistical analysis of the resulting data (UO 02), and the epidemiological study of the impact of STEMI on a specific territory (UO 03). In the scientific literature in the medical field it is possible to find a lot of papers about the usefulness of registries and administrative databases as evaluation instrument of the reality beside the more traditional clinical trials. The relationship between clinical trials, guidelines and registries has been sum up very well by Lukas Kappenberger *"Science tells us what we can do; guidelines what we should do; and registries what we are actually doing"*. In the last ten years, many registries have been constructed and focused on ACS, with particular attention to the Acute Myocardial Infarction with or without ST-segment elevation (see, among others, [152], [163], [165], [190], [192], [193] and [219]). On the initiative of scientific societies many databases have been collected in Lombardia about the treatment of AMI (for example: GESTIMA, LOMBARDIMA, MOMI^2, etc., for details see Section 3.2). These pilot studies have proved the importance of collected information for the analysis and improvement of clinical activity and medical planning, completing the knowledge of the number of cases, the incidence of diseases/complications, the time and performance of treatments, the short term and long term survival, etc.

The purpose of the SP-P1 project is the study the treatment of the ACS in Regione Lombardia, with particular attention to the patients affected by AMI. The aim is to exploit and integrate the epidemiological knowledge with the already collected patient's data: the Regione Lombardia and scientific communities gave us the permission to extract and study these data from the administrative regional datawarehouse and clinical databases. The actual limit in the knowledge is due to the fact that no statistical analyses have never been carried out on the the already existing data, or that these data let us have only a partial and spot picture of the reality. The new perspective pointed out by SP-P1 concerning the share of already existing information, their integration with suitable new records and a whole analysis of data should produce a general picture of the real clinical world in Lombardia

and drive the planned activities both in analysis and in improvement of clinical performance and scheduling of regional activities.

The main objective of the project is then to give a global picture of data in epidemiology and clinical treatment of Acute Myocardial Infarction. Quantitative analyses are necessary for an adequate scheduling of medical activities in order to intercept the new needs of the health, to verify the effectiveness of innovations and to measure the epidemiological outcomes. Traditionally, clinical trials are used in order to give scientific evidence of efficacy of new drugs, treatments, technologies and procedures respect to traditional ones. Nevertheless, often, even studies characterized by high scientific impact factor do not give a guarantee to us that it is possible to generalize results from particular experimental conditions and sample sizes. So the observational clinical studies become crucial, in fact they let us measure the effectiveness of procedures and interventions, since they describe the real clinical action on the treated population. The existence of the "Piano Cardio Cerebro Vascolare" in Regione Lombardia has been the stimulus to the integration of different clinical databases in order to help medical planning. Moreover it has been also possible to integrate these particular databases with the regional datawarehouse. In fact many databases concerning ACS are already available and the principal object of this project is to find the way such that, with a little supplementary amount of cost and work, the integration of all these information can be made immediately available and can also be analyzed in a relatively short time, so that an audit and a supervision of these data can be carried out without wide efforts. Such a work enables each clinical reality to detect those parameters upon which it can act in order to improve its own performance; moreover it let Regione Lombardia to develop, strengthen and extend to the whole territory the network model for ACS, so that it is possible to check and verify if network works well, as suggested by the rules in the consensus document: "La rete interospedaliera per l'emergenza coronarica" [169]. Moreover a complete monitoring let the patient to be guaranteed with a better relief continuity between pre-hospital phase, hospital and home care.

This study represents a new form of collaboration among scientific societies and institutional decision makers, in order to attain a uniformity in high quality treatment on the whole regional area. The law [36] contains instructions for the setting of a Regional Archive for STEMI as an instrument to improve the efficacy of the cardiological network as well as to monitor efficacy and efficiency of single provider, so that changes in operating policies can be implemented starting from a real time and ongoing instrument of control. It is fundamental that this new methodology (see Figure 2.2) in collecting and analyzing data, measuring indicators, giving a feedback to institutions, going throw a synergic integration of different systems, will be in the future a standard method in other cardiological pathologies and other medical topics to carry out evaluations and decision about healthcare, as it is in Europe. This research project will involve mainly public hospitals, so the transferability to all the National Health Service would be an easy goal. In particular we think that the project will improve the share of information coming from registries devoted to the epidemiological studies and the diffusion of a new culture about the quality of collected data, unfortunately now not very deep in different areas of Italian National Health Service.
Specifically concerning the UO 02 - Politecnico di Milano, the contribution of this unit is to give a methodological support to the analyses of the collected data, in order to improve health care quality and clinical performances in the treatment of patients affected by cardiovascular diseases, with particular attention to Acute Myocardial Infarction with or without ST-segment elevation. This support activity consists of different topics:

- the systematic study of the literature about determination of optimal treatment approach to clinical practice, guidelines and quality indicators for the clinical performances;

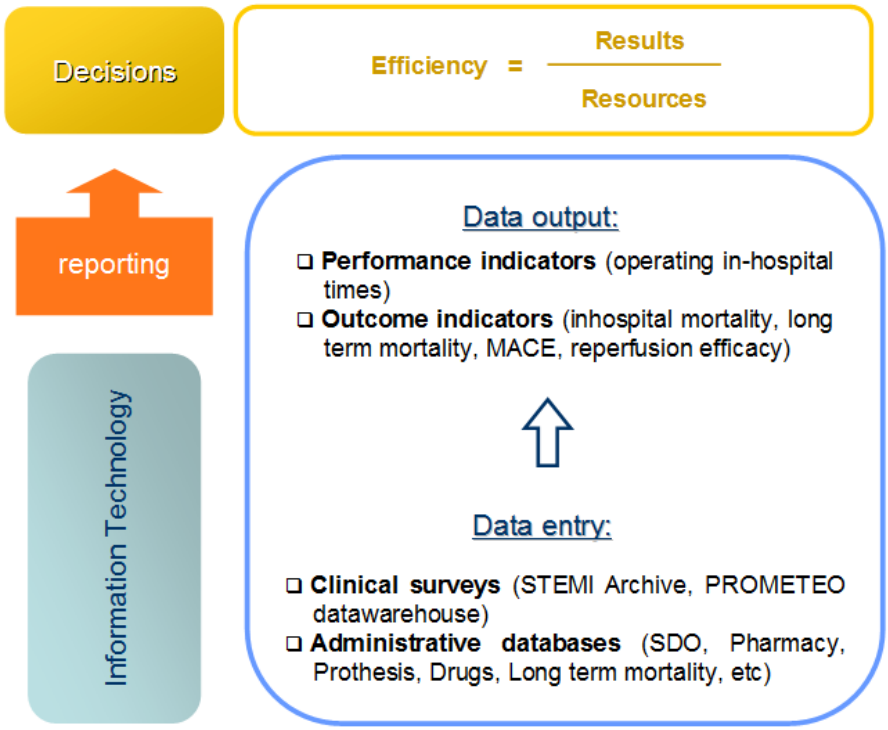


Figure 2.2: Workflow for measuring performances and outcomes and taking decisions in cardiovascular healthcare process.

- the identification of primary and secondary endpoints that must be investigated in order to catch a realistic picture of the treatment of cardiovascular patients in all areas of the Regione Lombardia;
- the identification of quality indicators for clinical performances, specific for the context where the Cardiological Network operates;
- the identification of useful data that must be collected in the patients clinical history;
- the support activity to other participating units, in what concerns the extraction of suitable records from the already existing databases;
- the management and quality control of data as well as the construction of suitable statistical models to analyse them;
- the project management, coordination of the different groups working in this project and management of the website of the project;
- the comprehension of knowledge/information gap and identification of organizational barriers and related costs;
- the description and communication of the results to the scientific community.

All this issues will be addressed in the analyses presented in Chapter 7.

2.2.2 Cardiovascular process indicators: time to treatment and time to intervenction.

We said that systems for STEMI care based on networks of medical institutions connected by an emergency medical service are pivotal for providing good performances in the treatment and management of STEMI patients. Concerning ACS in general, efficiency is a matter of time, the shorter the better. There are several ingredients that could combine to bring about delays in such a complex process like the one from symptoms onset to treatment. In order to reduce these delays as much as possible, different strategies can be adopted, depending on aims and scopes that are being pursued. Firstly it is necessary to minimize the patient's delay in seeking care, then to dispatch properly staffed and equipped rescue units to make diagnosis on scene, to deliver suitable drug therapies or surgical treatments, and to transport the patient in the most appropriate (not necessarily the closest) cardiac facility. Strong cooperation between cardiologists and emergency medicine doctors is mandatory for optimal pre-hospital STEMI care. Scientific societies have an important role in guideline implementation as well as in developing quality indicators and performance measures; health care professionals must overcome existing barriers to optimal care together with political and administrative decision makers.

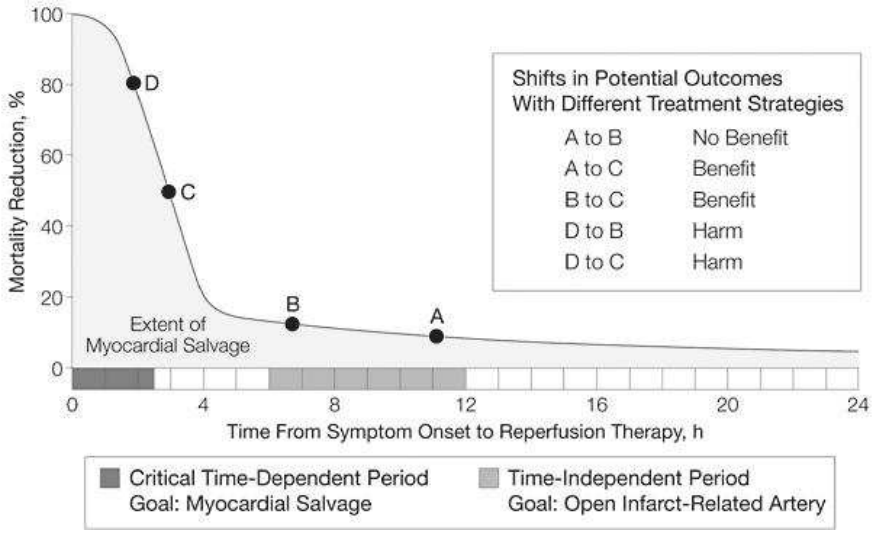


Figure 2.3: Extent of Myocardial Salvage along time. *Gersh (2005), JAMA **293**, 979–986.*

In order to properly evaluate the efficiency of the delivered service, it is necessary to define suitable process indicators. Since, as we said before, efficiency in STEMI care is a matter of time (Figure 2.3 quantifies the impact of time on capability of rescueing tissues as detailed in [172]), this forces to define proper time indicators. They must deal both with pre-hospital and in-hospital times (shown in Figures 2.4 and 2.6), which are defined according to the issues in [36] as:

- **Onset**: time of symptoms onset;
- **Call**: time of patient's call for 118 rescue;
- **First Medical Contact (FMC)**: time of first electrocardiogram (ECG) that allows for STEMI diagnosis, irrespectively of the setting and of the presence of a physician. It consists of the time of first pre-hospital ECG carried out by Advanced Rescue Units or by Basic Rescue

Units (the common ambulances) for patients who call 118, whereas consits of first ECG time in ER for self-presented patients;

- **Door**: time of admission to ER/emergency department;
- **ECG**: time of first electrocardiographic diagnosis;
- **Needle**: time of pharmacological treatment (if any);
- **Balloon**: time of inflaction of the balloon of the catheter in PCI treatment.

Figure 2.4: Pre-hospital times in STEMI.

The definitions proposed above define time intervals (see figure 2.5) to be evaluated for assessing care and Emergency Medical System (EMS) efficiency. The possible total delay is then composed by delays stored at each interval and, according to this, different actions can be implemented in order to reduce them. It is proved that longer delays to reperfusion are associated with higher mortality and that the implementation of regional systems of care for STEMI can increase the rate of reperfusion and reduce long term mortality (see [200]). Different logistics and patients' ways of admission to ER can modify the sequence of these times (see Figure 2.6), anyway the following process indicators can be defined:

- **Onset to Balloon (OB)**: total ischaemic time for patients undergone PCI;
- **Onset to first ECG (OfECG)**: time from symptoms onset and first STEMI diagnosis;
- **Onset to Door (OD)**: time from symptoms onset and admission at ER or at a triage in an emergency department;
- **Door to Needle (DN)**: time from admission at ER or at a triage in an emergency department and pharmacological treatment;
- **Door to Balloon (DB)**: time from admission at ER or at a triage in an emergency department and inflaction of the balloon of the catheter in PCI treatment;
- **first ECG to Balloon (EB)**: time from first STEMI diagnosis and inflaction of the balloon of the catheter in PCI treatment.

It is clear from Figure 2.5 that the optimal management of STEMI is possible only through an emergency systems based on a pre-hospital phase characterized by diagnostic and therapeutic capacities, a good level of networking and an efficient delivery system.

Pre-hospital care is of outstanding importance for patients' outcome. Decisions in the pre-hospital setting are pivotal in STEMI care, as delays can not be compensated later on. Systems of care need to address not only delays from first medical contact to treatment, but also the total delay from

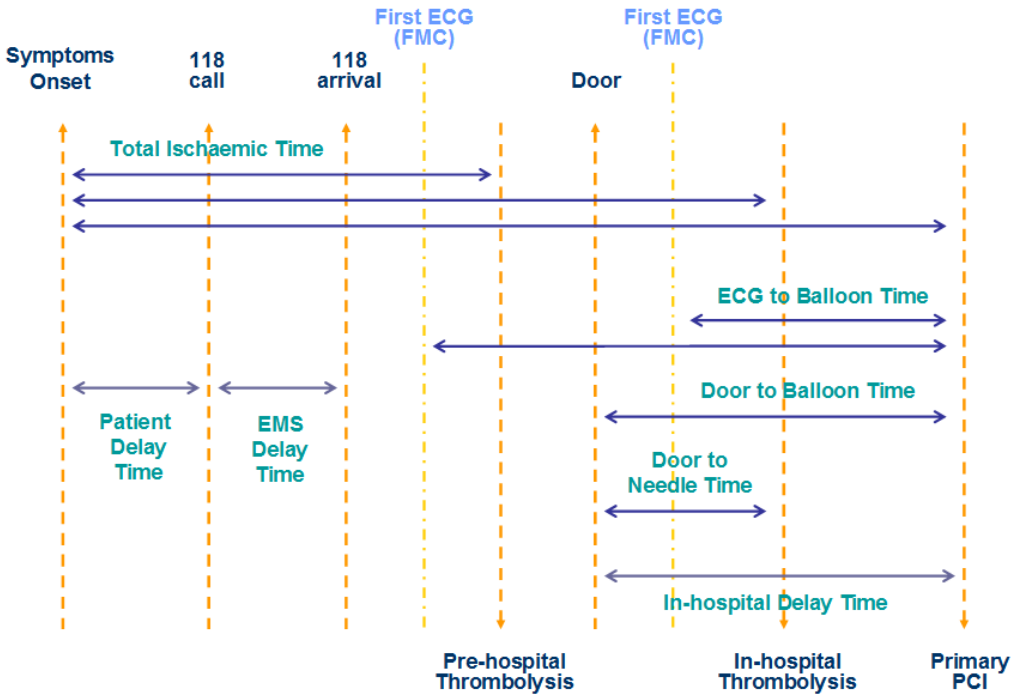


Figure 2.5: Components of total ischaemic time in STEMI.

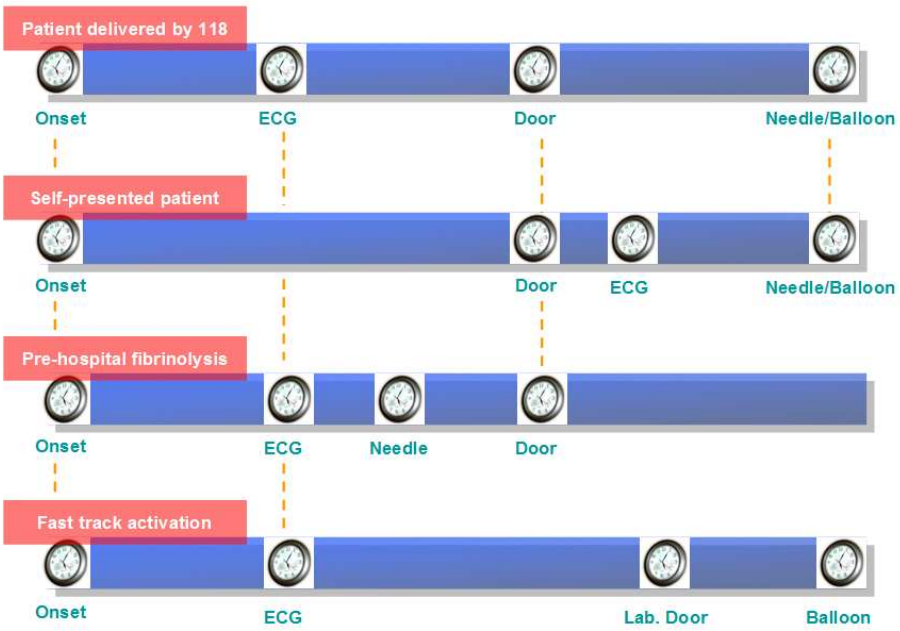


Figure 2.6: Time sequences stratified according to different ways of admission and/or pattern of care.

symptom onset to reperfusion (i.e., the total ischaemic time). The patients' decision time is usually a critical period. An early first call is desirable, since it allows for a rapid diagnosis that prevents complications, but often not realized. Emergency medical services vary in their approach to re-

ceiving and prioritising emergency calls. The ideal Emergency Medical Dispatcher triages should rest on the availability of advanced rescue units, being both rescue units with physicians on board and/or ambulances staffed by paramedics working with agreed protocols supplemented by physicians' direction. The use of pre-hospital ECG has already been shown to be able to reduce time of reperfusion therapy (as shown, for example, in [180], [198] and [204]) and possibly to decrease mortality (as sustained, for example, in [164]). Performing a high-quality diagnostic ECG is a specific process of care, requiring education, training and maintenance of competency for emergency medical service providers. A focused study and analysis of data arising from this activity is the topic of Chapters 6, and Sections 9.1 and 9.2 respectively.

Chapter 3

Data sources

3.1 The administrative datawarehouse of Regione Lombardia

In this section we describe structure, aim and use of the Regione Lombardia Public Health Database (PHD), the datawarehouse the STEMI Archive has been designed to be integrated with, as well as all the clinical surveys we will consider for the analyses presented in Part III.

3.1.1 The star scheme: an overview of complexity

Administrative health care databases play today a central role in epidemiological evaluation of Lombardia healthcare system because of their widespread diffusion and low cost of information. Public health care regulatory organizations can assist decision makers in providing information based on available Electronic Health Records, promoting the development and the implementation of methodological tools suitable for the analysis of administrative databases and answering questions oriented to disease management. The aim of this kind of evaluation is to estimate adherence to best practice (in the setting of evidence based medicine) and potential benefits and harms of specific health policies. Health care databases can be analysed in order to calculate measures of quality of care (quality indicators); moreover the implementation of disease and intervention registries based on administrative databases could enable decision makers to monitor the diffusion of new procedures or the effects of health policy interventions.

Health information systems in Lombardia experienced a rapid growth as a consequence of the introduction in the Italian health management of Diagnosis Related Groups (DRGs) in 1995. The development of health care measures for the specific aim of health system financing, gave rise to the availability of information useful for evaluating the efficiency of the providers and the efficacy of their activities. The development of health information systems was particularly pronounced in hospitals, and this extended the possibilities of measuring their activities: from the "classic" indicators (average length of stay, occupancy rate, turnover interval), measuring bare hospitality, to more meaningful evaluations linked to patient classification systems and to the actual opportunity of calculating quality indicators. Several regional and national rules introduced in recent years a large number of indicators in the Italian national health system.

The Regione Lombardia Public Health Database (PHD), called "BDA" (*Banca Dati Assistito*), contains a huge amount of data and requires specific and advanced tools and structures for data mining and data analysis. This is an on going datawarehouse, which up to now has been used only for administrative purposes, since decision makers of healthcare organizations need information about efficacy and costs of health services. The structure adopted by Regione Lombardia is

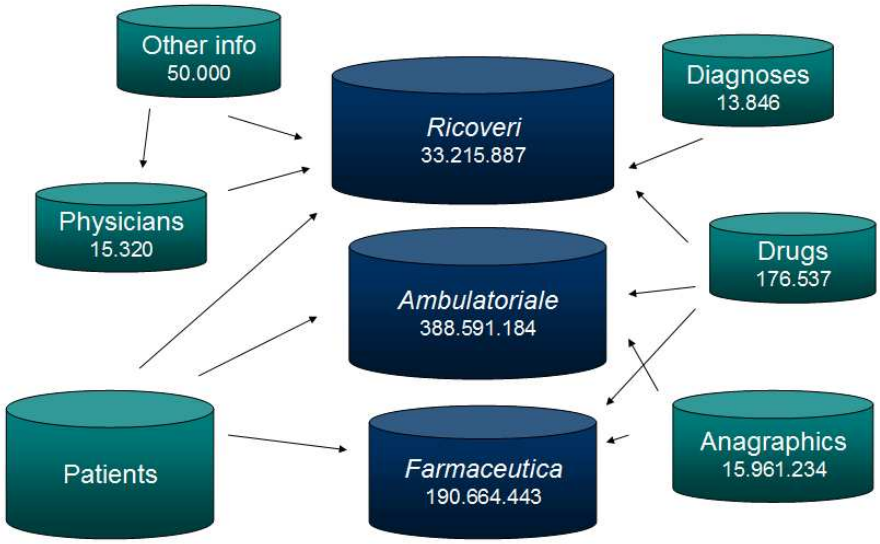


Figure 3.1: Star Scheme structure of Public Health Database (PHD) of Regione Lombardia. Numbers are referred to a two-years time period between 2003 and 2005

called Star scheme (see [86] and Figure 3.1), since it is centered on three main databases (*Ambulatoriale*, *Farmaceutica*, *Ricoveri*) - containing informations about visits, drugs, hospitalizations, surgical procedures that took place in hospitals in Lombardia - while being supported by secondary databases (*Assistibili*, *Medici*, *Strutture e Farmacie*, *Farmaci*, *Codici Diagnosi* e *Procedure Chirurgiche*) which contain specific information about procedures coding or personal data about people involved in the care process. The star scheme does not allow for repetitions in records entering. Considering, for example, the database *Ricoveri*, only one record for each episod of disease is allowed and each record finishes with patient discharge. In this case, we call *Event* the total amount of admissions and discharges related to the same episode of disease.

Inside the PHD, several records may correspond to the same patient over time: for example, a patient may have several events during years, and each event could consist of multiple admissions. As we said before, for each admission/discharge path, one record is produced in PHD. Records related to the same subject may be linked in order to achieve the correct information about the basic observation unit. However each of the above databases has its own dimension and structure, and data are different and differently recorded from one database to another. Suitable techniques are therefore required to make information coming from different databases uniform and to carry out data mining and data analysis. In particular, the longitudinal data that we will analyse will be generated by deterministic record linkage between STEMI Archive and the databases *Ambulatoriale*, *Farmaceutica* and *Ricoveri* of the PHD, as described in Section 3.3. Regione Lombardia, as data manager and owner, provides an encrypted code for each patient in order to protect citizen's privacy. This encrypted code represents the key to obtain the deterministic linkage between the databases.

3.1.2 Data mining of administrative databanks

Dealing with the huge amount of data the PHD contains is a very challenging scientific and mining enterprise. In general, over recent years there has been an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases, as

testified in [153], [154], [168], [176], [186] and [218]; this motivated Regione Lombardia to use its own administrative database for clinical and epidemiological aims.

Administrative healthcare databases can be analysed in order to calculate measures of quality of care (quality indicators). The importance of such databases for clinical purposes depends on the fact that they provide all the relevant information that decision makers need to know, in order to evaluate the implications of particular policies affecting medical therapies (for example, information about applicability of a trial findings to the settings and patients of interest, effectiveness and diffusion of new surgical techniques, estimation of adherence to best practice and potential benefits/harms of specific health policies, etc.). Moreover, administrative healthcare databases play today a central role in epidemiological evaluation of healthcare systems because of their widespread diffusion and low cost of information.

The most critical issue when using administrative databases for observational studies is represented by the selection criteria of the observation units: several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. Statistical analysis can be performed by means of multiple logistic regression models for studying outcomes and by means of survival analysis when studying failure times (hospital readmissions, continuity of drug prescriptions, survival times). Multilevel models (see Part II) can also be adopted if structural and organizational variables are measured. When outcomes are the main focus of the observational study, appropriate risk adjustment tools are needed. Hospital discharge records may be analysed with the indicators developed by the Agency for Health care Research and Quality (AHRQ) that include efficient risk adjustment tools within a multiple logistic regression model. In disease management programs the Johns Hopkins Adjusted Clinical Groups (ACG) methodology and the Classification Reasearch Group (CRG) classification system have been proposed (see [173], [179] and [205]).

3.2 Clinical surveys

Clinical Surveys (also named Clinical Registries of pathology) are databases that systematically collect health-related information on individuals who are:

- treated with a particular surgical procedure, device or drug (i.e., joint replacement surgery);
- diagnosed with a particular illness (i.e., Acute Myocardial Infarction); or
- managed via a specific healthcare resource (i.e., treated in an intensive care unit).

Information in clinical registries is captured on an ongoing basis from a defined population. Clinical registries provide the most suitable and accurate method of providing monitoring and benchmark data and provide the greatest potential to improve healthcare performance across institutions and providers.

In what follows, the cardiological surveys carried out in the last years in Regione Lombardia will be presented, together with other sources of clinical data related to caridovascular sydromes. Then details will be given on STEMI Archive and PROMETEO datawarehouse, since they are the main object of analyses presented in Part III.

3.2.1 Past cardiological surveys of Regione Lombardia

On February 11^{th} 2005, the Piano Cardio-Cerebro Vascolare was introduced in Regione Lombardia [35]. This law sets favourable conditions for using clinical registries in health-care process planning. Regione Lombardia is very sensitive to cardiovascular diseases, as proved by the huge amount of

social and scientific projects concerning these syndromes that have been funded and promoted over recent years. Among others, the following three projects can be mentioned to be the reference experiences the SP drew inspiration from:

→ **GestIMA** - *Gestione dello STEMI in Lombardia*. Bimonthly data collection (Oct - Nov, 2003), 612 patients with STEMI diagnosis were enrolled. See [194].

→ **LombardIMA** - *A regional registry for coronary angioplasty in ST-Elevation Myocardial Infarction*. 3901 STEMI patients underwent PCI procedure within 12 hours of the onset of symptoms. See [196].

→ **Nuove Reti Sanitarie** (2004-2009). Tele-monitoring activities for patients affected by Chronic cardiac insufficiency and those concerned with in-home care after cardiac admittance to hospital.

→ **MOMI^2** - *MOnth MOnitoring Myocardial Infarction in MIlan* (2006-2008). Observational study collecting 841 STEMI patients in 6 monthly/bimonthly collections along 2 years. See [57], [59], [62], [67], [74] and [78].

In the next section we will focus on the last project, detailing results achieved and statistical methods implemented on data arising from this clinical survey on STEMI. It that can be considered a pilot experiment on the "smaller scale" of Milanese urban area of what SP strives to do for Regione Lombardia, since it enabled us to test feasibility of network and PM&E process concerning the issues the STEMI Archive is also focused on.

3.2.2 The MOMI^2 experience on Milan area

A significative preliminar experience of data collection on STEMI patients is represented by the MOMI^2 survey, an observational study hold between 2006 and 2008 where five 30-60 days collections have been performed on the Milanese urban area in order to assess the impact and the efficiency of the cardiological network in treatment of STEMI patients. In the city of Milano the coordination of emergency resources servicing 1.4 million residents plus 1 million of daily commuters was centralized in 2001. It consists in a network connecting hubs that receives in-coming calls, variously equipped ambulances and 23 receiving hospitals, all with cardiology departments and Coronary Care Units, 18 with a 24-hour catheterization laboratory service able to perform primary PCI.

All details on setting, statistical analysis of data and results in terms of healthcare policy and updating of STEMI care patterns can be found mainly in [74], then in [57], [59], [62], [67] and [78]. Anyway, two main aspects arisen from MOMI^2 experience should be highlighted: the first is that the experience of the Milan network for Cardiac Emergency shows how a network coordinating community, rescue units and hospitals in a complex urban area and making use of medical technology contributes to the improvement of healthcare delivery concerning STEMI patients; the second is the seminal idea for PROMETEO project (see Paragraph 3.2.4).

In MOMI^2 survey, most of the 841 patients received a reperfusion therapy (82%), in particular 73% of them underwnt PCI treatment, in-hospital mortality was low (6.3%) and door-to-balloon time was less than 90 minutes in nearly 64% of cases. Moreover, in that context a number of variables related to the outcomes were identified, both non-modifiable and modifiable, i.e., variables we cannot/can act upon. Among the latter ones, for example, pre-hospital and in-hospital times (like total ischemic time, which was inversely related to in-hospital mortality, door-to-balloon time and symptom onset time, which were inversely related to treatment efficacy) were used as quality indicators in the PM&E process.

It is from the results arisen from this pilot study that the idea of PM&E process should have been carried out using time process indicators. In fact, total ischaemic time and symptom onset time depend on the time to first medical contact, time from medical contact to admittance in hospital and door-to-balloon time. Time to first medical contact depends on the time the patient and/or other people take to call the emergency services, which can be shortened by awareness campaigns. Time from medical contact to admittance in hospital and door-to-balloon depend on pre-hospital and hospital logistics respectively. In MOMI^2 study four modifiable variables were found to have an impact on door-to-balloon time: the most statistically significant predictors were mode of arrival and time to first ECG, then fast track organization in hospital and time of arrival (i.e., on hours/off hours). The mode of arrival had a major influence on door-to-balloon time, because the receptivity and hospital response were different in terms of triage and direct transport to the catheterization laboratory for patients whose arrival was expected or was managed by an Advanced Rescue Unit. Worthy of note is that a significant difference was found only when the Advanced Rescue Unit was equipped with 12-lead ECG recorder and transmitter. The difference between Basic and Advanced Supports did not offer any additional advantage. Thus, the crucial factor was ECG transmission to hospital staff, who, after having made the diagnosis of STEMI while the patient was still en route, could alert the cardiology team and catheterization laboratory, which could then be prepared to receive the patient by fast track and perform PCI immediately. This is consistent with the literature results, which have shown that pre-hospital 12-lead ECG reduces door-to-balloon time and mortality; it also suggests that pre-hospital ECG may either be transmitted for interpretation by hospital staff or can be interpreted locally by paramedics who then communicate their diagnosis to the hospital, with an acceptable false positive diagnosis rate. We speculate that the beneficial effects of recording and transmitting a 12-leads ECG in the pre-hospital phase of STEMI may not be restricted to patients treated with PCI; they may extend also to patients treated with thrombolysis by implementng effective pre-hospital management and by reducing the in-hospital delay.

These findings were made within the context of an Emergency Service, in which a hub is connected to rescue units and receiving hospital in real time. Such a network ensures that an ambulance is directed to the nearest appropriate hospital. The efficiency of the organization is reflected by the very low transfer rate.

As to the use of different survey periods to validate the advantages and limits of the Milan network, it is worth pointing out that the analysis of data collected for short periods of time by different observers has already proved to be a reliable and easy implementing method. Moreover, repeated data collections enabled continual updates that fueled debates on logistics designed to optimize the system. A limitation of the study was its observational nature that did not allow any intervention to ensure appropriate management of the patient. Another limitation was its sample size and short observation periods, which did not enable an adequate estimate of the mortality rate of the patients. Anyway, MOMI^2 pointed out significative results in terms of effective management of Milan Cardiological Network (see [84]) and enabled to test the feasibility of performing a PM&E analysis based on clinical registry.

The main result consists of having proved that, in the presence of suspected Acute Myocardial Infarction, the immediate performance of an ECG is essential to document STEMI and alert a catheterization laboratory at the nearest hospital available. In order to achieve this it is essential to equip rescue units with devices able to record and transmit ECG to experienced staff. This significantly shortens door-to-balloon time, which not only was found to contribute to effective reperfusion, but also to be an important factor, as component of total ischaemic time, associated with mortality. These findings, made within the context of an emergency service that efficiently connects a hub to rescue units and receiving hospitals, resulted in a better organization of STEMI patterns of

Figure 3.2: Example of final result of filled sheet of STEMI Archive.

care on Milanese urban area. Moreover, as shown in [62], [66] and [78], advanced statistical analyses carried out on these data within the Strategic Program made it possible to tune suitable process indicators to be used in the monitoring process of performances of the whole regional network of cardiologies.

3.2.3 The STEMI Archive

All the previous experiences described in this chapter concurred in the definition of the new clinical registry on STEMI pointed out within the Strategic Program, i.e., the STEMI Archive. STEMI Archive enlarges the MOMI^2 paradigm to the whole territory of Regione Lombardia, standardizing data collection systems (the same standard for all hospitals involved based on Electronic Health Records provided by SISS). It is also presented as a candidate for becoming the performance evaluation instrument for Regione Lombardia cardiovascular policy.

As presented in [83] and shown in Figure 3.2, STEMI Archive is a multicenter observational clinical registry planned within the Strategic Program (see Section 2.2). This is an observational clinical registry that collects clinical indicators, process indicators and outcomes concerning STEMI patients admitted to any hospital of the Regional district, one of the most advanced and intensive-care area in Italy. This registry is arranged to be automatically linked to the PHD presented in Section 3.1. Its main goal is to enhance the integration of different sources of health information in order to automate and streamline clinicians' workflow, so that data collected once can be used multiple times for different aims, and especially for measuring performances of healthcare system, to understand how hospitals work and to increase efficacy of healthcare offer in terms of costs and patterns of care. In fact, integrated systems enable people in charge with healthcare government to obtain data for billing or performance evaluations, as well as they allow clinicians to see trends in the effectiveness of treatments or to compare patterns of care. Finally, they let researchers to analyse the efficacy and efficiency of system on patients' outcomes. In other words, integrated systems play

a fundamental role in complex clinical environments.

The STEMI Archive consists of clinical information collection related to patients admitted in all hospitals of Regione Lombardia with STEMI diagnosis. The STEMI Archive, as well as every survey on specific disease, enables researchers to point out a subpopulation of interest for clinical and scientific inquiries. Starting from these subpopulations, studies on effectiveness of different patterns of care and then provider profiling can be carried out, adopting models for explaining outcomes by means of suitable process indicators and adjusting for different case mix. In our case, a primary outcome measure is incidence of Major Adverse Cardiovascular Events (MACE) defined as any one of the following events: in-hospital mortality, Acute myocardial reinfarction, Cardiogenic shock, Stroke, Long term Mortality, Major bleeding. A secondary outcome is reperfusion effectiveness measured quantifying the reduction of ST segment elevation one hour after the treatment: if the reduction is larger than 50% in the case of thrombolysis and 70% in the case of angioplasty we could consider the procedure effective. Process indicators and patients covariates can be resumed in the following four categories:

- Personal data: *Codice Fiscale* (the alpha-numeric identity code used to identify people who have fiscal residence on Italian territory), date of birth, sex, weight, height, hospital of admission;
- Risk factors: diabetes, smoking, high blood pressure, high cholesterol level, history of cardiac pathology;
- Admission data: time and type of symptoms onset, time of first medical contact, time to call for rescue, type of rescue unit sent (advanced or basic rescue unit, that is with or without pre-hospital ECG tele-transmission), time of first ECG, site of infarction on ECG, mode of hospital admittance, Fast Track activation, Killip class (which quantify in four categories the severity of infarction), systolic blood pressure, cardiac frequency, ejection fraction and creatinine value at admittance, site of ST-elevation, number of leads with ST-elevation, pre-hospital hearth failure;
- Therapeutic data: time of thrombolysis (Door to Needle time), time of angioplasty (Door to Balloon time), culprit lesion, Ejection Fraction and therapy at discharge;
- Outcomes: in-hospital survival, ST-resolution after 60 minutes from treatment, bleedings, shock, re-AMI, acute pulmunary edema, arrhythmias, Mytral regurgitation.

The eligible cohort consists in all patients admitted to any hospitals of the Regione Lombardia Network with STEMI diagnosis.

In addiction to what is provided by STEMI Archive as typical clinical registry, the innovative contents of this survey are represented by process indicators recorded in it: the main idea is to evaluate treatment times with the aim of designing a preferential therapeutic path to reperfusion in STEMI patients, and to direct the patient flow trough different pathways according, for example, to on hours vs off hours of working time table, or to clinical conditions such severity of infarction. In this sense, this survey represents an instrument both for epidemiological enquiries and for organizational optimization of the cardiological healthcare networks.

Moreover, personal data are collected not only for administrative purposes, but also so that the patient can be univocally identified also within administrative datawarehouse and a longitudinal electronic record containing his/her previous clinical history and follow-up can be traced, thanks to

the potential of Electronic Health Record. The link between the two databases generates the primary platform for the study of impact and care of STEMI on the whole territory of Regione Lombardia. Finally, information concerning outcomes are recorded, so that they can be returned to clinicians and institutions appropriately exploited in terms of patient's case-mix and care pattern, in order to support healthcare decisions and clinical policies through monitoring and analysing data. These steps may be carried out through suitable statistical monitoring and modelling. Statistical models, in fact, are able to capture complexity, variability and grouped nature of these data, providing an evidence based decisional support as well as pursuing the optimization of healthcare offer.

The STEMI Archive should overcome the difficulties faced in previous pilot data collections (i.e., MOMI^2, GestIMA, LombardIMA) related to non-uniformity, inaccuracy of filling and data redundance. In particular non-uniformity of data collection among different structures, or among successive surveys, and inaccuracy in filling dataset fields ceased to be a problem because the Archive procedure for collecting data has become mandatory for all hospitals through a directive issued by the lawmaker [36]. All centers fill in the registry along the same protocol and with the same software, thanks to the help of Lombardia Informatica (`http://www.lispa.it`), the Information & Communication Technology (ICT) society which Regione Lombardia leans on for implementation of Electronic Health Record. Opinion leaders and Scientific Societies of cardiology agreed upon all fields to be recorded and a unique data collector was identified in the Governance Agency for Health, that is also the data owner. Moreover, since this registry is designed to be automatically linked with administrative databases, inaccuracy of information will be partially overcome by the fact that, after the linkage, all information contained in it are checked for coherence with those contained in PHD. Then only information of interest will be extracted, avoiding redundance and achieving greater accuracy and reliability (for further details on record linkage, see [39]).

Three data collections have been planned within the end of Strategic Program (December 2011). The first has been performed during the time slot of January-December 2010, to set, test and calibrate the STEMI Archive; the second one, from January 2011 to July 2011, represents the first official period of data collection and is the one we will refer to in the analyses of Chapter 7; finally a third collection period has been realized during October-December 2011. For this latter, anyway, data from STEMI Archive are already available, but the integration with administrative datawarehouse will be ready only at the end of January. The time planning is dictated by the need of providing all clinical providers involved in the project with a suitable assistance that enables them to overcome software and technical hitches, especially concerning SISS system.

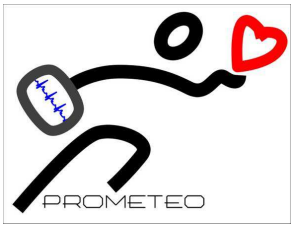
3.2.4 The PROMETEO database

As we said in Paragraph 3.2.2, the ECG tele-transmission is crucial for the efficient and quick management of STEMI patients. This Paragraph is then focused on a different source of clinical data, arising from 118 Electrocardiographic traces. Statistical methods for dealing with these data will be presented in Chapter 6, and their application to these data will be the focus of Chapter 9.

Since 2008, a project named PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero) has been started with the aim of spreading the intensive use of ECG as pre-hospital diagnostic tool and of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. In fact, anticipating diagnostic time, reducing infarction complications and optimizing the number of hospital admissions are the three main goals of PROMETEO. Thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara Rangoni Europe s.r.l., ECG recorder with GSM transmission

have been installed on all Basic Rescue Units (BRUs) of Milanese urban area. Persuading people to call 118 Dispatch Center when needed, equipping all Milan rescue units with ECG recorder and training rescuers to acquire ECG correctly to all people which call 118 for rescue, regardless of symptoms declared, is the way to strongly reduce delays in treatments and then in reperfusion. In fact this could be the way to obtain early diagnosis and then a quicker delivery of patients from territory to Intensive Cardiac Care Units (ICCU) of Hospitals, i.e., a better service for patients affected by ACS, enabling them to avoid to spend time in the ER and to go directly to PCI.

In [80], [81] and [82], data coming from PROMETEO datawarehouse are analysed. The aims of these analyses are, among others, to point out a semi automatic diagnostic tool for Bundle Branch Blocks (details on pathology are given later on this paragraph) based on statistical unsupervised classification algorithms [80] and inferential analysis on ECGs, considered as multivariate functional signals, coming from different population of physiological and pathological patients [81], [82].



PROMETEO datawarehouse contains all the ECG traces recorded by Basic and Advanced Rescue Units (ambulances and advanced units with physicians on board) on Milanese urban area, concerning patients who call 118 because supposed to be affected by infarction. Each file contained in PROMETEO datawarehouse is in correspondence to three sub-files. The first one is called *Details* and contains technical information, useful for signal processing and analysis, such as times of waves' repolarization and depolarization, landmarks indicating onset and offset times of main ECG's subintervals and automatic diagnoses, established by Mortara-Rangoni VERITAS^{TM} algorithm[1]. We used these automatic diagnoses to label ECG traces we analysed, in order to validate the performances of our unsupervised clustering algorithm. The challenge of the work proposed in [80], in fact, consists of tuning and testing a real time procedure which enables semi automatic diagnosis of the patients' disease based only on ECG traces morphology, then not dependent on clinical evaluations. The second sub-file is called *Rhythm* and contains the ECG signal sampled for 10 seconds (10000 sampled points). The third one is called *Median*. It is built starting from *Rhythm* file, and depicts a *reference* beat lasting 1.2 seconds (1200 points). Technical details on signal filtering are reported in the Mortata Rangoni *Physician's Guide to VERITAS with adult and pediatric resting EG interpretation* (available at `www.mortara.com`). We carried out the analysis considering only the *Median* files, obtaining 8 curves (one for each ECG lead) for each patient, which represents his/her "Median" beat for that lead. Examples of *Rhythm* and *Median* files of a patient are reported in Figures 3.3 and 3.4 respectively.

The main goal of the analysis of these data is then to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. To this aim, it is extremely important to understand the link between cardiac physiology and ECG trace shape. As detailed in the following, we focus on physiological traces in contrast to Right and Left Bundle Branch Block (RBBB and LBBB respectively) traces. Bundle Branch Block (BBB) is a cardiac conduction abnormality seen on the ECG. In this condition, activation of the left (right) ventricle is delayed, which results in the one ventricle contracting later than the other.

[1]Mortara Rangoni Europe s.r.l. is the leading provider of ECG algorithms and components for various clinical applications, see `http://www.mortara.com`.

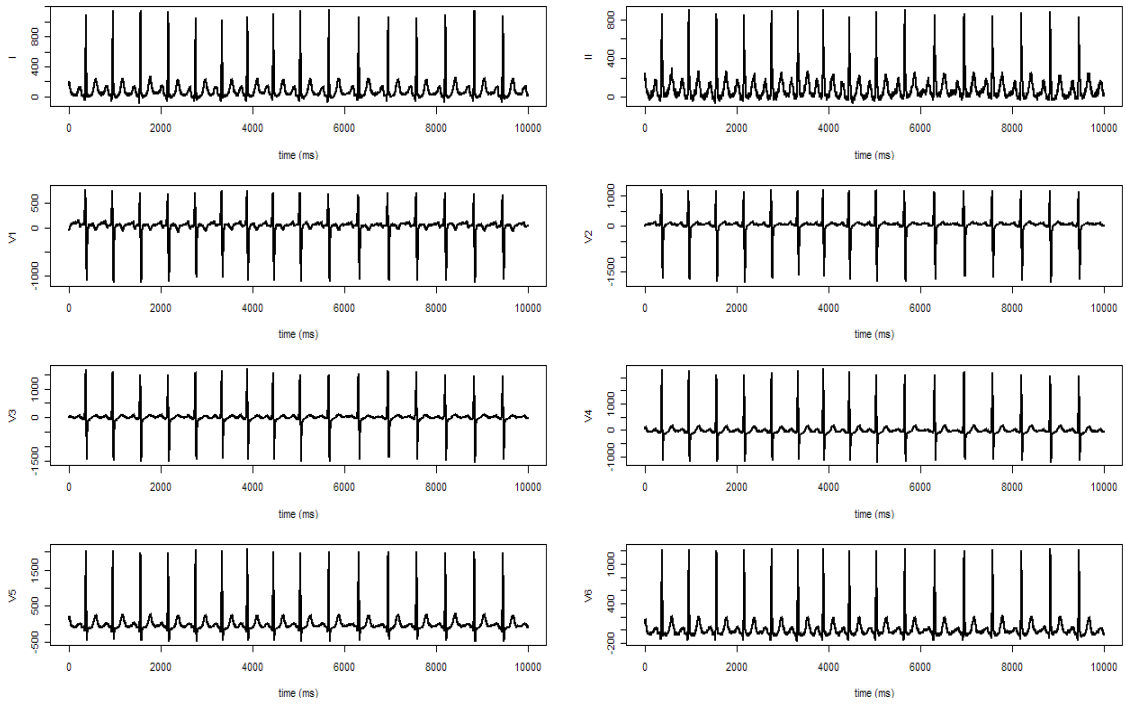


Figure 3.3: An example of file *Rhythm*.

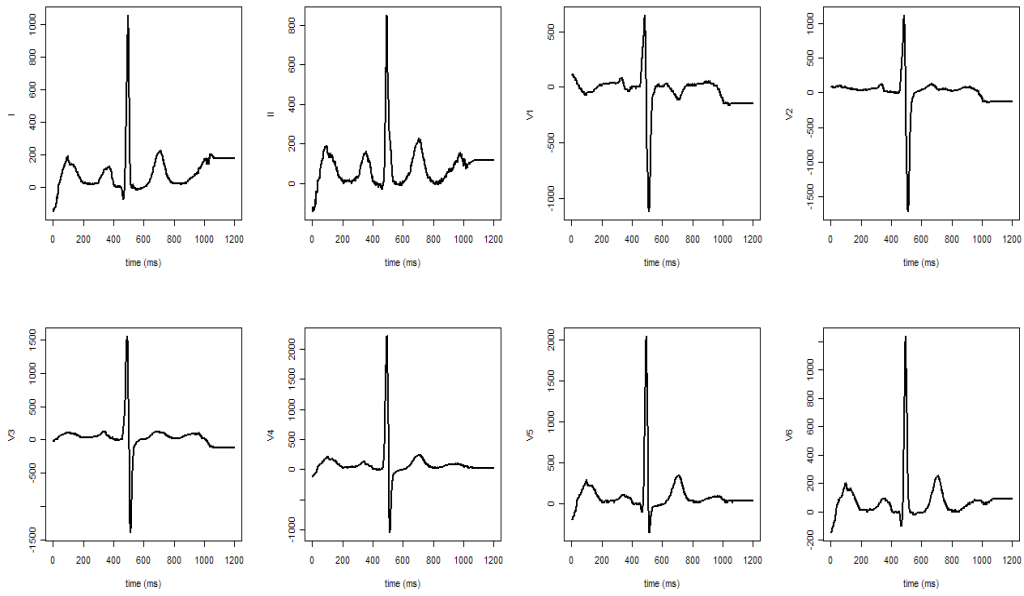


Figure 3.4: An example of file *Median*.

Electrocardiography and Bundle Branch Block

Electrocardiography is a transthoracic recording of the electrical activity of the heart over time captured and externally recorded through skin electrodes. The ECG works mostly by detecting and

amplifying the tiny electrical changes on the skin that are caused when the heart muscle depolarises during each heart beat (for further inquiry about clinical details, see [185]). First attempts of measuring ECG signals date back to Willem Einthoven (see [166], [167]). The Einthoven *limb leads* (standard leads) are illustrated in Figure 3.5 and are defined in the following way:

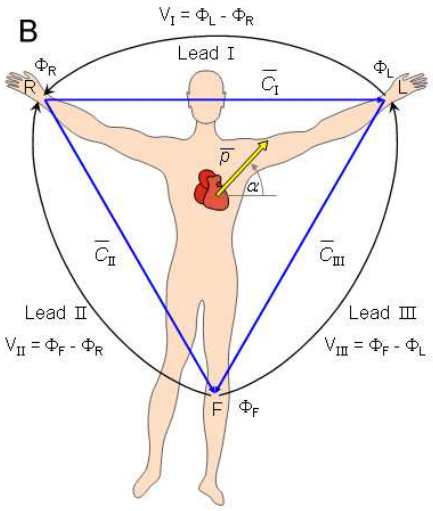


Figure 3.5: Einthoven limb leads

$$\begin{aligned} \text{Lead I} &: V_I = \Phi_L - \Phi_R, \\ \text{Lead II} &: V_{II} = \Phi_F - \Phi_R, \\ \text{Lead III} &: V_{III} = \Phi_F - \Phi_L; \end{aligned}$$

where

$$\begin{aligned} V_I &= \text{voltage of Lead I} \\ V_{II} &= \text{voltage of Lead II} \\ V_{III} &= \text{voltage of Lead III} \\ \Phi_L &= \text{potential at the left arm} \\ \Phi_R &= \text{potential at the right arm} \\ \Phi_F &= \text{potential at the left foot} \end{aligned}$$

These lead voltages satisfy the following relationship:

$$V_I + V_{III} = V_{II}, \tag{3.1}$$

hence only two of these three leads are independent. A simple model results from assuming that the cardiac sources are represented by a dipole located at the center of a sphere representing the thorax, hence at the center of an equilateral triangle. With these assumptions, the voltages measured by the three limb leads are proportional to the projections of the electric heart vector on the sides of the lead vector triangle. The voltages of the leads are obtained from Equation (3.1).

Nowadays, the most commonly used clinical ECG-system, the 12-lead ECG system, consists of the following 12 leads: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6. The main reason for recording all 12 leads is that it enhances pattern recognition (see [174] and [175]; [187] and [216]). Of these 12 leads, the first six are derived from the same three measurement points. Therefore, any two of these six leads include exactly the same information as the other four. So, the ECG traces analysed in the following sections will consist of leads I, II, V1, V2, V3, V4, V5 and V6 only.

Figure 3.6 shows a scheme of the stylized shape of a physiological single beat, recorded on ECG graph paper; main relevant points, segments and waves are highlighted. Deflections in this signal are denoted in alphabetic order starting with the letter P, which represents atrial depolarization. The ventricular depolarization causes the QRS complex, and repolarization is responsible for the T-wave. Atrial repolarization occurs during the QRS complex and produces such a low signal amplitude that it cannot be detected, with the exception of physiological ECGs (see [201]). The direction of travel of the wave of depolarization is named the *heart electrical axis*.

In the case of interest, the file *Rhythm* of our dataset represents the output of an ECG recorder. From this curve, a representative heartbeat for each patient is obtained and it is provided in the file

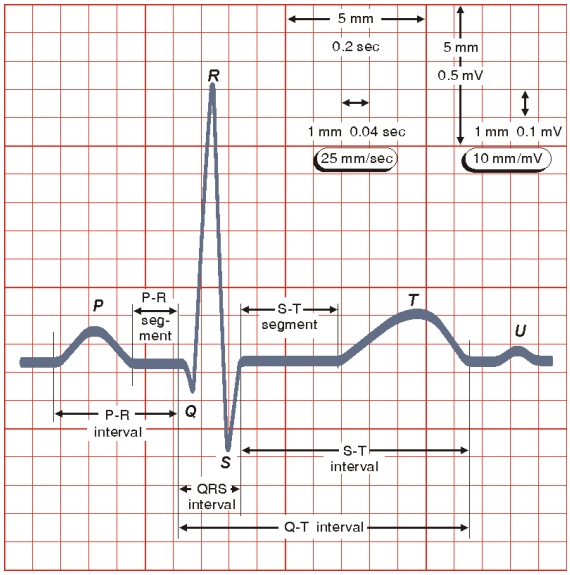


Figure 3.6: Scheme of the stylized shape of a physiological single beat, recorded on ECG graph paper. Main relevant points, segments and waves are highlighted.

Median. As we said before, it consists of a trace of a single cardiac cycle (heartbeat), i.e., of a *P* wave, a *QRS* complex, a *T* wave, and a *U* wave, which are normally visible in 50% to 75% of ECGs.

The heart's electrical activity begins in the sinoatrial node (the heart's natural pacemaker, n.1 in Figure 3.7), which is situated on the upper right atrium. The impulse travels next through the left and right atria and summates at the AV node (n.2 in Figure 3.7). From the AV node the electrical impulse travels down the Bundle of His (n.3 in Figure 3.7) and divides into the right and left bundle branches (n.4 and n. 10 in Figure 3.7). The right bundle branch contains one fascicle. The left bundle branch subdivides into two fascicles: the left anterior fascicle and the left posterior fascicle (n.4 and 5 in Figure 3.7). Ultimately, the fascicles divide into millions of Purkinje fibres which in turn interdigitise with individual cardiac myocytes, allowing for rapid, coordinated, and synchronous physiologic depolarization of the ventricles.
Bundle branch or fascicle injuries result in altered pathways for ventricular depolarization. In this case, there is a loss of ventricular synchrony, ventricular depolarization is prolonged, and there may be a corresponding drop in cardiac output. From a clinical perspective, a RBBB typically causes prolongation of the last part of the QRS complex, and may shift the heart electrical axis slightly to the right. LBBB widens the entire QRS, and in most cases shifts the heart electrical axis to the left. Another usual finding with bundle branch block is appropriate T wave discordance: this means that the T wave will be deflected opposite the terminal deflection of the QRS complex. From a statistical point of view, we will focus our analysis on shape modifications induced on the ECG curves and their first derivatives by the BBB pathology, and we will investigate these shape modifications only in a statistical perspective, i.e., not using clinical criteria to classify ECGs. The exploitation of these morphological modifications in the clustering procedure is the focus of the [80].

3.3 More complex data

Most of the indicators arising from clinical surveys or administrative datawarehouse only measure partial aspects of the health system: in the first case they take a clinical, real time but partial snapshot

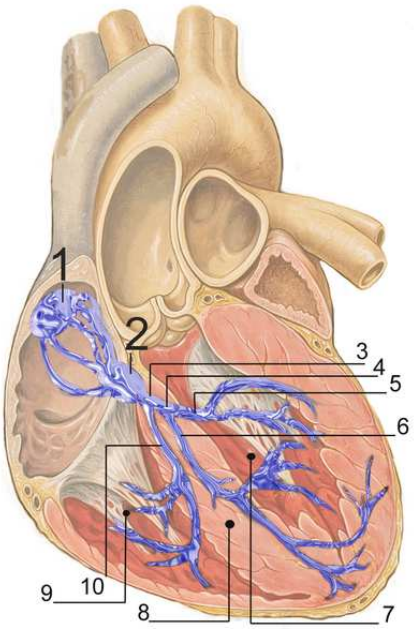


Figure 3.7: Conduction system of the heart: 1. Sinoatrial node; 2. Atrioventricular node; 3. Bundle of His; 4. Left bundle branch; 5. Left posterior fascicle; 6. Left-anterior fascicle; 7. Left ventricle; 8. Ventricular septum; 9. Right ventricle; 10. Right bundle branch.

of the population of interest, in terms of unavoidable biases and missing of longitudinal sight; on the other hand, administrative data collections give the researchers/government an idea of costs, degree and characteristics of supply, organizational factors, access to health care, population health status, but not much about the processes within the hospitals. Indications about criteria for the definition of such measures are scanty and research about the validation of the indicators has not been properly developed. On the basis of these considerations the National Agency for Regional Health Services in Lombardia (Agenzia per i Servizi Sanitari Regionali - ASSR) developed a set of quality measures (outcome and process indicators) in the context of the Strategic Program founded by the Ministry of Health, as we said in Chapter 2, Section 2.2. Indeed, one of the main goals of the Strategic Program is finding a set of indicators useful for comparison and classification of health care providers and for the identification of factors which can produce different outcomes, using both sources of data. The way this can be accomplished is explained in the following sections.

3.3.1 Data Mart of Regione Lombardia datawarehouse

A Data Mart (DM) is the access layer of the datawarehouse (DW) environment that is used to get data out to the users. The DM is a subset of the datawarehouse, usually oriented to a specific target. Within the Strategic Program, Lombardia Informatica S.p.A. (LISPA), provided a Data Mart called DWRETEIMA to enable the data collection of STEMI Archive and the linkage of this data with the PHD of Regione Lombardia. Details of software and services it provides are described in [161]. This software has been produced thanks to a strict collaboration among cardiologists (who provided the epidemiological needs and clinical expertise), statisticians (who provided scientific knowledge about experimental design and who are the intermediate user of data) and staff of LISPA (who provided the technical support), according to the issues highlighted in Section 1.2. The goal of DWRETEIMA is then to allow the clinical survey STEMI Archive to be as complete as possible

without being redundant with respect to information contained in the PHD. In Figure 3.8 the structure of DWRETEIMA supporting the STEMI Archive is shown (with a zoom on one sub folders in Figure 3.9). In the next paragraph will be shown how this instrument enables researchers to obtain integrated longitudinal data for each patient.

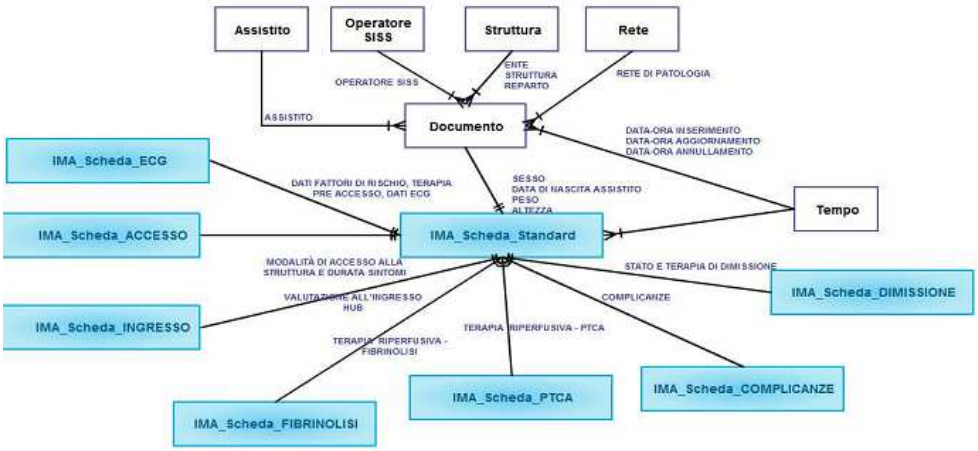


Figure 3.8: The source DWRETEIMA aggregates all sources of data requested by STEMI Archive in the single record connected to each patient admission.

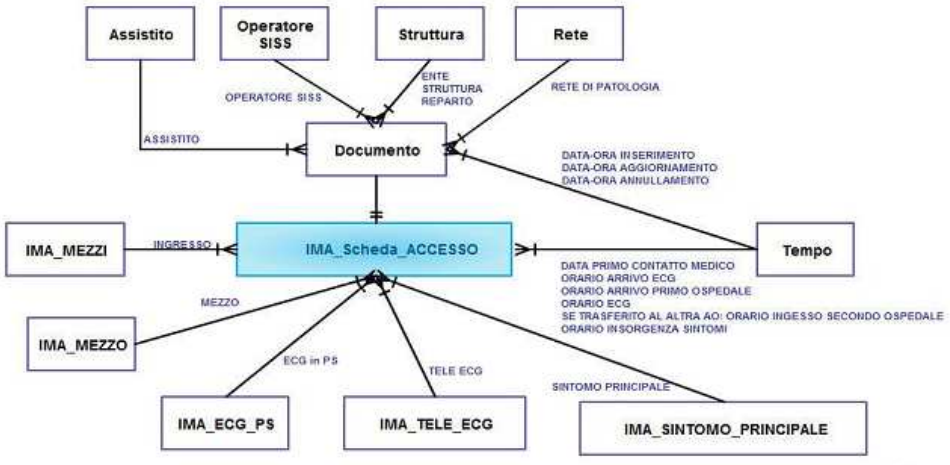


Figure 3.9: Example of subfolder collecting way of admission, symptoms and times.

3.3.2 Integrated system: examples of complex longitudinal data

Once different sources of data have been linked, it is possible for clinicians, researchers and people involved in healthcare governance to answer epidemiological questions such: is the trigger event of the STEMI Archive the first cardiological event for the observed patient? If not, how many cardiovascular events have been recorded in the previous history of this patient? These information are provided by the integration of the STEMI Archive with *Ricoveri* administrative database. Moreover,

if a patient is already known to the healthcare systems in terms of cardiovascular hospital admissions, was his/her compliance to the therapy good, i.e., did he/she assumed correct quantities of drugs and received a convenient treatment in terms of visits and clinical practice? These information are provided by the integration of the STEMI Archive with *BDA* and *Farmaci* administrative databases. Moreover, how the previous clinical history of each patient affects his/her outcome observed in the STEMI Archive gathering? These questions ask for a proper statistical modelling and represent real and new challenges of the Strategic Program. Finally, the long term mortality of each patient can be obtained through the linkage with *Anagrafica*. Even if the long term mortality is to be intended as the mortality due to any cause, not only to cardiovascular events, if considered up to 30/60 days it is strongly realated to cardiovascular causes. Anyway it is the first time that the long term mortality and follow up can be achieved for a clinical registry of STEMI in Regione Lombardia.

All information coming from the integration enable the researchers to point out new prognostic factors to be considered for better explain the main outcomes the hospitals are evaluated on. Then, the longer is the time slot on which the integration can be performed, the richer, the more complete and the more reliable is the information which can be used in order to built the outcome measures. Regione Lombardia enabled us to look at the administrative datawarehouse up to 8 years ago. In such time slot, a single patient could have order of dozens admissions, hundreds of visits, drugs and procedures. Dealing with such complex and high dimensional data is the challenge of the statistical analysis.

In this section we discuss the results of integration, in terms of the longitudinal electronic records obtained for each patient inserted in the STEMI Archive. We said that, over recent years, there has been an increasing agreement among epidemiologists on the validity of disease and intervention registries based on administrative databases and that this motivated Regione Lombardia to use its own administrative databases for clinical and epidemiological aims. Research using disease and intervention registries, outcome studies using administrative databases and performance indicators adopted by quality improvement methods can all shed light on who is most likely to benefit, what the important tradeoffs are and how policy makers might promote the safe, effective and appropriate use of new interventions.

When in the PHD we look for events related with a patient belonging to the population selected by a clinical registry, for example the STEMI Archive (see Figure 3.10), we find all his clinical history in term of healthcare utilization (visits, hospital admissions, drugs, etc). Since we are not interested in all this huge amount of information, but only in cardiovascular events, criteria for adequately choose only the hospital discharge records effectively related to cardiovascular events of the patient of interest are needed. In fact, the most critical issue when using administrative databases within observational studies is represented by the selection criteria of the discharge records: several different criteria may be used, and they will result in different images of prevalence or incidence of diseases. Among the most accepted criteria, those referring to the Agency for healthcare Research and Quality (AHRQ) methodology, the ones of Johns Hopkins Adjusted Clinical Groups (ACG) and Classification Research Groups (CRG) have been considered (for further details, see [13]).
As we said before, integrating clinical surveys on specific diseases with administrative databanks, enable us to select subpopulation of interest for observational studies, focused on answering to specific epidemiological needs. In fact, the main point and the novelty of Strategic Program is the proposal of an epidemiological research for specific subpopulation of interest pointed out by clinical registries, which is different from the classical epidemiological inquiry since it is conducted starting from the Electronic Health Records, then it is faster and cheaper, and moreover it is real

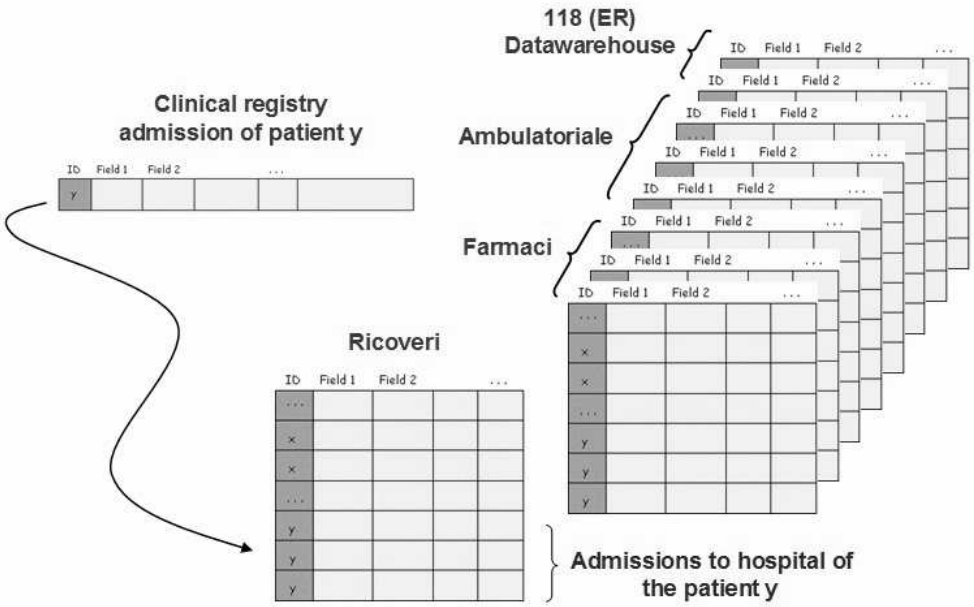


Figure 3.10: Sketch of integration between STEMI Archive and Public Health Database.

time achieving. For this new epidemiology, new methods for inquiry and analysis must be pointed out, and adequate information media must be provided. The STEMI Archive described in Paragraph 3.2.3 and statistical models proposed in Part II are some of the instruments to be adopted to this aim, and the Strategic Program is the first official set in Italy where they have been considered.

As previously mentioned, when integration of different sources of data is performed, attention must be paid to a carefully selection of covariates and data of interest. In this sense, several further problems arise: firstly it is necessary to select only cardiovascular events and events in some way related to this pathology; then a dimensional reduction is needed, pointing out just covariates which can be of interest in exploiting outcomes by means of suitable covariates and process indicators. This is the challenge of the statistician, and it is strongly related with the clinical questions that physicians want to investigate. In this sense, several analyses can performed on such rich and complex data. A review of techniques to be applied to such data can be found in [13] and [83]

Dealing with integrated data provides us rich longitudinal data containing lots of different information about each statistical unit (patient) of the subpopulation of interest. On the other hand, the more information, the more complexity, then suitable statistical methods must be developed in order to manage this information in the most fruitful way. For example, we could be interested in using information provided by the past clinical history of a patient to predict his/her risk of re-hospitalization. The use of hospitalizations (arising from integration of STEMI Archive with the database *Ricoveri* of the PHD) to study the risk of a new event or to quantify issues concerned with healthcare assessment is an innovative approach, since no standard methodology exists to exploit this kind of data.

Thinking to the integrated data like a longitudinal (functional) observation for each patient or like a realization of a stochastic counting process led us to implement models like the ones explained in Section 6.3 mentioned in [15]. The idea is that database integration, counting process modelling

of hospitalizations and generalized functional mixed models are methodologies that can be applied to the study of many different pathologies, thanks to their flexibility and capability of dealing with complex data. Although it can seem contradictory to define functional data as “synthetic”, it is clear that complex, heterogeneous data are easier to study if their effect is resumed with a process that represents their combined effect on instantaneous risk. Moreover, specific epidemiological enquires can be addressed using integrated database, starting from the compliance to prescribed therapy up to the impact of using specific treatments, drugs and devices. Actually we are still working on feasibility of analyses on integrated systems, but it is clear that such a font of real time information has a great potential within the context of both in clinical and economic assessment of Regione Lombardia. In fact, this kind of methodology has led to interesting preliminar results that could have an impact on the planning of this care strategy. Further development of this framework in cooperation with medical staff could lead to the definition of a useful guidelines for supporting long term decisions and performing health care assessment concerning policies to be adopted with patient affected by STEMI.

3.4 Data mining of integrated databases

Capabilities of both generating, collecting and storing data have been increasing dramatically in the last two decades. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge. Moreover, the shift toward evidence-based practice and outcomes research presents significant opportunities and challenges to extract meaningful information from massive amounts of clinical data to transform it into the best available knowledge to guide clinical practice. In fact, healthcare has been no exception: modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database becomes increasingly necessary. Databases are increasing in size in two ways: (1) the number of records or objects in the database and (2) the number of fields or attributes to an object. Examples of these can be found in the administrative datawarehouse of Regione Lombardia as well as the integration among clinical and on going data collections.

Data mining, a step in the process of Knowledge Discovery in Databases (KDD) (as shown in Figure 3.11), is a method of unearthing information from large data sets. Built upon statistical analysis, it can analyse massive amounts of data and provide useful and interesting information about patterns and relationships that exist within the data that might otherwise be missed. In medicine, data mining can improve the management level of hospital information and promote the development of telemedicine and community medicine. Because the medical information is characteristic of redundancy, multi-attribution, incompletion and closely related with time, medical data mining differs from other one.

In general, the basic problem addressed by the KDD process is one of mapping low-level data into other forms that might be more compact, more abstract, or more useful. The distinction between the KDD process and the data-mining step (within the process) is crucial. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data, especially when deling with high dimensional data. In our case, in fact, not only there is often a large number of records in the database, but there is also a large number of fields within each record; so, the dimensionality of the problem is high. A

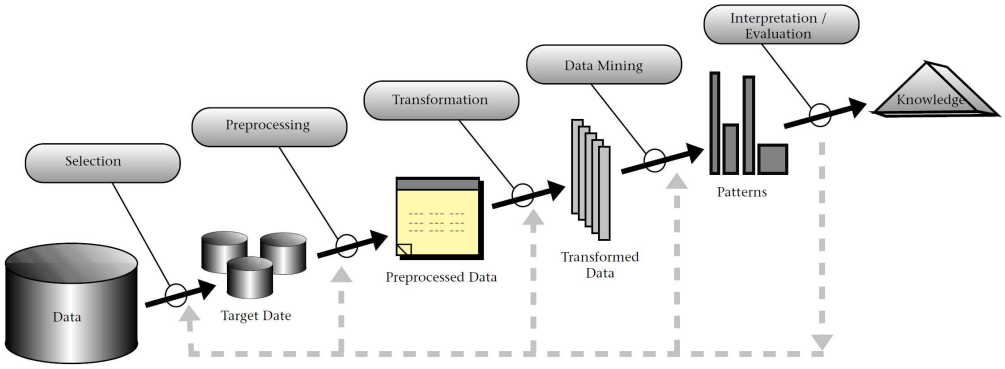


Figure 3.11: An Overview of the Steps That Compose the KDD Process.

high dimensional data set creates problems in terms of increasing the size of the search space for model induction in a combinatorially explosive manner. In addition, it increases the chances that a data-mining algorithm will find spurious patterns that are not valid in general. Approaches to this problem include methods to reduce the effective dimensionality of the problem and the use of prior knowledge to identify irrelevant variables. Another problem arising when dealing with high dimensional data is the problem of missing and/or noisy data: important attributes can be missing if the database was not designed with discovery in mind, or field can be wrongly filled. This calls for an accurate design of experiment, leading to a focused data collection.

Finally, the two high-level primary goals of data mining in practice tend to be description and prediction. Description focuses on finding human-interpretable patterns describing the data, whereas prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Although the boundaries between prediction and description are not sharp, the distinction is useful for understanding the overall discovery goal. The relative importance of prediction and description for particular data-mining applications can vary considerably. The goals of prediction and description can be achieved using a variety of particular data-mining methods.

In the context we are interested in, we focus on description when we perform monitoring of process indicators and analyse patterns of dependence among variables that influence them. On the other hand, we also develop models for survival and other outcomes mentioned in previous sections, using them in order to make predictions and to quantify losses/gains in terms of probability of success once we adjusted for all the external influences. In this way it is possible to quantify the providers' effect on outcomes, as detailed in Part II from a theoretical point of view.

Part II

Statistical models and methods for healthcare data

Social and medical researchers have long been concerned about the need properly to model complex data structures, especially those where there is a hierarchical structure such as pupils nested within schools or measurements nested within individuals. Failure to take account of such structures in standard models can lead to incorrect inferences. What has been less well appreciated is that a failure to properly model complex data structures makes it impossible to capture the complexity that exists in the real world.

Harvey Goldstein

Key words: Linear Mixed Effects Models; Generalized Linear Mixed Effects Models; Nonlinear Mixed Effects Models; Nonparametric Random Effects; EM algorithm; Longitudinal data; Overdispersion; Dirichlet Process; Dependent Dirichlet Process; Bayesian Decision Theory; Multivariate Functional Data; Depth Measures; Outlier Detection; Rank tests.

Chapter 4

Statistical models for healthcare: the frequentist approach

In this chapter, an overview of the principal frequentist statistical methods for dealing with grouped and hierarchical data is presented. The main goal is to understand how they can be useful in modelling problems such those arising from the context presented in Part I and how they can be applied to data that tipically come out from the healthcare context. We focus in particular on mixed effects models [55].

4.1 Motivations

Mixed effects models provide a flexible and powerful tool for the analysis of grouped data, which arise in many areas as agriculture, biology, economics, manufacturing, geophysics and so on. Examples of grouped data include longitudinal data, repeated measures, and multilevel data. The increasing popularity of mixed effects models is explained by the flexibility they offer in modelling the within-group correlation often present in grouped data, by the handling of balanced and unbalanced data in a unified framework, and by the availability of reliable and efficient software for fitting them [42] [49]. The mixed effects approach is based on the assumption that, for every group of observations, the response can be modeled by a linear regression model, but with group/(subject)-specific regression coefficients. In the case of repeated measures, the subject is in fact the grouping factor. Many common statistical models can be expressed as linear models that incorporate both fixed effects, which are parameters associated with an entire population or with certain repeatable levels of experimental factors, and random effects, which are associated with individual experimental units drawn at random from a population. A model with both fixed effects and random effects is called a mixed effects model. Mixed effects models are primarily used to describe relationships between a response variable and some covariates in data that are grouped according to one or more classification factors.

4.2 Linear parametric mixed effects models

Linear parametric Mixed Effects (LME) models are mixed effects models in which both the fixed and the random effects occur linearly in the model function and with suitable parametric assumption on the random effects distribution. They extend linear models by incorporating random effects, which can be regarded as additional error terms, to account for correlation among observations

within the same group.

4.2.1 Single and multi level of grouping

For a single level of grouping, the linear mixed effects model expresses the n_i-dimensional response vector $\mathbf{y}_i$ for the i-th group as

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i \qquad i = 1,\ldots,M \tag{4.1}$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0},\Sigma) \qquad \varepsilon_i \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbb{I})$$

where β is the p-dimensional vector of fixed effects, $\mathbf{b}_i$ is the q-dimensional vector of random effects, $\mathbf{X}_i$ (of size $n_i \times p$) and $\mathbf{Z}_i$ (of size $n_i \times q$) are known fixed effects and random-effects regressor matrices, and ε_i is the n_i-dimensional within-group error vector with a Normal distribution. The assumption $Var(\varepsilon_i) = \sigma^2\mathbb{I}$ can be relaxed, extending to the case of nonconstant variances or special within-group correlation structures (for further details, see [122]). The random effects $\mathbf{b}_i$ and the within-group errors ε_i are assumed to be independent for different groups and to be independent of each other for the same group. Because the distribution of the random effects vectors $\mathbf{b}_i$ is assumed to be Normal with zero mean, it is completely characterized by its variance-covariance matrix Σ. This matrix must be symmetric and positive semi-definite; that is, all its eigenvalues must be non-negative. We will make the stronger assumption that it is positive-definite which is to say that all its eigenvalues must be strictly positive. We can make this restriction because a singular model can always be re-expressed as a positive-definite model of lower dimension. The columns of $\mathbf{Z}_i$ are usually a subset of the columns of $\mathbf{X}_i$. When computing with the model it is more convenient to express the variance-covariance matrix in the form of a relative precision factor, $\mathbf{\Delta}$, which is any matrix that satisfies the following equality:

$$\frac{\Sigma^{-1}}{1/\sigma^2} = \mathbf{\Delta}^T\mathbf{\Delta}$$

If Σ is positive-definite then such a $\mathbf{\Delta}$ will exist, but it need not be unique. The Cholesky factor of $\sigma^2\Sigma^{-1}$ is one possible $\mathbf{\Delta}$. The matrix $\mathbf{\Delta}$ is called a *relative precision factor* because it factors the precision matrix of the random effects (Σ^{-1}), expressed relative to the precision, $1/\sigma^2$, of the ε_i.

The formulation for single level LME models presented above can be extended to multiple, nested levels of random effects. In the case of two nested levels of random effects the response vectors at the innermost level of grouping are written $\mathbf{y}_{im}$, $i = 1,\ldots,M$, $m = 1,\ldots,M_i$ where M is the number of first-level groups and M_i is the number of second-level groups within first-level group i. The length of $\mathbf{y}_{im}$ is M_{im}. The fixed effects model matrices are $\mathbf{X}_{im}$, $i = 1,\ldots,M$, $m = 1,\ldots,M_i$ of size $M_{im} \times p$. Using first-level random effects $\mathbf{b}_i$ of length q_1 and second-level random effects $\mathbf{b}_{im}$ of length q_2 with corresponding model matrices $\mathbf{Z}_{i,m}$ of size $M_i \times q_1$ and $\mathbf{Z}_{im}$ of size $M_i \times q_2$, we write the model as

$$\mathbf{y}_{im} = \mathbf{X}_{im}\beta + \mathbf{Z}_{i,m}\mathbf{b}_i + \mathbf{Z}_{im}\mathbf{b}_{im} + \varepsilon_{im} \qquad i = 1,\ldots,M, \quad m = 1,\ldots,M_i \tag{4.2}$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0},\Sigma_1), \quad \mathbf{b}_{im} \sim \mathcal{N}(\mathbf{0},\Sigma_2), \quad \varepsilon_{im} \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbb{I})$$

The level-1 random effects $\mathbf{b}_i$ are assumed to be independent for different i, the level-2 random effects $\mathbf{b}_{im}$ are assumed to be independent for different i or m and to be independent of the level-1 random effects, and the within group errors ε_{im} are assumed to be independent for different i or m and to be independent of the random effects. Extensions to an arbitrary number Q of levels

of random effects follow the same general pattern. As with a single level of random effects, the variance-covariance matrix Σ_q, $q = 1, \ldots, Q$ will be expressed in terms of relative precision factors $\mathbf{\Delta}_q$.

In what follows, we will always assume a single level of grouping, assuming every time a grouped data framework (where subjects are units within groups) or longitudinal data structure (where subjects are the grouping factor with respect to their own data).

4.2.2 Estimation in LME models

Consider first the model (4.1) that has a single level of random effects. The parameters of the model are β, σ^2 and whatever parameters determine $\mathbf{\Delta}$. We use θ to represent an unconstrained set of parameters that determine $\mathbf{\Delta}$, assuming for instance that a suitable parametrization has been chosen for it (for further discussion on this topic, see [73] and [122], Paragraph 2.2.7). The likelihood function for the model (4.1) is the probability density for the data given the parameters, but regarded as a function of the parameters with the data fixed, instead of as a function of the data with the parameters fixed. That is,

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = f(\mathbf{y}|\beta, \theta, \sigma^2)$$

where L is the likelihood, f is a probability density, and $\mathbf{y}$ is the entire N-dimensional response vector, $N = \sum_{i=1}^{M} n_i$. Because the nonobservable random effects $\mathbf{b}_i$, $i = 1, \ldots, M$ are part of the model, we must integrate the conditional density of the data given the random effects with respect to the marginal density of the random effects to obtain the marginal density for the data. We can use the independence of the $\mathbf{b}_i$ and the ε_i to express this as

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{M} f(\mathbf{y}_i|\beta, \theta, \sigma^2) \\ &= \prod_{i=1}^{M} \int f(\mathbf{y}_i|\mathbf{b}_i, \beta, \sigma^2) f(\mathbf{b}_i|\theta, \sigma^2) d\mathbf{b}_i \end{aligned} \tag{4.3}$$

where the marginal density of $\mathbf{y}_i$ is multivariate Normal

$$f(\mathbf{y}_i|\mathbf{b}_i, \beta, \sigma^2) = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\mathbf{b}_i\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{n_i/2}} \tag{4.4}$$

and the marginal density of $\mathbf{b}_i$ is also multivariate Normal

$$\begin{aligned} f(\mathbf{b}_i|\theta, \sigma^2) &= \frac{\exp\left(-\mathbf{b}_i^T \Sigma^{-1} \mathbf{b}_i\right)}{(2\pi)^{q/2}\sqrt{|\Sigma|}} \\ &= \frac{\exp\left(-\|\mathbf{\Delta}\mathbf{b}_i\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{q/2} abs|\mathbf{\Delta}|^{-1}} \end{aligned} \tag{4.5}$$

Substituting (4.3) and (4.4) in (4.5) provides the likelihood as

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{M} \frac{abs|\mathbf{\Delta}|}{(2\pi\sigma^2)^{n_i/2}} \int \frac{\exp\left[-\left(\|\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\mathbf{b}_i\|^2 + \|\mathbf{\Delta}\mathbf{b}_i\|^2\right)\right]}{(2\pi\sigma^2)^{q/2}} d\mathbf{b}_i \\ &= \prod_{i=1}^{M} \frac{abs|\mathbf{\Delta}|}{(2\pi\sigma^2)^{n_i/2}} \int \frac{\exp\left(-\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\mathbf{b}_i\|^2\right)}{(2\pi\sigma^2)^{q/2}} d\mathbf{b}_i \end{aligned} \tag{4.6}$$

where

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{X}}_i = \begin{bmatrix} \mathbf{X}_i \\ \mathbf{0} \end{bmatrix}, \quad \tilde{\mathbf{Z}}_i = \begin{bmatrix} \mathbf{Z}_i \\ \boldsymbol{\Delta} \end{bmatrix}, \tag{4.7}$$

are augmented data vectors and model matrices. This approach of changing the contribution of the marginal distribution of the random effects into extra rows for the response and the design matrices is called a pseudodata approach because it creates the effect of the marginal distribution by adding "pseudo" observations. The exponent in the integral of (4.6) is in the form of a squared norm or, more specifically, a residual sum-of-squares. We can determine the conditional modes of the random effects given the data, written $\hat{\mathbf{b}}_i$, by minimizing this residual sum-of-squares. This is a standard least squares problem for which we could write the solution as

$$\hat{\mathbf{b}}_i = \left(\tilde{\mathbf{Z}}_i^T \tilde{\mathbf{Z}}_i\right)^{-1} \tilde{\mathbf{Z}}_i^T \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \beta\right).$$

The squared norm can then be expressed as

$$\begin{aligned} \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\mathbf{b}_i\|^2 &= \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\|^2 + \|\tilde{\mathbf{Z}}_i\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)\|^2 \\ &= \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\|^2 + \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^T \tilde{\mathbf{Z}}_i^T \tilde{\mathbf{Z}}_i \left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right) \end{aligned} \tag{4.8}$$

The first term in (4.8) does not depend on $\mathbf{b}_i$ so its exponential can be factored out of the integral in (4.6). Integrating the exponential of the second term in (4.8) is equivalent, up to a constant, to integrating a multivariate normal density function. Note that

$$\begin{aligned} &\frac{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \int \frac{\exp\left[-\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^T \tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}} d\mathbf{b}_i \\ &\qquad = \frac{1}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \int \frac{\exp\left[\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)^T \tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i\left(\mathbf{b}_i - \hat{\mathbf{b}}_i\right)/2\sigma^2\right]}{(2\pi\sigma^2)^{q/2}/\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \\ &\qquad = \frac{1}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} = \frac{1}{\sqrt{|\mathbf{Z}_i^T\mathbf{Z} + \boldsymbol{\Delta}^T\boldsymbol{\Delta}|}} \end{aligned} \tag{4.9}$$

By combining (4.8) and (4.9) we can express the integral in (4.6) as

$$\int \frac{\exp\left[-\left(\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\mathbf{b}_i\|^2\right)\right]}{(2\pi\sigma^2)^{q/2}} d\mathbf{b}_i = \int \frac{\exp\left[-\left(\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\hat{\mathbf{b}}_i\|^2\right)\right]}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \tag{4.10}$$

to give

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \left(\frac{-\sum_{i=1}^M \|\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\hat{\mathbf{b}}_i\|^2}{2\sigma^2}\right) \prod_{i=1}^M \frac{abs\boldsymbol{\Delta}}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \tag{4.11}$$

The expression (4.11) could be used directly in an optimization routine to calculate the maximum likelihood estimates for β, θ, and σ^2. However, the optimization is much simpler if we first *profile* the likelihood so it is a function of θ only. That is, we calculate the conditional estimates $\hat{\beta}(\theta)$ and $\hat{\sigma}^2(\theta)$ as the values that maximize $L(\beta, \theta, \sigma^2)$ for a given θ. Notice that the parts of (4.11) involving β and σ^2 are identical in form to the likelihood for a linear regression model so $\hat{\beta}(\theta)$ and $\hat{\sigma}^2(\theta)$ can be determined from standard linear regression theory.

We do need to be careful because the least squares estimates for β will depend on the conditional modes $\hat{\mathbf{b}}_i$ and these, in turn, depend on β. Thus, we must determine these least squares values jointly as the least squares solution to

$$\left(\hat{\mathbf{b}}_1^T,\ldots,\hat{\mathbf{b}}_M^T,\hat{\beta}_1^T\right)^T = \arg\min_{\mathbf{b}_1,\ldots,\mathbf{b}_M,\beta} \|\mathbf{y}_e - \mathbf{X}_e(\mathbf{b}_1,\ldots,\mathbf{b}_M,\beta)^T\|^2$$

where

$$\mathbf{X}_e = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{X}_1 \\ \boldsymbol{\Delta} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \ldots & \mathbf{0} & \mathbf{X}_2 \\ \mathbf{0} & \boldsymbol{\Delta} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{Z}_M & \mathbf{X}_M \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Delta} & \mathbf{0} \end{bmatrix}, \quad \text{and} \quad \mathbf{y}_e = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{0} \\ \mathbf{y}_2 \\ \mathbf{0} \\ \vdots \\ \mathbf{y}_M \\ \mathbf{0} \end{bmatrix} \tag{4.12}$$

Conceptually we could write

$$\left(\hat{\mathbf{b}}_1^T,\ldots,\hat{\mathbf{b}}_M^T,\hat{\beta}^T\right)^T = \left(\mathbf{X}_e^T\mathbf{X}_e\right)^{-1}\mathbf{X}_e^T\mathbf{y}_e$$

but we definitely would not want to calculate these values this way. The matrix $\mathbf{X}_e$ is sparse and can be very large. If possible we want take advantage of the sparsity and avoid working directly with $\mathbf{X}_e$. Linear regression theory also gives us the conditional maximum likelihood estimate for σ^2

$$\hat{\sigma}^2(\theta) = \frac{\|\mathbf{y}_e - \mathbf{X}_e(\hat{\mathbf{b}}_1^T,\ldots,\hat{\mathbf{b}}_M^T,\hat{\beta}^T)\|^2}{N} \tag{4.13}$$

Substituting these conditional estimates back into (4.11) provides the *profiled* likelihood

$$L(\theta) = L(\hat{\beta}(\theta),\theta,\hat{\sigma}^2(\theta)) = \frac{\exp(-N/2)}{[2\pi\hat{\sigma}^2(\theta)]^{N/2}}\prod_{i=}^{M}\frac{abs|\boldsymbol{\Delta}|}{\sqrt{|\tilde{\mathbf{Z}}_i^T\tilde{\mathbf{Z}}_i|}} \tag{4.14}$$

We do not actually need to calculate the values of $\hat{\mathbf{b}}_1,\ldots,\hat{\mathbf{b}}_M,\hat{\beta}(\theta)$ to evaluate the profiled likelihood. We only need to know the norm of the residual from the augmented least squares problem. There is, in fact, a decomposition methods that provide us with fast, convenient way of calculating this. Referring to the *Orthogonal-triangular decompositions* of rectangular matrices described in [122] Paragraph 2.2.2, we have that the *QR* decomposition of a general matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, $n > p$ is given by

$$\mathbf{X} = \mathbf{Q}\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$$

being $\mathbf{Q}$ a $n \times n$ orthogonal matrix, $\mathbf{R}$ a $p \times p$ upper triangular matrix. An important property of orthogonal matrixes is that they preserve the norm of vectors they are applied to. So, if we apply this to the residual vector of a least squares problem we get

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\beta\|^2 &= \|\mathbf{Q}^T(\mathbf{y} - \mathbf{X}\beta)\|^2 \\ &= \|\mathbf{Q}^T\mathbf{y} - \mathbf{Q}^T\mathbf{X}\beta\|^2 \\ &= \|\mathbf{c} - \mathbf{Q}^T\mathbf{Q}\begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}\beta\|^2 \\ &= \|\mathbf{c}_1 - \mathbf{R}\beta\|^2 + \|\mathbf{c}_2\|^2 \end{aligned}$$

where $\mathbf{c} = (\mathbf{c}_1^T, \mathbf{c}_2^T) = \mathbf{Q}^T\mathbf{y}$ is the rotated residual vector. The components of $\mathbf{c}_1$ and $\mathbf{c}_2$ are of lengths p and $n-p$ respectively. If $\mathbf{X}$ has rank p, the $p \times p$ matrix $\mathbf{R}$ is non singular and upper-triangular. The least-squares solution $\hat{\beta}$ is easily evaluated as the solution of

$$\mathbf{R}\hat{\beta} = \mathbf{c}_1$$

and the residual sum of squares is $\|\mathbf{c}_2\|^2$. Notice that the residual sum of squares can be evaluated without having to calculate β. Now, applying these last considerations to the framework of linear mixed effects models with single level grouping, we take an orthogonal decomposition of the augmented model matrix $\tilde{\mathbf{Z}}_i$

$$\tilde{\mathbf{Z}}_i = \mathbf{Q}_{(i)} \begin{bmatrix} \mathbf{R}_{11(i)} \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{Q}_{(i)}$ is $(n_i+q) \times (n_i+q)$ and $\mathbf{R}_{11(i)}$ is $q \times q$. Then

$$\begin{aligned} \|\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\mathbf{b}_i\|^2 &= \|\mathbf{Q}_{(i)}^T \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\beta - \tilde{\mathbf{Z}}_i\mathbf{b}_i\right)\|^2 \\ &= \|\mathbf{c}_{1(i)} - \mathbf{R}_{10(i)}\beta - \mathbf{R}_{11(i)}\mathbf{b}_i\|^2 + \|\mathbf{c}_{0(i)} - \mathbf{R}_{00(i)}\beta\|^2 \end{aligned}$$

where the $q \times p$ matrix $\mathbf{R}_{10(i)}$, the $n_i \times p$ matrix $\mathbf{R}_{00(i)}$, the q-vector $\mathbf{c}_{1(i)}$ and the n_i-vector $\mathbf{c}_{0(i)}$ are defined by

$$\begin{bmatrix} \mathbf{R}_{10(i)} \\ \mathbf{R}_{10(i)} \end{bmatrix} = \mathbf{Q}_{(i)}^T\tilde{\mathbf{X}}_i \quad \text{and} \quad \begin{bmatrix} \mathbf{c}_{1(i)} \\ \mathbf{c}_{0(i)} \end{bmatrix} = \mathbf{Q}_{(i)}^T\tilde{\mathbf{y}}_i$$

Returning to the integral in (4.6), we can now reduce it to

$$\begin{aligned} &\int \frac{\exp\left[-\left(\|\mathbf{y}_i - \mathbf{X}_i\beta - \mathbf{Z}_i\mathbf{b}_i\|^2 + \|\Delta\mathbf{b}_i\|^2\right)/2\sigma^2\right]}{\sqrt{2\pi\sigma^2}} d\mathbf{b}_i = \\ &\exp\left[\frac{\|\mathbf{c}_{0(i)} - \mathbf{R}_{00(i)}\beta\|^2}{-2\sigma^2}\right] \int \frac{\exp\left[\frac{\|\mathbf{c}_{1(i)} - \mathbf{R}_{10(i)}\beta - \mathbf{R}_{11(i)}\mathbf{b}_i\|^2}{-2\sigma^2}\right]}{(2\pi\sigma)^{q/2}} d\mathbf{b}_i \end{aligned} \tag{4.15}$$

and then providing the likelihood as

$$\begin{aligned} L(\beta, \theta, \sigma^2|\mathbf{y}) &= \prod_{i=1}^{M} \frac{\exp\left[-\|\mathbf{c}_{0(i)} - \mathbf{R}_{00(i)}\beta\|^2/2\sigma^2\right]}{(2\pi\sigma^2)^{n_i/2}} abs\left(\frac{|\Delta|}{|\mathbf{R}_{11(i)}|}\right) \\ &= \frac{\exp\left(-\|\sum_{i=1}^{M} \mathbf{c}_{0(i)} - \mathbf{R}_{00(i)}\beta\|^2/2\sigma^2\right)}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{M} abs\left(\frac{|\Delta|}{|\mathbf{R}_{11(i)}|}\right) \end{aligned}$$

Detailed procedure is reported in [122],§2.2.3. The term in the exponent has the form of a residual sum-of-squares for β pooled over all the groups. Forming another orthogonal-triangular decomposition

$$\begin{bmatrix} \mathbf{R}_{00(1)} & \mathbf{c}_{0(1)} \\ \vdots \quad \vdots & \\ \mathbf{R}_{00(M)} & \mathbf{c}_{0(M)} \end{bmatrix} = \mathbf{Q}_0 \begin{bmatrix} \mathbf{R}_{00} & \mathbf{c}_0 \\ \mathbf{0} & \mathbf{c}_{-1} \end{bmatrix} \tag{4.16}$$

produces the form

$$L(\beta, \theta, \sigma^2|\mathbf{y}) = (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{\|\mathbf{c}_{-1}\|^2 + \|\mathbf{c}_0 - \mathbf{R}_{00}\beta\|^2}{2\sigma^2}\right) \prod_{i=1}^{M} abs\left(\frac{|\Delta|}{|\mathbf{R}_{11(i)}|}\right) \tag{4.17}$$

For a given θ, the values of β and σ^2 that maximize (4.17) are

$$\hat{\beta}(\theta) = \mathbf{R}_{00}^{-1}\mathbf{c}_0 \quad \text{and} \quad \hat{\sigma}^2(\theta) = \frac{\|\mathbf{c}_{-1}\|^2}{N} \tag{4.18}$$

which give the profiled likelihood

$$\begin{aligned} L(\theta|\mathbf{y}) &= L(\hat{\beta}(\theta), \theta, \hat{\sigma}^2(\theta)|\mathbf{y}) \\ &= \left(\frac{N}{2\pi\|\mathbf{c}_{-1}\|^2}\right)^{N/2} \exp\left(-\frac{N}{2}\right) \prod_{i=1}^{M} abs\left(\frac{|\mathbf{\Delta}|}{|\mathbf{R}_{11(i)}|}\right) \end{aligned} \tag{4.19}$$

The profiled likelihood (4.19) is maximized with respect to θ, producing the maximum likelihood estimate $\hat{\theta}$. The maximum likelihood estimates $\hat{\beta}$ and $\hat{\sigma}^2$ are then obtained by setting $\theta = \hat{\theta}$ in (4.18), being θ included in the $\mathbf{R}_{00}$ and $\mathbf{c}_{-1}$ terms.

Although technically the random effects $\mathbf{b}_i$ are not parameters for the statistical model, they do behave in some ways like parameters and often we want to "estimate" their values [111]. The conditional modes of the random effects, evaluated at the conditional estimate of β, are the Best Linear Unbiased Predictors (BLUPs) of the $\mathbf{b}_i$, $i = 1, \ldots, M$. They can be evaluated as

$$\hat{\mathbf{b}}_i(\theta) = \mathbf{R}_{11(i)}^{-1}\left(\mathbf{c}_{1(i)} - \mathbf{R}_{10(i)}\hat{\beta}(\theta)\right) \tag{4.20}$$

Maximum likelihood estimates of variance components tend to produce underestimates. Many analysts prefer the REstricted Maximum Likelihood (REML) estimates for these quantities. There are several ways to define the REML estimation criterion. According to the one that refers to Laird and Ware in [97] and that provides a convenient computational form, this is

$$L_R(\theta, \sigma^2|\mathbf{y}) = \int L(\beta, \theta, \sigma^2|\mathbf{y}) d\beta$$

which, within a Bayesian framework, corresponds to assuming a locally uniform prior distribution for the fixed effects β and integrating them out of the likelihood. Using (4.17) in log-likelihood version, we have

$$l_R(\theta, \sigma^2|\mathbf{y}) = -\frac{N-p}{2}\log(2\pi\sigma^2) - \frac{\|\mathbf{c}_{-1}\|}{2\sigma^2} - \log\text{abs}|\mathbf{R}_{00}| + \sum_{i=1}^{M} \log\text{abs}\left(\frac{|\mathbf{\Delta}|}{|\mathbf{R}_{11(i)}|}\right)$$

This produces the conditional estimate $\hat{\sigma}_R^2(\theta) = \|\mathbf{c}_{-1}\|^2/(N-p)$ for σ^2, from which we obtain the profiled log-restricted-likelihood

$$\begin{aligned} l_r(\theta|\mathbf{y}) &= l_R(\theta, \hat{\sigma}^2|\mathbf{y}) \\ &= \text{const} - (N-p)\log\|\mathbf{c}_{-1}\| - \log abs|\mathbf{R}_{00}| + \sum_{i=1}^{M} \log abs\left(\frac{|\mathbf{\Delta}|}{|\mathbf{R}_{11(i)}|}\right) \end{aligned} \tag{4.21}$$

The evaluation of the restricted maximum likelihood estimates is done by optimizing the profiled log-restricted-likelihood (4.21) with respect to θ only, and using the resulting REML estimate $\hat{\theta}_R$ to obtain the REML estimate of σ^2 and $\hat{\sigma}_R^2(\hat{\theta}_R)$. Similarly, the REML estimated BLUPs of the random effects are obtained by replacing θ with $\hat{\theta}_R$.

An important difference between the likelihood function and the restricted likelihood function is that the former is invariant to one-to-one reparameterizations of the fixed effects (i.e., a change in the contrasts representing a categorical variable), while the latter is not. Changing the $\mathbf{X}_i$ matrices results in a change in $\log abs|\mathbf{R}_{00}|$ and a corresponding change in $l_R(\theta|\mathbf{y})$. As a consequence, LME models with different fixed effects structures fit using REML cannot be compared on the basis of their restricted likelihoods. In particular, likelihood ratio tests are not valid under these circumstances. For details

4.2.3 Classification and inference using LME models

Inference on the parameters of a linear mixed effects model usually relies on approximate distributions for the maximum likelihood estimators and the restricted maximum likelihood estimators derived from asymptotic results. In [121] Pinheiro shows that, under certain regularity conditions generally satisfied in practice, the maximum likelihood estimates in the general LME model are consistent and asymptotically normal. The approximate distributions of the maximum likelihood estimators in a LME model with Q levels of nesting are

$$\begin{gathered}\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \mathbf{R}_{00}^{-1} \mathbf{R}_{00}^{-T}\right) \\ \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_Q \\ \log\hat{\sigma} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta_1 \\ \cdots \\ \theta_Q \\ \log\hat{\sigma} \end{bmatrix}, \boldsymbol{I}^{-1}(\theta_1, \ldots, \theta_Q, \sigma^2) \right)\end{gathered} \tag{4.22}$$

where

$$\boldsymbol{I}(\theta_1, \ldots, \theta_Q, \sigma^2) = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1 \partial \theta_1^T} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1^T} & \cdots & \frac{\partial^2 l}{\partial \log\sigma \partial \theta_1^T} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 l}{\partial \theta_1 \partial \log\sigma} & \frac{\partial^2 l}{\partial \theta_2 \partial \log\sigma} & \cdots & \frac{\partial^2 l}{\partial^2 \log\sigma} \end{bmatrix}$$

and $l(\theta_1, \ldots, \theta_Q, \sigma^2)$ denotes the log-likelihood function profiled on the fixed effects, being $\boldsymbol{I}$ the empirical information matrix. $\mathbf{R}_{00}$ is defined as in (4.16). We use $\log\sigma$ in place of σ^2 in (4.22) to give an unrestricted parameterization for which the normal approximation tends to be more accurate. As shown by [121], the REML estimates in an LME model also are consistent and asymptotically normal, with approximate distributions identical to (4.22) but with l replaced by the log-restricted-likelihood l_R. In practice, the unknown parameters $\theta_1, \ldots, \theta_Q$ and σ^2 are replaced by their respective ML or REML estimates in the expressions for the approximate variance-covariance matrices in (4.22). Starting from these approximate distributions for the maximum likelihood estimates and REML estimates it is possible to produce hypothesis tests and confidence intervals for the LME model parameters, as shown in [122], Section 2.4.

In general, the greatest inferential effort about LME models is focused on the fixed effects and the variance components. Random effects are a sort of "residual noise" to be suitably taken into account, but which does not have any more appeal for inference purposes. Usually, graphical enquires are adopted a priori to test for the presence of the grouping factor effect on data, then a Normal random effect is included in the model, and finally Normality assumption is checked.

In our work, we look at random effects estimates from a different perspective. Although technically the random effects $\mathbf{b}_i$ are not parameters for the statistical model, they do behave in some ways like parameters and often we want to "estimate" their values as well as to use them for further analyses. When the assumptions made on random effects are not satisfied, in fact, this is often because there are some groups (individuals) whose effects are "more similar" than others. This leads to a sort of "clustering of clusters". In fact, our idea is to take advantage of these hints and to adopt mixed effects models as explorative tool for pointing out higher level of grouping on grouping factor, i.e., to investigate if grouping factor (for us, providers admitting patients affected by STEMI) can be thought as belonging to macro-groups with "similar behaviour". In other words, the point estimates of the random effects are telling us something about the providers' effect adjusted for the case mix it deals with, then the question is wheter any pattern of behaviour can be seen among different structures. This would be of great interest for people in charge with healthcare government,

to drive actions oriented to profile providers, to assess benchmarks of acceptability and so on. We generalize this idea also to a more complex models like generalized linear mixed effects models (see Section 4.3)

Starting from point estimates of random effects, is it possible to classify and to rank providers according to some acceptability criteria, labelling them as shown in [62], or to embody the pursuit of these labels within the estimate procedure, as done in [77].

4.3 Generalized linear parametric mixed effects models

Multilevel modelling is applied to logistic regression and other generalized linear models in the same way as linear regression. In fact, Generalized Linear Mixed Models (GLME Models) extend Generalized Linear Models (GLMs) by the inclusion of random effects in the predictor. In fact, a linear mixed model like (4.1) assumes that the relationship between the mean of the dependent variable $\mathbf{y}_i$ and the fixed and random effects can be modeled as a linear function, that the variance of $\mathbf{y}_i$ is not a function of the mean, and that the random effects follow a normal distribution. Any or all these assumptions may be violated for certain traits. A number of approaches have been taken to address the deficiencies of a linear mixed model (see for example [21]). Transformations have been used to stabilize the variance, to obtain a linear relationship, and to normalize the distribution. However the transformation needed to stabilize the variance may not be the same transformation needed to obtain a linear relationship. Further details on LME models with more general variance structure can be found in [73] and [122]. Anyway, all these options sidestep the issue that the linear mixed model is incorrect. It seems more reasonable to start with an appropriate model for the data and use an estimation procedure derived from that model. A generalized linear mixed model is a model which gives us extra flexibility in developing an appropriate model for the data. The generalized linear mixed model is the most frequently used random effects model in the context of discrete repeated measurements.

4.3.1 Model formulation for GLME models

Let y_{ij}, is the j-th outcome measured for cluster (subject) i, $i = 1,\ldots,M$, $j = 1,\ldots,n_i$ and $\mathbf{y}_i$ is the n_i-dimensional vector of all measurements available for cluster i. As introduced in Section 4.2, it is assumed that, conditionally on q-dimensional random effects $\mathbf{b}_i$, assumed to be drawn independently from the $\mathcal{N}(\mathbf{0},\Sigma)$, the outcomes y_{ij} are independent with densities belonging to the exponential family

$$f_{ij}(y_{ij}|\mathbf{b}_i,\beta,\tau) = \exp\left\{\tau\left[y_{ij}\zeta_{ij} - a(\zeta_{ij})\right] + c(y_{ij},\tau)\right\}$$

with $\mu_{ij} = \mathbb{E}[y_{ij}|\mathbf{b}_i]$ and $\eta_{ij} = h(\mu_{ij}) = \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i$ for a known link function $h(\cdot)$. Usually the canonical link function is chosen, i.e., such that $h(\mathbb{E}[y_{ij}|\mathbf{b}_i]) = \zeta_{ij}$, and since $\mathbb{E}[y_{ij}|\mathbf{b}_i] = a'(\zeta_{ij})$, we have $\zeta_{ij} = (a')^{-1}(\mathbb{E}[y_{ij}|\mathbf{b}_i])$. $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are respectively a p-dimensional and a q-dimensional vectors of known covariate values, and β is a p-dimensional vector of unknown fixed regression coefficients. Moreover, ζ_{ij} is the natural parameter of the exponential family, and τ the scale parameter. Finally, let $f(\mathbf{b}_i|\Sigma)$ be the density of the $\mathcal{N}(\mathbf{0},\Sigma)$ distribution for the random effects $\mathbf{b}_i$.
The hierarchical model formulation where the outcome is modeled conditionally on random effects, which are then modeled in an additional step, makes Bayesian methodology very appealing for fitting generalized linear mixed models. We will discuss this approach deeply later on for both LME Models and GLME Models in Chapter 5.

As already mentioned, random effects models can be fitted by maximization of the marginal likelihood, obtained by integrating out the random effects. The likelihood contribution of group i then becomes

$$f_i(\mathbf{y}_i|\beta,\Sigma,\tau) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i,\beta,\tau)\, f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i \tag{4.23}$$

from which the likelihood for β, Σ, and τ is derived as

$$\begin{aligned} L(\beta,\Sigma,\tau|\mathbf{y}_i) &= \prod_{i=1}^{M} f_i(\mathbf{y}_i|\beta,\Sigma,\tau) \\ &= \prod_{i=1}^{M} \int \prod_{j=1}^{n_i} f_{ij}(\mathbf{y}_{ij}|\mathbf{b}_i,\beta,\tau)\, f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i \end{aligned} \tag{4.24}$$

The key problem in maximizing (4.24) is the presence of M integrals over the q-dimensional random effects $\mathbf{b}_i$. In some special cases, these integrals can be worked out analytically. For example, as it has been shown in Section 4.2 on linear mixed models for outcomes belonging to the Gaussian family. In general, no analytic expressions are available for the integrals in (4.24) and numerical approximations are needed. There is a large statistical literature on various methods to do so. In Paragraph 4.3.2 we will just make an overview of the most frequently used ones, also implemented in commercially available software packages. For further discussions on these topics, see [142] and [143].

Although in practice one is usually primarily interested in estimating the parameters in the marginal distribution for $\mathbf{y}_i$, we are also deeply interested in obtaining estimates for the random effects $\mathbf{b}_i$ as well. They reflect between-subject variability, which makes them helpful for detecting special profiles (i.e., outlying groups) or groups with behaviour or patterns different from all the others. If groups are subjects, i.e., if we have repeated measures for each unit along time, this means to be interested in pointing out subjects evolving differently in time. Also, estimates for the random effects are needed whenever interest is in prediction of group (subject)-specific evolutions.

4.3.2 Estimation for GLME models

As mentioned before, the key problem in maximizing (4.24) is the presence of M integrals over the q-dimensional random effects $\mathbf{b}_i$. In some special cases, these integrals can be worked out analytically, but in general, no analytic expressions are available for them and numerical approximations are needed. These numerical approximations can be divided in those that are based on the approximation of the integrand, those based on an approximation of the data, and those that are based on the approximation of the integral itself.

Approximation of the integrand

When integrands are approximated, the goal is to obtain a tractable integral such that closed-form expressions can be obtained, making the numerical maximization of the approximated likelihood feasible. Several methods have been proposed, but basically all come down to Laplace-type approximations of the function to be integrated. The Laplace method [139] has been designed to approximate integrals of the form

$$I = \int e^{Q(\mathbf{b})} d\mathbf{b} \tag{4.25}$$

where $Q(\mathbf{b})$ is a known, unimodal, and bounded function of a q-dimensional variable $\mathbf{b}$. Let $\hat{\mathbf{b}}$ be the value of $\mathbf{b}$ for which Q is maximized. We then have that the second-order Taylor expansion of $Q(\mathbf{b})$ is of the form

$$Q(\mathbf{b}) \approx Q(\hat{\mathbf{b}}) + \frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T Q''(\hat{\mathbf{b}})(\mathbf{b}-\hat{\mathbf{b}}) \tag{4.26}$$

for $Q''(\hat{\mathbf{b}})$ equal to the Hessian of Q, evaluated at $\hat{\mathbf{b}}$. Replacing $Q(\mathbf{b})$ in (4.25) by its approximation in (4.26), we obtain

$$I \approx (2\pi)^{q/2} \left| -Q''(\hat{\mathbf{b}}) \right|^{-1/2} e^{Q(\hat{\mathbf{b}})} \tag{4.27}$$

Clearly, we can use this method when each integral in (4.24) can be written in the form (4.25), with suitable functions $Q(\mathbf{b})$. Note that the mode $\hat{\beta}$ of Q depends on the unknown parameters β, τ and Σ, such that in each iteration of the numerical maximization of the likelihood, $\hat{\beta}$ will be recalculated conditionally on the current values for the estimates for these parameters. The Laplace approximation will be exact when $Q(\mathbf{b})$ is a quadratic function of $\mathbf{b}$, i.e., if the integrands in (4.24) are exactly equal to normal kernels. More details can be found in [68].

Approximation of the data

A second class of approaches is based on a decomposition of the data into the mean and an appropriate error term, with a Taylor series expansion of the mean that is a non-linear function of the linear predictor. All methods in this class differ in the order of the Taylor approximation and/or in the point around which the approximation is expanded. More specifically, one considers the decomposition

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} = h^{-1}(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i) + \varepsilon_{ij} \tag{4.28}$$

in which $h^{-1}(\cdot)$ equals the inverse link function, and where the error terms have the appropriate distribution with variance equal to $\mathbb{V}ar(y_{ij}|\mathbf{b}_i)$. Now, let consider, for example, binary outcomes with the logistic natural link function and $\tau = 1$. One then has

$$\mu_{ij} = \mathbb{P}(y_{ij}=1) = \pi_{ij} = \frac{\exp(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i)}{1+\exp(\mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\mathbf{b}_i)}$$

and so ε_{ij} equals $1-\pi_{ij}$ with probability π_{ij} and equals $-\pi_{ij}$ with probability $1-\pi_{ij}$. Several approximations of the mean μ_{ij} in (4.28) can be considered.

The first one we discuss is a linear Taylor expansion of (4.28) around current estimates $\hat{\beta}$ and $\hat{\mathbf{b}}_i$ of the fixed effects and random effects, respectively. This yields

$$\begin{aligned} y_{ij} &\approx h^{-1}(\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}_i) \\ &+ h^{-1'}(\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}_i)\mathbf{x}_{ij}^T(\beta-\hat{\beta}) \\ &+ h^{-1'}(\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}_i)\mathbf{z}_{ij}^T(\mathbf{b}-\hat{\mathbf{b}}) + \varepsilon_{ij} \\ &= \hat{\mu}_{ij} + \mathbb{V}ar(\hat{\mu}_{ij})\mathbf{x}_{ij}^T(\beta-\hat{\beta}) + \mathbb{V}ar(\hat{\mu}_{ij})\mathbf{z}_{ij}^T(\mathbf{b}-\hat{\mathbf{b}}) + \varepsilon_{ij} \end{aligned}$$

where $\hat{\mu}_{ij}$ equals the current predictor $h^{-1}(\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}_i)$ for the conditional mean $\mathbb{E}[y_{ij}|\mathbf{b}_i]$. In vector notation it becomes

$$\mathbf{y}_i \approx \hat{\mu}_i + \hat{\mathbf{V}}_i\mathbf{X}_i(\beta-\hat{\beta}) + \hat{\mathbf{V}}_i\mathbf{Z}_i(\mathbf{b}_i-\hat{\mathbf{b}}_i) + \varepsilon_i$$

for appropriate design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$, and with $\hat{\mathbf{V}}_i$ equal to the diagonal matrix with diagonal entries equal to $\mathbb{V}ar(\hat{\mu}_{ij})$. Re-ordering the above expression yields

$$\mathbf{y}_i^* \equiv \hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\beta} + \mathbf{Z}_i\hat{\mathbf{b}}_i \approx \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i^* \tag{4.29}$$

for ε_i^* equal to $\hat{\mathbf{V}}_i^{-1}\varepsilon_i$, which still has mean zero. Note that (4.29) can be viewed as a linear mixed effects model for the pseudo data $\mathbf{y}_i^*$, with fixed effects β, random effects $\mathbf{b}_i$, and error terms ε_i^*. This immediately yields an algorithm for fitting the original generalized linear mixed effects model. Given starting values for the parameters β, Σ in the marginal likelihood, empirical Bayes estimates (see [143], Section 4.5) are calculated for $\mathbf{b}_i$, and pseudo data $\mathbf{y}_i^*$ are computed. Then, the approximate linear mixed model (4.29) is fitted, yielding updated estimates for β and Σ. These are then used to update the pseudo data and this whole scheme is iterated until convergence is reached. The resulting estimates are called *Penalized Quasi-Likelihood* (PQL) estimates because they can be obtained from optimizing a quasi-likelihood function which only involves first- and second-order conditional moments, augmented with a penalty term on the random effects. We refer to [19] for more details.

An alternative approximation is very similar to the PQL method, but is based on a linear Taylor expansion of the mean μ_{ij} in (4.28) around the current estimates $\hat{\beta}$ for the fixed effects and around $\hat{\mathbf{b}}_i = 0$ for the random effects. This yields very similar expressions as derived in the paragraph before, only is the current predictor $\hat{\mu}_{ij}$ now of the form $h^{-1}(\mathbf{x}_{ij}^T\hat{\beta})$, rather than $h^{-1}(\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}_i)$ as was the case before. The pseudo-data are now of the form $\mathbf{y}_i^* \equiv \hat{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \hat{\mu}_i) + \mathbf{X}_i\hat{\beta}$ and satisfy the approximate linear mixed effects model

$$\mathbf{y}_i^* \approx \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \varepsilon_i^* \tag{4.30}$$

Again, model fitting is done by iterating the calculation of the pseudo-data and the fitting of the approximate linear mixed model for these pseudo-data. The resulting estimates are called *Marginal Quasi-Likelihood* (MQL) estimates. As with the PQL estimates, they can be obtained by optimizing a quasi-likelihood function which only involves first- and second-order moments, but now evaluated in the marginal linear predictor $\mathbf{x}_{ij}^T\hat{\beta}$ rather than the conditional linear predictor $\mathbf{x}_{ij}^T\hat{\beta} + \mathbf{z}_{ij}^T\hat{\mathbf{b}}$. We refer to [19] and [52] for more details.

The essential difference between PQL and MQL is that the latter do not incorporate the random effects $\mathbf{b}_i$ in the linear predictor, but both methods are based on the same key idea and will, in general, have very similar properties. Obviously the accuracy of both approximations depends on the accuracy of the linear mixed effects model for the pseudo data $\mathbf{y}_i^*$. In each step of the iterative process $\prod_j f_{ij}(y_{ij}|\mathbf{b}_i, \beta, \tau)$ in (4.24) is replaced by the multivariate Normal density of $\mathbf{y}_i^*$. Note that

$$\begin{aligned}
\prod_j^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \beta, \tau) &= \exp\left\{\sum_j^{n_i} \tau\left[y_{ij}\zeta_{ij} - a(\zeta_{ij})\right] + \sum_j c(y_{ij}, \tau)\right\} \\
&= \exp\left\{\tau\left[\beta^T\sum_j^{n_i}\mathbf{x}_{ij}y_{ij} + \mathbf{b}_i^T\sum_j^{n_i}\mathbf{z}_{ij}y_{ij} - a(\zeta_{ij})\right] + \sum_j^{n_i} c(y_{ij}, \tau)\right\}
\end{aligned}$$

The sufficient statistics for β and $\mathbf{b}_i$ are $\sum_j \mathbf{x}_{ij}y_{ij}$ and $\sum_j \mathbf{z}_{ij}y_{ij}$, respectively. The approximation will be accurate whenever these sufficient statistics are approximately normally distributed, i.e., whenever the responses y_{ij} are "sufficiently" continuous and/or if the number n_i of measurements per group (subject) is sufficiently large. This explains why, as for the Laplace method, PQL and

MQL perform poorly in cases with binary repeated observations, with a relatively small number of repeated observations available for all groups.

Although similar in underlying key ideas, there are also some important differences between MQL and PQL. Obviously, MQL completely ignores the random effects variability in the linearization of the mean. Therefore, it will only provide a reasonable approximation when the variance of the random effects is (very) small. Even with increasing numbers of measurements per cluster, the bias in MQL remains. This is not the case for PQL which can be shown to be consistent when both the number of subjects as well as the number of measurements per subject approach infinity, even for binary outcomes. One way to improve the accuracy of the approximations is the inclusion of a second-order term in the Taylor expansions. This leads to the PQL2 and MQL2 methods, discussed, for example, in [53]. Finally, besides using higher orders in the Taylor expansions, some authors have advised the introduction of bias correction terms (for example [20]). Because the linearizations in the PQL and the MQL methods lead to linear mixed effects models, the implementation of these procedures is often based on feeding updated pseudo data into software for the fitting of linear mixed effects models. However, it should be emphasized that outputs resulting from such fittings, which are sometimes reported intermediately, should be interpreted with great care. For example, reported (log-)likelihood values correspond to the assumed normal model for the pseudo data and should not be confused with (log-)likelihood for the generalized linear mixed effects model for the actual data at hand. Also, as discussed in the previous sections, fitting of linear mixed effects models can be based on maximum likelihood (ML) as well as restricted maximum likelihood (REML) estimation. Hence, within the PQL and MQL frameworks, both methods can be used for the fitting of the linear model to the pseudo data, yielding (slightly) different results.

Approximation of the integral

Especially in cases where the above approximation methods fail, approximations to the integral, i.e., numerical integration, proves to be very useful. Of course, a wide toolkit of numerical integration tools, available from the optimization literature, can be used. Several of those have been implemented in various software tools for generalized linear mixed effects models. A general class of quadrature rules selects a set of abscissas and constructs a weighted sum of function evaluations over those. In the particular context of mixed effects models, so called adaptive quadrature rules can be used, were the numerical integration is centered around the Empirical Bayes estimates (see [143], Section 14.4) of the random effects, and the number of quadrature points is then selected in terms of the desired accuracy. To illustrate the main ideas, we consider Gaussian and adaptive Gaussian quadrature, designed for the approximation of integrals of the form

$$\int f(\mathbf{t})\phi(\mathbf{t})d\mathbf{t} \tag{4.31}$$

for a known function $f(\mathbf{t})$ and for $\phi(\mathbf{t})$ the density of the multivariate standard Normal distribution. We will therefore first standardize the random effects such that they get the identity covariance matrix. Let $\mathbf{s}_i = \Sigma^{-1/2}\mathbf{b}_i$. We then have that $\mathbf{s}_i$ is normally distributed with mean $\mathbf{0}$ and covariance $\mathbb{I}$, and the linear predictor becomes $\eta_{ij} = \mathbf{x}_{ij}^T\beta + \mathbf{z}_{ij}^T\Sigma^{1/2}\mathbf{s}_i$. Hence, the variance components in Σ have been moved to the linear predictor. The likelihood contribution for group i equals

$$\begin{aligned} f_i(\mathbf{y}_i|\beta,\Sigma,\tau) &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i,\beta,\tau)\, f(\mathbf{b}_i|\Sigma)d\mathbf{b}_i \\ &= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{s}_i,\beta,\Sigma,\tau)\, f(\mathbf{s}_i)d\mathbf{s}_i \end{aligned} \tag{4.32}$$

Obviously, (4.32) is of the form (4.31) as required to apply (adaptive) Gaussian quadrature. This class of methods, approximates integrals in the form of (4.31) by the weighted sum

$$\int f(\mathbf{t})\phi(\mathbf{t})dt \approx \sum_{k=1}^{K} w_k f(\mathbf{t}_k).$$

K here is the order of the approximation. The higher K, the more accurate the approximation is. Further, the so called nodes (or quadrature points) t_k are solutions to the K-th order Hermite polynomial, while the w_k are appropriately chosen weights. Further discussion on Gaussian quadrature, as well as methods for numerical integration of multivariate integrals are no directly concerned with the main aim of this thesis, then we remind to [143] for further details on this topic.

4.3.3 Inference for GLME models

Because the fitting of generalized linear mixed effects models is based on maximum likelihood principles, inferences for the parameters are readily obtained from classical maximum likelihood theory. Indeed, assuming the fitted model is appropriate, the obtained estimators are asymptotically normally distributed with the correct values as means, and with the inverse Fisher information matrix as covariance matrix. Hence, Wald-type tests, comparing standardized estimates to the standard normal distribution can easily be performed. Composite hypotheses can be tested using the more general formulation of the Wald statistic which is a standardized quadratic form, which is then compared to the chi-squared distribution. Alternatively, likelihood ratio and score tests can be used as well. As discussed in the previous section, the parameters in generalized linear mixed effects models are often estimated by fitting linear mixed effects models to pseudo-data. Therefore, precision estimates for the fixed effects and for the random effects are often calculated using linear mixed effects model methodology, yielding for example Z-, t- and F-tests for the fixed effects. A detailed discussion of inferential tool to be applied and used in the LME Models framework can be found in [142], Chapter 6.

Finally, when interest is also in inference for some of the variance components in Σ, classical asymptotic Wald, likelihood ratio, and score tests can be used, as long as the hypotheses to be tested are not on the boundary of the parameter space. For example, suppose one wishes to test whether the variance of a single random effect (let us denote it as σ_b^2) in a generalized linear mixed effects model equals zero. This is of crucial interest, since it assesses if the random effect should or not be included in the model. In this case, one has to test the null-hypothesis $H_0 : \sigma_b^2 = 0$ versus the alternative $H_1 : \sigma_b^2 > 0$, then rejecting H_0 allows for the presence of the random effects. Obviously, the null-hypothesis is on the boundary of the parameter space $\sigma_b^2 \geq 0$. None of the classical Wald, likelihood ratio, or score tests are still valid. This can most easily be seen from considering the classical Wald test that would be based on the standard Normal approximation to the standardized maximum likelihood estimate $\hat{\sigma}_b^2$.

Last but not least, as already mentioned before, the inference for binary data is often influenced by the sample balancing. If we deal we unbalanced samples in the overall population and/or within the single groups, we could get biased estimates of parameters of interest which determine uncorrect inference. A solution that involves modification of iterative algorithms generating estimates for GLME models is proposed in [51], but since this is the case we deal with in the context of inference and prediction of survival in patient affected by STEMI (for whom in-hospital empirical survival probability is around 94%), this topic is treated in detail in Section 4.7.

4.4 Linear nonparametric mixed effects models

At the start of the previous section, we said that, in linear mixed effects models, assumptions of linear relationship between the mean of the dependent variable and fixed and random effects, constant variance, and normality are questionable. Then in the previous section attention has been focused on pointing out more general and flexible models to solve problem mainly due to the lack of linearity assumption. Often happens that the Normality assumption on the random effect distribution, or in general parametric assumptions on it, are very restrictive and unrealistic for modelling real data. We mentioned a different approach to the analysis data not satisfying such assumption in Paragraph 4.2.3. In this section we then introduce a technique for modelling random effect distribution in a nonparametric way for linear and generalized linear mixed effects models, then we will move in the nex section to the most general case of parametric and nonparametric modelling of random effects within non linear models.

In [4], interest is focused on an EM algorithm for Non Parametric Maximum Likelihood (NPML) estimation in generalized linear mixed effects models with variance component structure, which provides an alternative analysis to approximate MQL and PQL estimates, mentioned in Section 4.3. The algorithm is initially derived as a form of Gaussian quadrature assuming a normal mixing distribution, but with only slight variation it can be used for a completely unknown distribution of the random effects, giving a straightforward method for the fully NPML estimation of this distribution. This is because the ML estimates of the GLME models parameters can be sensitive to the specification of a parametric form for the mixing distribution of the radnom effects. This can produce substantial computational saving compared with full numerical integration over a specified parametric distribution for the random effects parameters.

In the following we focus our attention on grouped data, such that the statistical unit i is no more the grouping factor (as for longitudinal data), but belongs to a group j, with $j = 1,\dots,J$. For easying the exposition, we begin with the simple two level model for a structure with upper- or second-level sampling units indexed by $j = 1,\dots,J$ and lower or first-level sampling units indexed by i sampled within each upper-level unit, where $i = 1,\dots,n_j$. On each first-level unit, we measure or record a response y_{ij} , and we have explanatory variables z, which can be measured at both upper (z_j) and lower (z_{ij}) levels. We want to represent the distribution of the response y by an exponential family member, with a link function and linear predictor involving the explanatory variables at both levels and perhaps their cross-level interactions. The nested structure of the responses y_{ij} induces an intraclass correlation between the lower-level responses on the same upper-level unit. A natural way of representing this common variation is by adding a common unobserved random effect to the linear predictor for each lower-level unit in the same upper-level unit. Thus, the common variation is modeled as an extra unobserved variable on the same scale as the linear predictor. If the distribution of this random effect belongs to the exponential family distribution, we already saw that then maximum likelihood (ML) is straightforward in principle from the marginal distribution of the observed data [98]. An appealing approach would be to assume a common distribution for the random effects across the exponential family; an obvious choice is the normal $\mathcal{N}(0,\sigma^2)$ distribution [19]. This is especially natural for link functions giving an unbounded space for the linear predictor. However, exponential family models other than the normal with a normal random effect have been difficult and slow to fit by ML because the resulting likelihood does not have a closed form.

Anyway, the main disadvantage of any approach using a specified parametric form for the mixing distribution of the unobserved random effects is the possible sensitivity of the conclusions to

this specification. This difficulty can be avoided by NPML estimation of the mixing distribution concurrently with the structural model parameters; the NPML estimate is known to be a discrete distribution on a finite number of mass points (it is discussed and proved in [93], [96] and [100] among others). An example of this approach, in the framework of a single level overdispersion model can be found in [25].

Finding the NPML estimate is widely regarded as computationally intensive, the particular difficulty being the location of the mass points. In [4], an exposition of how estimating both the mass-point locations c_k and the masses ω_k in a very straightforward way by ML within the framework of a finite mixture of GLMs is given for the two-level variance case, allowing the straightforward full NPML estimation of the mixing distribution. Convergence of the EM algorithm can then become very slow, as information in the data about the mixing distribution might be very limited, but the algorithm is easily programmed.

4.4.1 From parametric to nonparametric GLME models

Concerning the model specifications, for $i = 1\ldots,n_j$, $j = 1,\ldots,J$ and $\sum_j n_j = N$ let y_{ij} be from an exponential family distribution $f(y_{ij}|\zeta_{ij})$ with canonical parameter ζ_{ij}, mean μ_{ij} and explanatory variables $\mathbf{X} = (x_{ij})$, related to μ_{ij} through a link function $\eta_{ij} = h(\mu_{ij})$ with linear predictor η_{ij}. Here the $\mathbf{X}$ matrix is understood to include both upper- and lower-level explanatory variables. In the extension to mixed effect models, we have an unobserved common random effect b_j for each lower-level unit in the j-th upper-level unit, the b_j being initially assumed independently Normally distributed $b_j \sim \mathcal{N}(0,\sigma^2)$, and conditionally on b_j, the y_{ij} have independent GLMs with linear predictor $\eta_{ij} = \beta^T x_{ij} + b_j$. The random effect is modeled as acting on the same scale as the linear predictor. The likelihood is then

$$L(\beta,\sigma) = \prod_j \int \prod_i f(y_{ij}|\beta,\sigma,b_j)\, f(b_j) db_j$$

where $f(b)$ is the Normal density function. Because the integral does not have a closed form except for y normal, we approximate it by Gaussian quadrature: we replace the integral over the normal b_j by a finite sum over K Gaussian quadrature mass points c_k with masses ω_k. The likelihood is then

$$L(\beta,\sigma) = \prod_j \sum_{k=1}^{K} \omega_k \prod_i f(y_{ij}|\beta,\sigma,c_k) \tag{4.33}$$

The likelihood is thus (approximately) the likelihood of a finite mixture of exponential family densities with known mixture proportions ω_k at known mass points c_k, with the linear predictor for the ij-th observation in the k-th mixture component being

$$\eta_{ijk} = \beta^T x_{ij} + c_k.$$

This is inherently of interest because the NPML estimate of the mixing distribution is known to be a discrete distribution on a finite number of mass points. In the following we consider the joint estimation of β, the ω_k and the mass points c_k, but for the moment consider the latter quantities as fixed.

Now, the log-likelihood version of (4.33) is

$$l(\beta,\sigma) = \sum_j \log \sum_k \omega_k f_{jk} \tag{4.34}$$

being $f_{jk} = \prod_i f_{ijk}$, $f_{ijk} = f(y_{ij}|\beta,\sigma,c_k) = \exp\{y_{ij}\zeta_{ijk} - a(\zeta_{ijk}) + c(y_{ij})\}$ and $\eta_{ijk} = h(\mu_{ijk}) = \beta^T x_{ij} + c_k$. Then

$$\frac{\partial l}{\partial \beta} = \sum_j \frac{\sum_k \omega_k f_{jk} \frac{\partial \log f_{jk}}{\partial \beta}}{\sum_k \omega_k f_{jk}} = \sum_j \sum_i \sum_k w_{jk} s_{ijk}(\beta) \tag{4.35}$$

where w_{jk} is the posterior probability that observation y_{ij} comes from component k, i.e.,

$$w_{ik} = \frac{\omega_k f_{jk}}{\sum_l \omega_l f_{jl}}$$

and $s_{ijk}(\beta)$ is the β component of the score for observation ij in component k. Thus, c_k becomes another observable variable in the regression. Equating the score to zero gives likelihood equations that are simple weighted sums of those for an ordinary GLM with weights w_{jk}; alternately solving these equations for given weights w_{jk} and updating these weights from the current parameter estimates is an EM algorithm.

Because the model assumption for unobservable random variables cannot be directly assessed, we consider as a preferable modelling strategy the NPML estimation of the mixing distribution, together with the GLM parameters. The aim is not to estimate this distribution, but to avoid possibly misleading inferences from an inappropriate and unverifiable model assumption. Then we now treat the masses and mass points as unknown parameters; the number K of mass points is also unknown but is treated as fixed and sequentially increased until the likelihood is maximized. So in the linear predictor $\eta_{ijk} = \beta^T x_{ij} + c_k$, c_k are now considered as intercept parameters for the k-th component. Differentiating the log likelihood with respect to ω_k and using $\omega_K = 1 - \sum_1^{K-1} \omega_k$, we have directly

$$\frac{\partial l}{\partial \omega_k} = \sum_j \frac{f_{jk} - f_{jK}}{\sum_l \omega_l f_{jl}} = \sum_j \left\{ \frac{w_{jk}}{\omega_k} - \frac{w_{jK}}{\omega_K} \right\}$$

Equating this to zero gives

$$\hat{\omega}_k = \sum_j \frac{w_{jk}}{N} \tag{4.36}$$

a standard mixture ML result. The same EM algorithm applies with the additional calculation in each M-step of the estimate of ω_k from the weights. A distinctive feature of the weights is that they are calculated for each upper-level unit in the E step but applied to all lower-level units in this upper level unit in the M-step.

4.5 Nonlinear parametric mixed effects models

In this section, a quick overview of methods for dealing with theory and computations concerned with Nonlinear Mixed Effects (NLME) Models is presented (a deeper tractation can be found in [46] and [122]). A general formulation of a NLME model is given in the case of single level of grouping.

When choosing a regression model to describe how a response variable varies with covariates, one always has the option of using models that are linear in the parameters. By increasing the order of the model, one can get increasingly accurate approximations to the true, usually nonlinear, regression function, within the observed range of the data. In general, we deal with empirical models that are based only on the observed relationship between the response and the covariates and do not include any theoretical considerations about the underlying mechanism producing the data. Nonlinear models, on the other hand, are often mechanistic, i.e., based on a model for the mechanism

producing the response. As a consequence, the model parameters in a nonlinear model generally have a natural physical interpretation and generally uses fewer parameters than a competitor linear model, giving a more parsimonious description of the data. Even when derived empirically, nonlinear models usually incorporate known, theoretical characteristics of the data, such as asymptotic behaviour and monotonicity. This is typical of growth curves, for example, which are often employed in biomedical applications for description of numbers of different phenomena.

NLME models extend LME models by allowing the regression function to depend nonlinearly on fixed and random effects. Because of its greater flexibility, an NLME model is generally more interpretable and parsimonious than a competitor LME model. Also, the predictions obtained from an NLME model extend more reliably outside the observed range of the data. The greater flexibility of NLME models does not come without cost, however. Because the random effects are allowed to enter the model nonlinearly, the marginal likelihood function, obtained by integrating the joint density of the response and the random effects with respect to the random effects, does not have a closed-form expression, as in the LME model. As a consequence, an approximate likelihood function needs to be used for the estimation of parameters, leading to more computationally intensive estimation algorithms and to less reliable inference results. We won't enter into details of this topic, but several different options can be considered for numerical approximations of such likelihoods. An important practical difference between NLME and LME models is that the former require starting estimates for the fixed effects coefficients. Anyway, there are far more similarities than differences between LME and NLME models. Both models are used with grouped data and serve the same purpose: to describe a response variable as a function of covariates, taking into account the correlation among observations in the same group. Random effects are used to represent within-group dependence in both LME and NLME models, and the assumptions about the random effects and the within-group errors are identical in the two models.

4.5.1 Theory and computational methods of NLME models

Nonlinear mixed effects models are mixed effects models in which some, or all, of the fixed and random effects occur nonlinearly in the model function. They can be regarded either as an extension of linear mixed effects models in which the conditional expectation of the response given the random effects is allowed to be a nonlinear function of the coefficients, or as an extension of nonlinear regression models for independent data in which random effects are incorporated in the coefficients to allow then to vary by group, thus inducing correlation within the groups.

We will now focus on a basic NLME model, i.e., a nonlinear model with a single-level of grouping and classical assumption on random effects, and on already existing estimation methods for it (see for example [145]). Generalizations of this framework towards both relaxation of assumption on random effect distribution and more complex structure for modelling the error terms can be found in [122].

By far the most common application of NLME models is for repeated measures data (i.e., the individual i turns back to be the grouping factor), and in particular for longitudinal data, which can be thought of as a hierarchical models. At one level the j-th observation on the i-th group is modeled as

$$y_{ij} = g(\phi_{ij}, v_{ij}) + \varepsilon_{ij} \qquad i = 1,\ldots,M \quad j = 1,\ldots,n_i \tag{4.37}$$

where M is the number of groups, n_i is the number of observations on the i-th group, g is a general, real-valued, differentiable function of a group specific parameter vector ϕ_{ij} and a covariate vector $\mathbf{v}_{ij}$, and ε_{ij} is a normally distributed within-group error term. The function g is nonlinear in at least

one component of the group-specific parameter vector ϕ_{ij} , which is modeled as

$$\phi_{ij} = \mathbf{A}_{ij}\beta + \mathbf{B}_{ij}\mathbf{b}_i \qquad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Sigma) \tag{4.38}$$

where β is a p-dimensional vector of fixed effects and β_i is a q-dimensional random effects vector associated with the i-th group (not varying with j) with variance-covariance matrix Σ. The matrices $\mathbf{A}_{ij}$ and $\mathbf{B}_{ij}$ are of appropriate dimensions and depend on the group and possibly on the values of some covariates at the j-th observation. This assumption allows the incorporation of "time varying" covariates in the fixed effects or the random effects for the model. It is assumed that observations corresponding to different groups are independent and that the within-group errors ε_{ij} are independently distributed as $\mathcal{N}(0, \sigma^2)$ and independent of the $\mathbf{b}_i$. The assumption of independence and homoscedasticity for the within-group errors can be relaxed (see [122] for a deeper tractation of this topic). The general model in (4.37) and (4.38) can be written in matrix form as

$$\mathbf{y}_i = \mathbf{g}_i(\phi_i, \mathbf{v}_i) + \varepsilon_i \qquad \phi_i = \mathbf{A}_i\beta + \mathbf{B}_i\mathbf{b}_i \tag{4.39}$$

for $i = 1, \ldots, M$, where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \phi_i = \begin{bmatrix} \phi_{i1} \\ \vdots \\ \phi_{in_i} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix}, \mathbf{g}_i(\phi_i, \mathbf{v}_i) = \begin{bmatrix} g(\phi_{i1}, v_{i1}) \\ \vdots \\ g(\phi_{in_i}, v_{in_i}) \end{bmatrix},$$

$$\mathbf{v}_i = \begin{bmatrix} v_{i1} \\ \vdots \\ v_{in_i} \end{bmatrix}, \mathbf{A}_i = \begin{bmatrix} A_{i1} \\ \vdots \\ A_{in_i} \end{bmatrix}, \mathbf{B}_i = \begin{bmatrix} B_{i1} \\ \vdots \\ B_{in_i} \end{bmatrix} \tag{4.40}$$

This framework can be straightforwardly generalized to the case of grouped data with multiple, nested random effects, as shown in [122], Section 7.1.

Different methods have been proposed to estimate the parameters of NLME models. Here we consider only the case of likelihood-based methods. Because the random effects are unobserved quantities, maximum likelihood estimation in mixed effects models is based on the marginal density of the responses $\mathbf{y}$, which is calculated as

$$f(\mathbf{y}|\beta, \sigma^2, \Sigma) = \int f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2) f(\mathbf{b}|\Sigma) d\mathbf{b} \tag{4.41}$$

where $f(\mathbf{y}|\beta, \sigma^2, \Sigma)$ is the marginal density of $\mathbf{y}$, $f(\mathbf{y}|\mathbf{b}, \beta, \sigma^2)$ is the conditional density of $\mathbf{y}$ given the random effects $\mathbf{b}$, and the marginal distribution of $\mathbf{b}$ is $f(\mathbf{b}|\Sigma)$. For NLME model in (4.37), expressing the random effects variance-covariance matrix in terms of the precision factor $\mathbf{\Delta}$, so that $\frac{\Sigma^{-1}}{1/\sigma^2} = \mathbf{\Delta}^T\mathbf{\Delta}$, provides the marginal density

$$f(\mathbf{y}|\beta, \sigma^2, \mathbf{\Delta}) = \frac{|\mathbf{\Delta}|}{(2\pi\sigma^2)^{(N+M_q)/2}} \prod_{i=1}^{M} \int \exp\left\{ \frac{\|\mathbf{y} - \mathbf{g}_i(\beta, \boldsymbol{b_i})\|^2 + \|\mathbf{\Delta}\mathbf{b}_i\|^2}{-2\sigma^2} \right\} d\mathbf{b}_i \tag{4.42}$$

where $\mathbf{g}_i(\beta, \mathbf{b}_i) = \mathbf{g}_i[\phi_i(\beta, \mathbf{b}_\mathbf{i}), \mathbf{v}_i]$. Because the model function g can be nonlinear in the random effects, the integral in (4.41) generally does not have a closed-form expression. To make the numerical optimization of the likelihood function a tractable problem, different approximations to (4.41) have been proposed. Some of these methods consist of taking a first-order Taylor expansion of the model function g around the expected value of the random effects, or around the conditional (on $\mathbf{\Delta}$)

modes of the random effects. Gaussian quadrature rules have also been used. We describe briefly the method for approximating the likelihood function in the NLME model proposed by Lindstrom and Bates in [101], called LME approximation in [122]. It is the basis of the estimation algorithm currently implemented in the `nlme` function of `R`. Other methods can be found in [122], Paragraph 7.2.1.

The estimation algorithm described in [101] alternates between two steps, a Penalized Nonlinear Least Squares (PNLS) step, and a Linear Mixed Effects (LME) step, as described below. We will consider only the alternating algorithm for the single level NLME model in (4.41). In the PNLS step, the current estimate of $\mathbf{\Delta}$ (the precision factor) is held fixed, and the conditional modes of the random effects $\mathbf{b}_i$ and the conditional estimates of the fixed effects β are obtained by minimizing a penalized nonlinear least squares objective function

$$\sum_{i=1}^{M} \left[\|\mathbf{y}_i - \mathbf{g}_i(\beta, \mathbf{b}_i)\|^2 + \|\mathbf{\Delta}\mathbf{b}_i\|^2 \right] \tag{4.43}$$

The LME step updates the estimate of $\mathbf{\Delta}$ based on a first-order Taylor expansion of the model function g around the current estimates of β and the conditional modes of the random effects $\mathbf{b}_i$, which we will denote by $\hat{\beta}^{(w)}$ and $\hat{\mathbf{b}}_i^{(w)}$ respectively. Letting

$$\begin{aligned} \hat{\mathbf{X}}_i^{(w)} &= \frac{\partial \mathbf{g}_i}{\partial \beta^T}\Big|_{\hat{\beta}_i^{(w)}, \hat{\mathbf{b}}_i^{(w)}} \qquad \hat{\mathbf{Z}}_i^{(w)} = \frac{\partial \mathbf{g}_i}{\partial \mathbf{b}_i^T}\Big|_{\hat{\beta}_i^{(w)}, \hat{\mathbf{b}}_i^{(w)}} \\ \hat{\mathbf{w}}_i^{(w)} &= \mathbf{y} - \mathbf{g}_i(\hat{\beta}_i^{(w)}, \hat{\mathbf{b}}_i^{(w)}) + \hat{\mathbf{X}}_i^{(w)}\hat{\beta}^{(w)} + \hat{\mathbf{Z}}_i^{(w)}\hat{\mathbf{b}}_i^{(w)} \end{aligned}$$

the approximate log-likelihood function used to estimate $\mathbf{\Delta}$ is

$$\begin{aligned} l_{LME}(\beta, \sigma^2, \mathbf{\Delta}|\mathbf{y}) &= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{M}\{\log|\Sigma_i(\mathbf{\Delta})| \\ &\quad + \sigma^{-2}\left[\hat{\mathbf{w}}_i^{(w)} - \hat{\mathbf{X}}_i^{(w)}\beta\right]^T \Sigma_i^{-1}(\mathbf{\Delta})\left[\hat{\mathbf{w}}_i^{(w)} - \hat{\mathbf{X}}_i^{(w)}\beta\right]\Big\} \end{aligned}$$

where $\Sigma_i^{-1}(\mathbf{\Delta}) = \mathbb{I} + \hat{\mathbf{X}}_i^{(w)}\mathbf{\Delta}^{-1}\mathbf{\Delta}^{-T}\hat{\mathbf{X}}_i^{(w)^T}$. As we saw in Paragraph 4.2.2, one can express the optimal values of β and σ^2 as a function of $\mathbf{\Delta}$ and work with the profiled likelihood, greatly simplifying the optimization problem. In [101] a restricted maximum likelihood estimation for $\mathbf{\Delta}$ is proposed, consisting of replacing the log-likelihood in the LME step of the alternating algorithm by the log-restricted likelihood

$$\begin{aligned} l^R_{LME}(\sigma^2, \mathbf{\Delta}|\mathbf{y}) &= l_{LME}(\hat{\beta}(\mathbf{\Delta}), \sigma^2, \mathbf{\Delta}|\mathbf{y}) - \frac{1}{2}\sum_{i=1}^{M}\{\log|\Sigma_i(\mathbf{\Delta})| \\ &\quad + \sigma^{-2}\left[\hat{\mathbf{w}}_i^{(w)} - \hat{\mathbf{X}}_i^{(w)}\beta\right]^T \Sigma_i^{-1}(\mathbf{\Delta})\left[\hat{\mathbf{w}}_i^{(w)} - \hat{\mathbf{X}}_i^{(w)}\beta\right]\Big\} \end{aligned}$$

More details on maximum likelihood estimation in nonlinear mixed effect models can be found in [94].

4.6 Nonlinear nonparametric mixed effect models

We saw in the previous sections that nonlinear mixed effects models are mixed effects models in which at least one of the fixed or random effects appears nonlinearly in the model function. NLME

models are routinely and increasingly used in several biomedical applications, especially in population pharmacokinetics, pharmacodynamic, immune cells reconstruction and epidemiological studies (examples are [32], [34], [77], [132]). In these fields, statistical modelling based on NLME models takes advantage of tools that allow to distinguish overall population effects from drugs effects or unit specific influence. Indeed, mixed effects models address well these issues, since they include parameters associated with the entire population (fixed effects) and subject/group specific parameters (random effects). For this reason they are able to describe the dynamics of the phenomenon under investigation, even in presence of high between subjects variability. When the random effects represent a deviation from the common dynamic of the population, mixed effects models provide both estimates for the entire population's model and for each subject's one. We will now assume random effects to have a different meaning, i.e., to describe the common dynamic of different groups of subjects. In this framework, mixed effects models provide only estimates for each group-specific model. Thanks to this property, it will be possible to consider mixed effects models as an unsupervised clustering tool for longitudinal data and repeated measures, as mentioned in Paragraph 4.2.3. For this reason we focus our attention on the estimation of the distribution of the random effects $\mathscr{P}^*$.

A wide literature exists for parametric modelling of random effects distribution in linear and non linear mixed effects models, as discussed in the previous sections. In this framework, Maximum Likelihood (ML) estimators are generally preferred because of their consistency and efficiency. However, due to the non linearity of the likelihood, we are not always able to provide explicitly the parameter estimators. In summary, parametric models are widely used, but they rely on a normality assumption which may be too restrictive. In practice, this assumption is often checked using the empirical distribution of random effects' empirical Bayes estimates. Unfortunately, when data are sparse, this method is unreliable. Moreover, when the number of measurements for unit is small, predictions for random effects are strongly influenced by the parametric assumptions. For these reasons nonparametric (NP) framework, which allow $\mathscr{P}^*$ to live in an infinite dimensional space, is attractive. The discreteness of optimal nonparametric distribution enables to overcome some technicalities due to the numerical integration of likelihood functions conditioned to continuous random effects density functions (see for example [146], where a Laplace approximation for nonlinear random effects marginal distributions is introduced).

In [7] and [144], different nonparametric methods are compared with usually adopted parametric ones, since in literature several nonparametric methods have been proposed but their use is limited. That's because their practical and theoretical properties are unclear and they have a reputation for being computationally expensive. Nevertheless, nonparametric methods seem to be very useful when data are sparse. What we would like to do in the next paragraph, is to point out a new method for dealing at the same time with nonlinearities and relaxation of hypotheses on random-effects distribution inspired by works like [95], as introduced in [11] and [12].

4.6.1 NLNPEM: unsupervised classification in nonlinear nonparametric framework using random effects

We are interest in proposing a novel estimation method for nonlinear nonparametric mixed effects models, aimed at unsupervised classification. The proposed method is an iterative algorithm that alternates a nonparametric EM step and a nonlinear Maximum Likelihood step. We will call it NonLinear NonParametric Expectation Maximization algorithm (briefly NLNPEM).

We consider the following NLME model for longitudinal data, which is a particulr case of (4.37):

$$\begin{array}{ll} \mathbf{y}_i = g(\beta, \mathbf{b}_i, \mathbf{t}) + \boldsymbol{\varepsilon}_i & i = 1, \ldots, M \\ \varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n) & \text{i.i.d.} \end{array} \tag{4.44}$$

where $\mathbf{y}_i \in \mathbb{R}^n$ is the response variable evaluated at times $\mathbf{t} \in \mathbb{R}^n$ and g is a general, real-valued and differentiable function with $p+q$ parameters. Each parameter of g is treated either as fixed or as random. Fixed effects are parameters associated with the entire population whereas random effects are subject-specific parameters that allow to identify clusters of subjects. $\beta \in \mathbb{R}^p$ is a vector that contain all fixed effects and $\mathbf{b}_i \in \mathbb{R}^q$ is the vector for the i-th subject random effects. The function g is non linear at least in one component of the fixed or random effects. The errors ε_{ij} are associated with the j-th measurement of the i-th longitudinal data. They are normally distributed, independent between different subjects and independent within the same subject. In general, the proposed method could also take account of a different number of observations, located at different times, for different subjects (i.e., n_i not necessarily is equal to $n \; \forall i = 1, \ldots, M$). In (4.44) we chose not to consider this case in order to ease the notation, but the generalization is straightforward.

Usually random effects are assumed to be Normal distributed, $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$, with unknown parameters that, together with β and σ, can be estimated through methods based on the likelihood function. In this parametric framework the maximum likelihood estimators are generally favored by their statistical properties, i.e., consistency and efficiency. Nevertheless the parametric assumptions could be too restrictive to describe highly heterogeneous or grouped data, so it might be necessary to move to a non parametric approach. In our case, we assume $\mathbf{b}_i$, for $i = 1, ..., M$, independent and identically distributed according to a probability measure $\mathscr{P}^*$. Looking for the ML estimator $\hat{\mathscr{P}}^*$ of $\mathscr{P}^*$ in the space of all probability measures on $\mathbb{R}^q$, the discreteness theorem proved in [100], states that $\hat{\mathscr{P}}^*$ is a discrete measure with at most M support points. Therefore the ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_1, \ldots, \mathbf{c}_N)$, where $N \leq M$ and $\mathbf{c}_l \in \mathbb{R}^q$, and a set of weights $(\omega_1, \ldots, \omega_N)$, where $\omega_l \geq 0$ and $\sum_{l=1}^N \omega_l = 1$. As mentioned above, in this paper we propose an algorithm for the joint estimation of β, M, $(\mathbf{c}_1, \ldots, \mathbf{c}_N)$ and $(\omega_1, \ldots, \omega_N)$ in the non linear framework of model (4.44). The estimation of fixed effects β and variance σ^2 is performed through the maximization of the restricted likelihood:

$$L(\beta, \sigma^2 | \mathbf{y}) = f(\mathbf{y} | \beta, \sigma^2) = \sum_{l=1}^{N} \omega_l \frac{1}{(2\pi\sigma^2)^{(nM)/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{M} \sum_{j=1}^{n} \left(y_{ij} - g(\beta, \mathbf{c}_l, t_j)\right)^2}.$$

Notice that the number of support points N is estimated by the algorithm as well and we do not have to fix it a priori. Since we don't have to specify a priori the number of support points and in consequence the number of groups, the nonparametric mixed effects model could be interpreted as an unsupervised clustering tool for longitudinal data. This tool could be very useful in order to identify the groups of subjects to be used in the analysis.

The algorithm proposed for the estimation of the parameters of model (4.44) arises from the framework described in [129], and alternates two steps. The first one is a nonparametric EM step whereas the second one is a non linear maximum-likelihood step. The nonparametric EM step estimates the discrete q-dimensional distribution $(\mathbf{c}, \boldsymbol{\omega})$ of the random effects $\mathbf{b}_i$. The non linear maximum likelihood step provides an estimation of the fixed effects β and the variance σ^2, given $\mathbf{b}_i$. The nonparametric EM step consists in an update of the parameters of the discrete distribution $(\mathbf{c}, \boldsymbol{\omega})$ that increases the likelihood function. The property of increasing the likelihood was proved in [129].

The update is the following:

$$\begin{cases} \tilde{\omega}_l = \dfrac{1}{M}\displaystyle\sum_{i=1}^{M} W_{il} \\ \tilde{\mathbf{c}}_l = \arg\max\limits_{\mathbf{c}} \left[\displaystyle\sum_{i=1}^{M} W_{il} \ln\{ f(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c})\} \right] \end{cases} \tag{4.45}$$

where

$$W_{il} = \frac{\omega_l f(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_l)}{\sum_{k=1}^{N} \omega_k f(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_k)}$$

and

$$p(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_l) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_{ij}-g(\beta,\mathbf{c}_l,t_j)\right)^2}.$$

The coefficients W_{il} represent the probability of $\mathbf{b}_i$ being equal to $\mathbf{c}_l$ conditionally to the observation $\mathbf{y}_i$ and given the fixed effects β and the variance σ^2, that is

$$W_{il} = f(\mathbf{c}_l|\mathbf{y}_i,\beta,\sigma^2)$$

in fact,

$$W_{il} = \frac{f(\mathbf{c}_l) f(\mathbf{y}_i|\beta,\sigma^2,\mathbf{c}_l)}{f(\mathbf{y}_i|\beta,\sigma^2)} = \frac{f(\mathbf{y}_i,\mathbf{c}_l|\beta,\sigma^2)}{f(\mathbf{y}_i|\beta,\sigma^2)} = f(\mathbf{c}_l|\mathbf{y}_i,\beta,\sigma^2).$$

In order to estimate $\mathbf{b}_i$ for $i = 1,\ldots,M$ we want to maximize the conditional probability of $\mathbf{b}_i$ conditionally to the observations $\mathbf{y}_i$ and given the fixed effects β and the error variance σ^2. For this reason the estimation of the random effects, $\hat{\mathbf{b}}_i$, is obtained maximizing W_{il} over l, that is

$$\hat{\mathbf{b}}_i = \mathbf{c}_{\tilde{l}} \text{ if } \tilde{l} = \arg\max_{l} W_{il}.$$

During the nonparametric EM step, we could also reduce the support of the discrete distribution. The reduction of the support is performed in order to cluster the support of random effects. This support reduction consists in both making points very close to each other collapse and removing points with very low weight and not associated with any subject. In particular if two points are too close, that is $\|\mathbf{c}_l - \mathbf{c}_k\| < D$, where D is a tuning tolerance parameter, than we replace $\mathbf{c}_l$ and $\mathbf{c}_k$ with a new point $\mathbf{c}_{\min\{l,k\}} = (\mathbf{c}_l + \mathbf{c}_k)/2$ with weight $\omega_{\min\{l,k\}} = \omega_l + \omega_k$. Otherwise, if $\omega_l < \tilde{\omega}$, where $\tilde{\omega}$ is another tuning tolerance parameter, and the subset $\left\{ i : \hat{\mathbf{b}}_i = \mathbf{c}_l \right\}$ is empty, we remove the point $\mathbf{c}_l$. The thresholds D and $\tilde{\omega}$ are two complexity parameters that affect the estimation of the nonparametric distribution; the higher D is set, the lower is the number of groups. For this reason the two complexity parameters define a trade off between bias and high number of groups. We prefer setting D low in order to obtain an higher number of groups and, in case, cluster them later.

The non linear maximum likelihood step provides the estimation of the fixed effects β and the errors variance σ^2, given $\mathbf{b}_i = \hat{\mathbf{b}}_i$. In this step we maximize the non linear log-likelihood:

$$\ell(\beta,\sigma^2|\,\mathbf{y},\hat{\mathbf{b}}) = -\frac{nM}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{M}\sum_{j=1}^{n}\left(y_{ij} - g(\beta,\hat{\mathbf{b}}_i,t_j)\right)^2$$

where $\hat{\mathbf{b}}_i$ is the estimation of random effects for the i-th subject provided in the nonparametric EM step.

The algorithm, given a starting discrete distribution with M support points for the random effects and a starting estimate for the fixed effects, alternate the nonparametric EM step and the non linear maximum likelihood step until convergence. Technical details can be found in [12]. Here we report a sketch of the algorithm:

1. Define a starting discrete distribution for random effects with support on M points $(\mathbf{c}^{(0)}, \omega^{(0)})$, a starting estimate for the fixed effects $\beta^{(0)}$ and for $\sigma^{2(0)}$ and the tolerance parameters D and $\tilde{\omega}$;

2. given $(\mathbf{c}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})$, $\beta^{(k-1)}$ and $\sigma^{2(k-1)}$, perform the EM step (without the support reduction) in order to update the support points $\mathbf{c}^{(k)}$ and the weights $\boldsymbol{\omega}^{(k)}$ of the random effect distribution, according to equation (4.45);

3. given $(\mathbf{c}^{(k)}, \boldsymbol{\omega}^{(k)})$, perform the nonlinear maximum likelihood step in order to estimate the fixed effects $\beta^{(k)}$ and the error variance $\sigma^{2(k)}$;

4. iterate steps 2 and 3 until convergence;

5. reduce the support of the discrete distribution, according with the tuning parameters D and $\tilde{\omega}$;

6. given $(\mathbf{c}^{(k-1)}, \boldsymbol{\omega}^{(k-1)})$, $\beta^{(k-1)}$, $\sigma^{2(k-1)}$, D and $\tilde{\omega}$, perform the EM step with the support reduction in order to update the support points $\mathbf{c}^{(k)}$ and the weights $\boldsymbol{\omega}^{(k)}$ of the random effect distribution, according to equation (4.45);

7. given $(\mathbf{c}^{(k)}, \boldsymbol{\omega}^{(k)})$, perform the nonlinear maximum likelihood step in order to estimate the fixed effects $\beta^{(k)}$ and the error variance $\sigma^{2(k)}$;

8. iterate steps 6 and 7 until convergence.

The algorithm reaches convergence when parameters and discrete distribution stop changing or when there is no variation in the log-likelihood function.

In order to validate the proposed estimation algorithm and to compare it with different procedures, two simulation studies are proposed in [12] and detailed in Section 8.2. Since the main interest is in classifying curves in an unsupervised framework, attention is focused on the estimation of random effects distribution.

In the first simulation study, the testing framework is the linear one, in order to be able to compare results of our procedure with those obtained with the algorithm introduced in [3] and implemented in the `npmlreg` `R`-package (see [37]). In the second one, two classic non linear functions g in (4.44) are considered: the exponential and the logistic growth curves. In both these cases, the number of groups and distribution of random effects are correctly and effectively identified by our method, which performs better than the competitor not only when nonlinearities and high number of groups are present, but also in the linear case. Test sets of simulated curves and evaluation of the algorithm performances in the estimation of the random effects are detailed in [12], together with an application to real data arising from administrative data banks. Results of the analysis carried out with this method on data arising from PHD of Regione Lombardia are also presented in Chapter 8, Section 8.1.

4.7 Problems due to unbalanced share

In a binary logistic regression analysis with unequal sample frequencies of the two outcomes the less frequent outcome always has lower estimated prediction probabilities than the other one. This effect is unavoidable, and its extent varies inversely with the fit of the model, as given in [28] by a new measure that follows naturally from the argument. Unbalanced samples with a poor fit are typical for survey analyses in the social sciences and epidemiology, and there the difference

in prediction probabilities is most acute. It affects two common diagnostics: the within-sample percentage correctly predicted and the identification of outliers.

In this section, we deal with the problems that may arise in estimating and predicting success probabilities with logistic models when samples are unbalanced. The context is then set to be the standard binary logistic regression with Maximum Likelihood estimation of unknown parameters. Leaving the parameter estimates aside, we focus on the estimated within-sample probabilities $\hat{p}_i$ of the outcome $y_i = 1$. These probabilities are arranged in a vector $\hat{\mathbf{p}}$, with complement $\hat{\mathbf{q}}$; the outcomes are likewise recorded in $\mathbf{y}$, with complement vector $\mathbf{z}$. Starting from this setting, the crude residuals can be defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{p}}$, with the complement vector $\mathbf{z} - \hat{\mathbf{q}}$. The sample consists of n observations, m with $y_i = 1$ and $n - m$ with $y_i = 0$. We call *shares* the relative empirical means of $\mathbf{y}$ and $\mathbf{z}$. Let us denote them as π and $1 - \pi$ respectively. Whenever the two sample shares are unequal, π is by convention the larger share and the corresponding outcome is labelled $y_i = 1$. In the case of interest, π will be the in-hospital or long-term survival probability of a patient admitted in any hospital of Regione Lombardia with STEMI diagnosis.

The regressor matrix of the full model is $\mathbf{X}$, and the ML estimates $\hat{\mathbf{p}}$ of the logit model satisfy $\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{p}}) = \mathbf{X}^T\mathbf{e} = \mathbf{0}$. Since $\mathbf{X}$ is always taken to include a unit constant, in particular it holds the following identity $\mathbf{1}^T(\mathbf{y} - \hat{\mathbf{p}}) = \mathbf{1}^T\mathbf{e} = \mathbf{0}$, in other terms $\pi = \bar{p}$, where $\bar{p}$ is the overall mean of the elements of $\hat{\mathbf{p}}$ (i.e., $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i$). This property of the estimated probabilities will be called equality of the means. We shall make use of the estimated probability of the observed outcome $\mathbb{P}(i)$,

$$\mathbb{P}(i) = y_i\hat{p}_i + z_i\hat{q}_i \tag{4.46}$$

Note that the maximum of the log-likelihood function

$$\log(\hat{L}) = \sum_i \log\{\mathbb{P}(i)\}$$

The null model L_0 is the model with the unit constant as the sole regressor is nested in the full model with richer $\mathbf{X}$. In this model $\hat{p}_i$ and $\hat{q}_i$ are constant and equal to π and $1 - \pi$ respectively, with log-likelihood

$$\log(L_0) = m\log(\pi) + (n - m)\log(1 - \pi)$$

This is the lower limit of $\log(\hat{L})$; on average, the $\mathbb{P}(i)$ are at least equal to the estimated probabilities with the null model L_0, but it is of course hoped that they are substantially higher. This leads us to consider the ratio of $\mathbb{P}(i)$ to its null value

$$\mathbb{P}_r(i) = y_i(\hat{p}_i/\pi) + z_i\{\hat{q}_i/(1 - \pi)\} \tag{4.47}$$

$\mathbb{P}_r(i)$ reflects the improvement of the full model over the null model in predicting the *i*-th outcome; it is an index of performance for that particular observation. It is not a probability; it is non-negative, and its average should exceed 1. On taking logarithms and summing we find

$$\sum \log\{\mathbb{P}_r(i)\} = \log(\hat{L}) - \log(L_0) \tag{4.48}$$

Doubling this gives *LR*, the common likelihood ratio statistic for the significance of the full model.

4.7.1 Prediction probabilities in unequal sample size

In most survey data in the social sciences and epidemiology the sample shares of the two outcomes are unequal, for example in STEMI Archive we have a value for in-hospital survival which is around

0.94. Such values are much more common than equal shares. On fitting a logistic model it is then invariably found that the estimated prediction probabilities $\mathbb{P}(i)$ are quite high for $y_i = 1$, the outcome with the greater share, and very low for the outcome with the lesser share (inequality of sample proportions of the outcomes thus by itself leads to a high overall level of $\mathbb{P}(i)$ and to high log-likelihoods). If we distinguish two subsets among the $\hat{p}_i$, with $\hat{p}_i^+$ for $y_i = 1$ and $\hat{p}_i^-$ for $y_i = 0$, and likewise for $\hat{q}_i$, the $\hat{p}_i^+$ have a much higher overall level than the $\hat{q}_i^+$. This asymmetry in the prediction of $y_i = 1$ and $y_i = 0$ is well known to practitioners. As highlighted in [28], yet there is no clear reason why a rare outcome should be badly predicted; a good prediction must be simply a matter of choosing the right regressors. This is indeed so: even quite rare outcomes can have estimated probabilities all the way up to 1; but, whatever value they attain, on average the other, prevalent, outcome will always be predicted even better. The extent of this systematic difference varies with the fit of the model, and since outside controlled experiments the fit is usually mediocre a great contrast between the poor prediction of rare states and the good prediction of prevalent states is the rule.

This result can be explained by the following argument. Consider the averages of $\hat{p}_i^+$ and $\hat{p}_i^-$ for the two subsets of observations with $y_i = 1$ and $y_i = 0$ respectively, i.e.,

$$\begin{aligned} \bar{p}^+ &= \hat{p}^T y/m \\ \bar{p}^- &= \hat{p}^T z/(n-m) \end{aligned} \tag{4.49}$$

with the first refers to the outcome with the largest share. Similarly

$$\begin{aligned} \bar{q}^- &= \hat{q}^T y/m \\ \bar{q}^+ &= \hat{q}^T z/(n-m) \end{aligned} \tag{4.50}$$

The overall mean $\bar{p}$ is the weighted average of equations (4.49), or

$$\pi\bar{p}^+ + (1-\pi)\pi\bar{p}^- = \bar{p} = \pi \tag{4.51}$$

If the fitted model has any explanatory power, $\bar{p}^+$ exceeds $\bar{p}^-$, and the two will lie on either side of their (weighted) average π. Both mean probabilities are constrained to the interval $(0,1)$; $\bar{p}^-$ lies in $(0,\pi]$ and $\bar{p}^+$ in $[\pi,1)$. This suggests writing $\bar{p}^+$ as a linear combination of π and 1 with non-negative weights $1-\lambda$ and λ, or

$$\bar{p}^+ = (1-\lambda)\pi + \lambda = \pi + \lambda(1-\pi) \tag{4.52}$$

and by equation (4.51)

$$\bar{p}^- = (1-\lambda)\pi \tag{4.53}$$

so that $\bar{p}^-$ is a linear combination of 0 and π with the same weights λ and $1-\lambda$. Similar expressions hold for the $\hat{q}_i$. Making use of

$$\begin{aligned} \bar{p}^+ + \bar{q}^- &= 1 \\ \bar{p}^- + \bar{q}^+ &= 1 \end{aligned}$$

we find

$$\begin{aligned} \bar{q}^+ &= 1 - \pi + \pi\lambda \\ \bar{q}^- &= (1-\lambda)(1-\pi) \end{aligned} \tag{4.54}$$

Thus $\bar{q}^+$ and $\bar{q}^-$ are linear combinations like $\bar{p}^+$ and $\bar{p}^-$ with the same weights. The upshot is that all four means are determined by two parameters, the share π and the weight λ, with $0 < \pi < 1$

and $0 \leq \lambda < 1$. The limits of π are self-evident; λ is 0 for the null model with $\bar{p}^{+} = \bar{p}^{-} = \bar{p} = \pi$. Negative values are ruled out, as they would mean that the log-likelihood of the full model is less than that of the null model, and λ cannot attain its upper bound since this would imply that $\bar{p}^{+} = \bar{q}^{+} = 1$. Such perfect prediction is beyond logit probabilities and their estimate.

Combining equations (4.52) and (4.54) as

$$\bar{p}^{+} - \bar{q}^{+} = 2(\pi - 0.5)(1 - \lambda) \tag{4.55}$$

we have the answer to the initial question why the level of predicted probabilities varies with the sample share. Unless $\pi = 0.5$, $\bar{p}^{+}$ exceeds $\bar{q}^{+}$, and this excess varies inversely with λ. Large values of λ therefore limit its size; but in practice this is of little help, as λ is usually quite small, as proved by simulation studies carried out in [28]. In these conditions estimated probabilities are a poor measure of within-sample predictive performance: they may lead to the absurd conclusion that success is predicted very well whereas failure is predicted badly, as if we can simultaneously predict survival with precision but death not at all. From equations (4.52)-(4.54) we also find

$$\bar{p}^{+} - \bar{p}^{-} = \bar{q}^{+} - \bar{q}^{-} = \lambda \tag{4.56}$$

λ can therefore be seen as a crude measure of fit since it indicates the discrimination of $\hat{p}_i$ (and of $\hat{q}_i$) between the two observed outcomes. Further comments on this consideration can be found in [28].

4.7.2 Undesirable effects for unbalanced samples

In this paragraph we come back to unbalanced samples with widely different levels of $\hat{p}_i^{+}$ and $\hat{q}_i^{+}$ in order to examine the effect of unbalancing on two common diagnostics. Usually, after fitting a logistic regression model, the percentage correctly predicted in the sample is computed. Estimated $0-1$ predicted values $\hat{y}_i$ are assigned to the observations according to whichever is the greater of $\hat{p}_i^{+}$ and $\hat{q}_i^{+}$, i.e., if $\hat{p}_i^{+} \geq 0.5$, $\hat{y}_i = 1$; $\hat{q}_i^{+} \geq 0.5$, $\hat{y}_i = 0$. But the number of correct prediction carried out in this way reflects the composition of the sample rather than the performance of the model. This incongruous result hinges on the cut-off point of 0.5. This choice is usually defended by the argument that it is optimal if the $\hat{y}_i$ determine a course of action and if moreover the cost of misclassification is the same for either form that this may take. But if the cut-off point is optimal for the use of the predictions in actual decisions it need not also be optimal for assessing the within-sample performance of the fitted model. For the latter purpose it is natural to use a prediction that is optimal in the sense that, for given $\hat{p}_i$, it maximizes $\mathbb{P}_r(i)$ of equation (4.47), and hence the fit of $\hat{y}$ to the given $\hat{p}$. This is achieved by a cut-off point of π. We will adopt this criterion for choosing the classification threshold in our logistic regression models, as can be observed in [62] and [66] among others.

Chapter 5

Statistical models for healthcare: the Bayesian approach

In this chapter, the Bayesian approach to problems and data proposed in the previous part is presented. Both parametric and nonparametric mixed effect models are considered. In particular, we are interested to inference on grouping factors provided by the predictive distributions in parametric models and to in built clustering of random effects arising from the choice of Dirichlet priors in semi-parametric ones. This is because one of the main aim of this work is to detect eventual patterns among the grouping factors. Finally, the Bayesian decision theory is also considered, since it can provide support to healthcare governance in term of identification of structures that need to improve their performances.

5.1 Motivations

In order to address the issues arising from the problems treated in Part I, we saw that it is important to incorporate scientists' expertise into making decisions related to the data. It is also clear that prediction plays a central role in the decision making process itself. Bayesian statistical analysis is based on the premise that all uncertainty should be modeled using probabilities and that inferences arising from statistical models should be logical conclusions based on the laws of probability. Models typically involve parameters that are presumed to be related to characteristics of the sampled population, and for them the Bayesian approach mandates an additional probability model, incorporating the "a priori" information, obtaining a hierarchy of models. A key question for this type of approach becomes then how the causal structure that operates at one level of analysis varies across a higher level of analysis. The Bayesian approach to statistical inference is extremely well-suited to answering this question. The idea in Bayesian approach is that parameters are always random variables, typically in the sense that the researcher is unsure as to their value, but can characterize that uncertainty in the form of a *prior* density replacing the prior with a stochastic model formalizing the researcher's assumptions about the way the parameters might vary across groups, perhaps as a function of observable characteristics of the groups. The model is then comprised of a nested hierarchy of stochastic relations: the data from each group are modeled as functions of covariates and parameters, while cross-groups heterogeneity in parameters is modeled as a function of group-specific covariates and *hyperparameters*. Starting from prior probabilities, which describe the current state of knowledge of the researcher about the phenomenon to be modeled, the Bayesian approach incorporates information through collection of data, leading to new probabilities (*posterior* distributions) to describe the state of knowledge after combining prior and data. Inference is then carried out on

parameters through computation of *posterior* or *marginal* distributions of the parameters, as well as *predictive* distribution for new observations. The Bayes theorem, the assumption of *exchangeability* and the MCMC algorithm for simulation and sampling will be the instruments that allow us to reach this goal in the Bayesian setting.

In most analyses of the following chapters, we want to learn about a continuous parameter, like the mean of a continuous variable, a proportion (that is a continuous parameter on an interval) or a regression coefficient. In general, let $\boldsymbol{\psi}$ be the vector of unknown parameters and $\mathbf{y}$ the vector of data available for the analysis. In this case, the beliefs about the parameters are presented as probability density functions. Denoting $f(\boldsymbol{\psi})$ the prior density and $f(\boldsymbol{\psi}|\mathbf{y})$ the posterior, the *Bayes theorem* states that

$$f(\boldsymbol{\psi}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})}{\int f(\mathbf{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})d\boldsymbol{\psi}}$$

often written as $f(\boldsymbol{\psi}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\psi})f(\boldsymbol{\psi})$, since the constant of proportionality (the denominator) does not depend on $\boldsymbol{\psi}$. In some cases, the posterior distribution belongs to the same class of parametric densities of the prior. When it happens, the prior is said to be *conjugate* with respect to the likelihood, and then the posterior belongs to the same family of the prior, and its parameters will result in an updated version of the prior ones through data. In a Bayesian approach, the best information one can ever have about the parameters is their posterior density, since it allows for any inference purpose the researcher has. In fact, this is the strenght of the Bayesian paradigm, i.e., the fact it provides a distribution for the unknown parameters which allows for any inference to be carried out. This has become particulary true and feasible since the introduction of sampling method for posterior distributions that were not analytics, that is the MCMC methods (see [47], [48] and [128] among others for introduction and a deeper tractation of the topic).

Moreover, another inferential issue of interest is the prediction of a new set of observations, say $\mathbf{y}_{new}$, independent of $\mathbf{y}$ given $\boldsymbol{\psi}$. The predictive density $f_p(\mathbf{y}_{new}|\mathbf{y})$ of the future observations given the past ones is given by

$$f_p(\mathbf{y}_{new}|\mathbf{y}) = \int f_p(\mathbf{y}_{new}|\boldsymbol{\psi})f(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi}$$

To see this, it must be used the fact that by conditional independence $f_p(\mathbf{y}_{new}|\mathbf{y},\boldsymbol{\psi}) = f_p(\mathbf{y}_{new}|\boldsymbol{\psi})$ and that for any measurable set A, by the law of total probability

$$\begin{aligned}
\mathbb{P}(\mathbf{y}_{new} \in A|\mathbf{y}) &= \int \mathbb{P}(\mathbf{y}_{new} \in A|\mathbf{y},\boldsymbol{\psi})f(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \\
&= \int \left[\int_A f_p(\mathbf{y}_{new}|\mathbf{y},\boldsymbol{\psi})d\mathbf{y}_{new}\right] f(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} = \int \left[\int_A f_p(\mathbf{y}_{new}|\mathbf{y},\boldsymbol{\Psi})f(\boldsymbol{\Psi}|\mathbf{y})d\mathbf{y}_{new}\right] d\boldsymbol{\Psi} \\
&= \int_A \left[\int f_p(\mathbf{y}_{new}|\mathbf{y},\boldsymbol{\psi})f(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi}\right] d\mathbf{y}_{new} = \int_A f_p(\mathbf{y}_{new}|\mathbf{y})d\mathbf{y}_{new}
\end{aligned}$$

Even if, as we said before, the best one can obtain when the purpose is to carry out inference is the posterior distribution of the parameters of interest, it is often necessary to report such wider information in few point estimates or summary statistics. The choice of which point summary of the posterior distribution to report can be rationalized by drawing on Bayesian *decision theory*, which may address the specific problem of choosing suitable summary of posterior distribution, but also, more generally, can answer the question of how to make rational choices under conditions of uncertainty.

Definition 5.1.1 *Let Ψ be a set of possible states of parameter $\boldsymbol{\psi}$, and let $d \in \mathscr{D}$ be decision available to the researcher. Then define $L(\boldsymbol{\psi},d)$ as the loss to the researcher from taking decision d when the parameter is $\boldsymbol{\psi}$.*

Averaging the losses over beliefs about $\boldsymbol{\psi}$, makes it possible to defined the *expected loss* as

Definition 5.1.2 *If $f(\boldsymbol{\psi})$ is the probability density for $\boldsymbol{\psi} \in \Psi$ at time of decision making, the Bayesian expected loss of a decision d is*

$$\rho(f(\boldsymbol{\psi}), d) = \mathbb{E}[L(\boldsymbol{\psi}, d)] = \int_{\Psi} L(\boldsymbol{\psi}, d) f(\boldsymbol{\psi}) d\boldsymbol{\psi}$$

An interesting case is when f in Definition 5.1.2 is a posterior density. In fact, given a posterior density for $\boldsymbol{\psi}$, say $f(\boldsymbol{\psi}|\mathbf{y})$, the posterior expected loss of a decision d is $\rho(f(\boldsymbol{\psi}), d) = \int_{\Psi} L(\boldsymbol{\psi}, d) f(\boldsymbol{\psi}|\mathbf{y}) d\boldsymbol{\psi}$.

A Bayesian rule for choosing among decisions $\mathscr{D}$ is to select $d \in \mathscr{D}$ so to minimize posterior expected loss. If the chosen loss function is convex, the corresponding Bayes estimate is unique, so the choice of what Bayes estimate to report usually amounts to what (convex) loss function to adopt. It can be proved, in fact, that the mean of the posterior density is the best choice for the parameter $\boldsymbol{\psi}$ arising from the use of quadratic loss. The same for suitable quantiles of posterior density in case of linear loss. See [88] for a deeper tractation and further references on the topic.

Once we presented motivation and theoretical context for Bayesian paradigm, in the next sections we will analyze the Bayesian modelling approach for parametric, nonparametric, generalized linear and nonlinear mixed-effects models.

5.2 Parametric Models

The rationale for applying multilevel models to grouped data is well established. When units at lower levels are nested within one or more higher level strata, conventional single level regression analysis is not appropriate since observations are no longer independent: pupils in the same schools, households in the same communities, patients in the same hospital tend to be more similar one another than pupils in different schools, households in different communities and patients in different hospitals. Such dependency means standard errors are underestimates if the nesting is ignored, and spurious inference regarding predictor effects may be made (see [72] and [73]). This asks for proper descriptions of variables that may differ between groups. A way to address this issue is to join regression models which describe within-group variation with models that describe heterogeneity among regression coefficients across groups, i.e multilevel models.

In multilevel analysis, predictors may be introduced at any level and the interest focuses on adjusting predictor effects for the simoultaneous operation of contextual and individual variability in the outcome. This may be important in health applications, for example, if the impact of individual level risk factors varies according to geographic context or organizational one. Another major goal is variance partitioning: for example, what proportion of area variations in mortality is due to the characteristics of those areas (*contextual variation*), and how much is due to the characteristics of the individuals who live in these areas (*individual variation*), as detailed in [21] and [22]. As well as predictor effects at any level, a multilevel model is likely to involve random effects defined over the clusters at higher level(s), and makes possible to carry out inferences about cluster effects. We enhanced all these goals within the context of data arising from STEMI Archive.

5.2.1 Hierarchical linear mixed effects models

To model hierarchical data where units belonging to each group are different in terms of sample size and/or variance, we use an ordinary regression model to describe the within-group heterogeneity of

observations, while we describe the between-group heterogeneity using a sampling model for the group specific regression parameters [26]. It is known, in fact, that the smaller is the sample size for the group, the more probable that unrepresentative data are sampled and an extreme estimate is produced. In order to overcome this problem, hierarchical mixed effects models are introduced in order to stabilize estimates for small sample size groups by sharing information across groups. This is the reason why hierarchical models estimates can be regarded as realizations of "shrinkage" estimators ([88], Paragraph 7.1.4). Focusing on grouped data for which i is the statistical unit index belonging to the j-th group, the linear model results to be

$$Y_{ij} = \beta_j^T \mathbf{w}_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{5.1}$$

where, for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, $\mathbf{w}_{ij}$ is a $p \times 1$ vector of regressors for observation i in group j and β_j is a p dimensional random vector. Expressing $Y_{1j}, \ldots, Y_{n_j j}$ as a vector $\mathbf{Y}_j$ and combining $\mathbf{w}_{1j}, \ldots, \mathbf{w}_{n_j j}$ into a $n_j \times p$ matrix $\mathbf{W}_j$, the within-group sampling model can be equivalently expressed as $\mathbf{Y}_j \sim \mathcal{N}(\mathbf{W}_j \beta_j, \sigma^2 \mathbb{I})$ with the group-specific data vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_J$ being conditionally independent given $\beta_1, \ldots, \beta_J$ and σ^2.

The eterogeneity among the regression coefficients $\beta_1, \ldots, \beta_J$ will be described with a between-group sampling model. If we have no prior information distinguishing the different groups, we can model them as being exchangeable, i.e., the joint probability of any permuted sequence is the same as the joint probability distribution of the original one. Otherwise, we can consider them (roughly) equivalently as being i.i.d from some distribution representing the sampling variability across groups. The *Normal Hierarchical Regression Model* describes the across-groups heterogeneity with a multivariate Normal model, where β_j are drawn from a population of regression parameters with random mean θ and random variance-covariance matrix Σ, i.e.,

$$\beta_1, \ldots, \beta_J | \theta, \Sigma \overset{i.i.d}{\sim} \mathcal{N}(\theta, \Sigma) \tag{5.2}$$

A graphical representation of the hierarchical model appears in Figure 5.1, and makes clear that the multivariate Normal distribution for $\beta_1, \ldots, \beta_J$ is not a prior distribution representing uncertainty about fixed but unknown quantity. Rather, it is a sampling distribution representing heterogeneity among collection of objects.

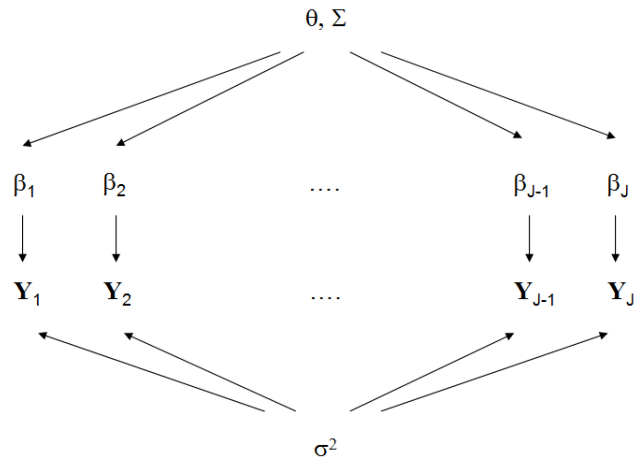


Figure 5.1: Graphical representation of the hierarchical model in 5.2.

The hierarchical regression model in (5.1)-(5.2) can be reparametrized in a different way that makes clear why these models are called *linear mixed effect models*. In fact, rewriting (5.1) and (5.2), a slightly different between-groups sampling model can be obtained:

$$\begin{aligned} \beta_j &= \theta + \mathbf{b}_j \\ \mathbf{b}_1, \ldots, \mathbf{b}_J | \Sigma &\overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \Sigma), \end{aligned}$$

where Σ is random as before. Plugging this into within-group regression model gives

$$\begin{aligned} Y_{ij} &= \beta_j^T \mathbf{w}_{ij} + \varepsilon_{ij} \\ &= \theta^T \mathbf{w}_{ij} + \mathbf{b}_j^T \mathbf{w}_{ij} + \varepsilon_{ij} \end{aligned}$$

In this parametrization, θ is referred to as *fixed effect* as it is constant across groups, whereas $\mathbf{b}_1, \ldots, \mathbf{b}_J$ are called *random effects*, as they vary. Although for our particular example the regressors corresponding to the fixed and random effects are the same, this is not compulsory. Splitting the covariates contained in the design matrix $\mathbf{W}_j$ in a matrix $\mathbf{X}_j$ (containing only covariates related to fixed effects) and in another matrix $\mathbf{Z}_j$ (containing those related to random effects only), a more general model comes out and can be written as

$$Y_{ij} = \theta^T \mathbf{x}_{ij} + \mathbf{b}_j^T \mathbf{z}_{ij} + \varepsilon_{ij} \tag{5.3}$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ could be vectors of different lengths which may or may not contain overlapping variables. The prior of random effects $(\mathbf{b}_1, \ldots, \mathbf{b}_J)$ is assumed to be exchangeable.

Given a prior distribution for $(\theta, \Sigma, \sigma^2)$ and having observed $\mathbf{Y}_1 = \mathbf{y}_1, \ldots, \mathbf{Y}_J = \mathbf{y}_J$, a Bayesian analysis proceeds by computing the posterior distribution $f(\mathbf{b}_1, \ldots, \mathbf{b}_J, \theta, \Sigma, \sigma^2 | \mathbf{y}_1, \ldots, \mathbf{y}_J)$. If conjugate prior distributions are used for θ, Σ and σ^2, then the posterior distribution can be approximated quite easily with Gibbs sampling. Common prior distributions for θ, Σ and σ^2 are

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_0, \Sigma_0) \\ \Sigma^{-1} &\sim Wishart(\mathbf{S}_0, \eta_0) \\ \sigma^2 &\sim inverse - Gamma(\nu_0/2, \nu_0 \sigma_0^2/2) \end{aligned} \tag{5.4}$$

A common alternative for level 1 variance σ^2 is represented by a Uniform distribution (for detailed discussion about sensitivity of results to the specification of priors hyperparameters, see [50]). We adopted this framework and this choice in [63] and [66].

Computing the posterior distribution for so many parameters may seem danting, and effectively is analytically possible only for conjugate models. Even in this case, computational effort requested for high-dimensional parameters vectors are not straightforward. On the other hand, posterior distributions of parameters of interest are the main focus of our inferential efforts. The problem has been overcome thanks to the development of MCMC methods (see [47] and [128]), which rest upon the Markov Chain theory to get samples from the joint posterior distribution (or to approximate it), using the full conditional distributions of the parameters themselves. The most frequently adopted algorithm for getting and simulating full conditional distributions is the Gibbs Sampler [23], as it is implemented in `R` [124] or `jags` [123]. It is a version of the more general family of Metropolis-Hastings algorithms [24], that iteratively sample from full conditionals to approximate the joint posterior distribution. Then the focus moves to how getting samples from full conditional distributions of parameters of interest.

Full conditionals for $\mathbf{b}_1, \ldots, \mathbf{b}_J$

Hierarchical regression model shares information across groups via the parameters $(\theta, \Sigma, \sigma^2)$. As a result, conditional to θ, Σ and σ^2 the regression coefficients $\beta_1, \ldots \beta_J$ are a priori independent. This helps in computing the full conditionals in the Gibbs Sampler algorithm, since it can be shown (see [71] and [72], Paragraph 9.2.1) that

$$\mathbf{b}_j|\mathbf{y}_j, \theta, \Sigma, \sigma^2 \sim \mathscr{N}\left((\mathbf{Z}_j^T\mathbf{Z}_j + \sigma^2\Sigma^{-1})^{-1}\mathbf{Z}_j^T(\mathbf{y}_j - \mathbf{X}_j\theta), \sigma^2(\mathbf{Z}_j^T\mathbf{Z}_j + \sigma^2\Sigma^{-1})^{-1}\right)$$

Full conditional for Σ

We have assumed that $\mathbf{b}_j$s conditioned to Σ are independent Normal random variables with $\mathbf{0}$ mean and variance-covariance matrix Σ. The standard non informative prior for Σ is the one we assumed in (5.4), then, the full conditional distribution of Σ^{-1} given $\mathbf{b}_j$ $(j = 1, \ldots, J)$ at each iteration, denoted by $\mathbf{b}_j^{(k)}$ follows a Wishart distribution with the following parameters:

$$\Sigma^{-1}|\mathbf{b}_1^{(k)}, \ldots, \mathbf{b}_J^{(k)} \sim Wishart(S^{(k)}, \eta_0 + J)$$

where $S^{(k)} = \sum_{j=1}^{J} \mathbf{b}_j^{(k)}\mathbf{b}_j^{(k)^T}$ and $\eta_0 + J$ are the degrees of freedom (see [148]). Note that $S^{(k)}$ must be recomputed each time the vectors $\mathbf{b}_j^{(k)}$ of the random effects is updated in Gibbs sampler Markov chain.

Full conditionals for σ^2

The parameter σ^2 represents the error variance, assumed to be common across all groups. As such, conditional on $\mathbf{b}_1, \ldots, \mathbf{b}_J, \theta$, the data provide information about σ^2 via the sum of squared residuals from each group:

$$\sigma^2|\mathbf{b}_1, \ldots, \mathbf{b}_J, \theta \sim inverse - Gamma([\nu_0 + \sum_j n_j]/2, [\nu_0\sigma_0^2 + SSR]/2)$$

where $SSR = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - (\theta^T\mathbf{x}_{ij} + \mathbf{b}_j^T\mathbf{z}_{ij}))^2$. Again, note that SSR also depends on the value of $\mathbf{b}_j$, and so must be recomputed in each scan of the Gibbs sampler before σ^2 is uploaded.

5.2.2 Hierarchical generalized linear mixed effects models

As the name suggests, a generalized linear mixed effects model combines aspects of linear mixed effects models with those of generalized linear models. Such models are useful when we have hierarchical data structure but the Normal model for the within-group variation is not appropriate. Within this setting is the work [66], which uses Bayesian generalized linear mixed effects models with parametric additive random effect on grouping factor (the hospital of admission) to account for overdispersion induced by the grouped nature of STEMI Archive data, as well as to classify providers.

Suppose Y_{ij}, $i = 1, \ldots, n_j$ to be a conditionally independent sample, drawn from a distribution belonging to the exponential family. In the analysis of STEMI Archive data and in general in models

proposed in Part III, we will be interested in GLME models for binary response, in particular logistic regression, i.e.,

$$f(y_{ij}|\theta,\mathbf{b}_j) = exp\left\{\mathbf{y}_{ij}(\mathbf{x}_{ij}^T\theta + \mathbf{z}_{ij}^T\mathbf{b}_j) - \log\left(1+e^{\mathbf{x}_{ij}^T\theta+\mathbf{z}_{ij}^T\mathbf{b}_j}\right)\right\}$$

Notice that GLME models imitates the Normal hierarchical mixed effects model in that we assume that, conditional on the random effect $\mathbf{b}_j$, the observations of group j are independent. Thus the likelihood for J groups in the GLME model is

$$f(\mathbf{y}|\mathbf{X},\mathbf{Z},\theta,\mathbf{b}) = \prod_{j=1}^{J} f(\mathbf{y}_j|\mathbf{X}_j,\mathbf{Z}_j,\theta,\mathbf{b}_j) \propto \prod_{j=1}^{J}\prod_{i=1}^{n_j} f(y_{ij}|\theta^T\mathbf{x}_{ij},\mathbf{b}_j^T\mathbf{z}_{ij}) \tag{5.5}$$

where $\mathbf{b} = (b_1,\ldots,b_J)$ and $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_J)$. Usually in parametric framework, the model is completed by the specification of the following priors:

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_0,\Sigma_0) \\ \mathbf{b}_1,\ldots,\mathbf{b}_J|\Sigma &\overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0},\Sigma) \\ \Sigma^{-1} &\sim Wishart(\mathbf{S}_0,\eta_0) \end{aligned}$$

where (μ_0,Σ_0) are considered known and fixed.

Bayesian estimation in the linear mixed effects model is straightforward because the full conditional distributions of each parameter are standard, allowing for easy implementation of the Gibbs sampler. In contrast, for non-Normal generalized linear mixed effects models, tipically only θ and Σ have standard full conditional distributions. This suggests to use a Metropolis-Hastings within Gibbs algorithm [140] to approximate the posterior distribution of the parameters, using a combination of Gibbs steps for updating (θ,Σ) with a Metropolis step to generate from the full conditionals of each $\mathbf{b}_j$. In what follows we will assume the context of logistic regression. If necessary, a more general framework can be assumed, where it is inserted as further parameter to be updated using the Gibbs sampler if the conditional distribution is available, and Metropolis if it is not.

Gibbs step for θ, and Σ

The conditional law of θ given $\Sigma,\mathbf{b}_j,\mathbf{y}_j$ is independent of Σ, as long as they are assumed to be a priori independent. This is the case we are considering.

Given the $\mathbf{b}_j^{(k)}$s, the random effects model reduces to a generalized linear model with offset $\mathbf{z}_{ij}^T\mathbf{b}_j^{(k)}$ for each observation. Assuming a flat prior for θ, $f(\theta|\mathbf{b}_j^{(k)},\mathbf{y})$ is proportional to the likelihood function $\prod_{ij} f(y_{ij}|\beta^T\mathbf{x}_{ij},\mathbf{b}_j^T\mathbf{z}_{ij})$. In larger samples, this can be closely approximated by a Gaussian distribution with mean $\hat{\theta}^{(k)}$, the maximum likelihood estimator, and variance $V_{\hat{\theta}}^{(k)}$, the inverse of the Fisher information. That is, to sample from $f(\theta|\mathbf{b}^{(k)},\mathbf{y})$, we find $\hat{\theta}^{(k)}$ and $V_{\hat{\theta}}^{(k)}$ by performing GLM regression of y_{ij}, on $\mathbf{x}_{ij}$, using the simulated values $\mathbf{z}_{ij}^T\mathbf{b}_j^{(k)}$s as offsets and then generate a random $\theta^{(k+1)}$ from a multivariate Gaussian distribution, $\mathcal{N}(\hat{\theta}^{(k)},V_{\hat{\theta}}^{(k)})$. The preceding Gaussian approximation may not be adequate in smaller samples. A sample from the exact distribution can be obtained with little additional effort using rejection sampling [127]. Denote the Gaussian density by $g(\theta)$ and the true density by $f(\theta)$. To perform rejection sampling, a constant $c \geq 1$ is chosen so that $c\cdot g(\theta) \geq f(\theta)$ over the range of θ. The following steps result in a random variate $\theta^{(k+1)}$ with density $f(\theta)$:

1. Sample $\theta^* \sim g(\theta)$;
2. Sample $u \sim Unif(0,1)$;
3. Set $\theta^{(k+1)} = \theta^*$ if $f(\theta)/(c \cdot g(\theta)) < u$, otherwise return to step 1.

That $\theta^{(k+1)}$ has density $f(\theta)$ is shown in [127]. Note the additional computation is only to evaluate the likelihood function at one or a few θ^*s. The choice of c involves a tradeoff of accuracy and computational effort. It is difficult a priori to decide whether a Gaussian approximation to $f(\theta|\mathbf{b}_j, \mathbf{y})$ is adequate or whether rejection sampling is needed. In practice, therefore, the rejection sampling is always used.

Concerning the Gibbs update for Σ, the scheme presented in the paragraph before can be assumed.

Metropolis step for $\mathbf{b}_j$s

Generating $\mathbf{b}_j^{(k+1)}$s from $f(\mathbf{b}_j|\theta^{(k)}, \Sigma^{(k)}, \mathbf{y}_j)$ is the most time-consuming step. We saw that in the linear mixed-effects model $f(\mathbf{b}_j|\theta, \Sigma, \mathbf{y}_j)$ is a multivariate Normal. Unfortunately, this conditional distribution does not have a closed form for the entire GLM family and must usually be evaluated by numerical techniques. Its density is given by

$$f(\mathbf{b}_j|\theta, \Sigma, \mathbf{y}) = \frac{f(\mathbf{y}_j|\mathbf{b}_j, \theta) f(\mathbf{b}_j|\Sigma) f(\theta, \Sigma)}{\int f(\mathbf{y}_j|\mathbf{b}_j, \theta) f(\mathbf{b}_j|\Sigma) f(\theta, \Sigma) d\mathbf{b}_j} \tag{5.6}$$

The numerator can be easily evaluated (see [148]), but the scale factor in the denominator involves the same integral with respect to $\mathbf{b}_j$. Anyway, in Gibbs sampling, only a simulated value from $f(\mathbf{b}_j|\theta^{(k)}, \Sigma^{(k)}, \mathbf{y}_j)$ is needed. Again, it can be obtained using rejection sampling without evaluating the integral in the denominator. Then the more the mode and curvature of the numerator of (5.6) will match a Gaussian kernel arising from the prior on $\mathbf{b}_j$, the more efficient the rejection sampling will be (for deeper tractation of this topic, see [148]).

5.3 Semiparametric models

In the previous section, we saw how generalized linear mixed-effects models can address the wide range of problems where the assumption of independence among observations is no longer valid, because of the presence of grouping factors that induce stronger correlations between units of the same group. In such models, a group-specific covariance structure is generated by assuming that each group has a unique set of coefficients, the random effects, distributed around the mean regression coefficients for the population, the fixed effects. As we saw, conditional on the random effects, observations are considered independent, while marginalizing over the random effect, a unique covariance structure for the observations within each group is obtained. We will now consider a semi-parametric Bayesian model for generalized linear models with random effects, where these have non-parametric prior distribution. The corresponding frequentist attempt is presented in Section 4.4.

The semi-parametric Bayesian approach for the random effects is to specify a prior distribution on the space of all possible distribution functions of the random effects themselves. Instead of assuming that the random effects parameters are conditionally independent i.i.d. from a parametric distribution, we will assume them from a random distribution function, namely the Dirichlet Process (DP). We will assume such a nonparametric prior within this context because it handles well

relaxations on parametric assumption on random effects distribution and provides an in-built classification of random effects themselves. In fact, the DP prior results in what MacEachern [107] calls a "cluster structure" among the $\mathbf{b}_j$s. This cluster structure partitions the J $\mathbf{b}_j$s into k sets or clusters, $0 \leq k \leq J$. All of the observations in a cluster share an identical value of $\mathbf{b}_j$ and observations in different clusters have differing values of $\mathbf{b}_j$. We will exploit this "in-built classification" provided by discreteness of DP trajectories in order to cluster hospitals (our grouping factors) with similar behaviour in the sense above. In other words, unlike the usual goals of Bayesian inference which is focused on fixed parameters estimation and inference, we will focus our attention on the bias of random effects posteriors, since we want to use them to classify clinical structures.

In our works, an example of the use of DP priors within a GLME model for survival where STEMI patients are grouped by hospital of admission is [67].

5.3.1 Dirichlet Process for clustering

Unlike the parametric case, where we have a prior on a finite dimensional space Ψ and, given ψ, the observations are assumed i.i.d. from a parametric probability distribution P_ψ, in the nonparametric case, we have a prior P on the space $\mathscr{P}(\mathbb{R}^m)$ of all probability distributions on $(\mathbb{R}^m, \mathscr{B}, \mathbb{R}^m)$ and, given P, the observations are assumed i.i.d. from P. Under the assumption of exchangeability, de Finetti's Representation Theorem gives a validation of the Bayesian setting.
Let consider an infinite sequence of observations $(X_n)_{n\geq 1}$ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with each X_i taking values on $\mathbb{R}^m$ endowed with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^m)$. This last hypothesis can be relaxed and we could consider observations which take values in a complete metric and separable space $\mathbb{X}$. Here it is enough to consider $\mathbb{X} = \mathbb{R}^m$. There are several types of dependence among a sequence of observations $(X_n)_{n\geq 1}$. Under the exchangeability assumption, the information that the observations X_is provide is independent of the order in which they are collected. A random element defined on $(\mathbb{R}^m, \mathscr{B}, \mathbb{R}^m)$, with values in $\mathscr{P}(\mathbb{R}^m)$, is called random probability measure (r.p.m.). The most popular r.p.m. classes in literature are Dirichlet Processes, Polya Trees and Bernstein Polynomials. A complete review of the main r.p.m. classes appears in [114].

The Dirichlet Process is a stochastic process introduced in Ferguson [40]. Its distribution consists of a probability law on the space of all probability measures on $\mathbb{R}^m$, for some integer m, which induces finite-dimensional Dirichlet distributions when the data are grouped. Since grouping can be done in many different ways, reduction to finite-dimensional Dirichlet distribution should hold under any grouping mechanism. In practice, this means that for any finite measurable partition $\{B_1, \ldots, B_k\}$ of $\mathbb{R}^m$, the joint distribution of the probability vector $(P(B_1), \ldots, P(B_k))$ is a finite-dimensional Dirichlet distribution. A more formal definition can be obtained thinking the Dirichlet Process as infinte dimensional generalization of the finite-dimensional Dirichlet distribution.

Definition 5.3.1 *Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ with $\alpha_i > 0$ for $i = 1, \ldots, k$. The random vector $P = (P_1, \ldots, P_k)$, $\sum_i P_i = 1$, has Dirichlet distribution with parameter $\boldsymbol{\alpha}$ if $(P_1, \ldots, P_{k-1})$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{k-1}$ with density*

$$f(p_1, \ldots, p_{k-1}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)} p_1^{\alpha_1 - 1} p_1^{\alpha_1 - 1} \ldots p_{k-1}^{\alpha_{k-1} - 1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k - 1}$$

where $0 \leq p_i \leq 1$, $0 \leq p_1 + \cdots + p_{k-1} \leq 1$, 0 otherwise. We will write $P \sim D(\alpha)$.

Starting from this definition, also the one for Dirichlet Process can be given:

Definition 5.3.2 *Let α be a finite measure on $\mathbb{R}^m$, $a := \alpha(\mathbb{R}^m)$. Moreover, let $\alpha_0(\cdot) = \alpha(\cdot)/a$. A random probability measure P with values in the space $\mathscr{P}(\mathbb{R}^m)$ of all probability distributions is a Dirichlet process on $\mathbb{R}^m$ with parameter α if, for any finite measurable partition $B_1, \ldots, B_k$ of $\mathbb{R}^m$*

$$(P(B_1), \ldots, P(B_k)) \sim D(\alpha(B_1), \ldots, \alpha(B_k)).$$

In what follows, we will write $P \sim DP(\alpha)$ for short, or equivalently $P \sim DP(a\alpha_0)$. If $P \sim DP(\alpha)$, it follows that $\mathbb{E}[P(A)] = \alpha_0(A)$ for any Borel set A, and then we say that α_0 is the prior expectation of P. It can be proved that such process exists. The two parameters defining the DP are then the weight parameter a, and the base measure α_0. Their role can be better understood thinking that, as we said above, for any Borel set A, $\mathbb{E}(P(A)) = \alpha_0(A)$, where following the Definition 5.3.2 we have that $\alpha_0(A) = \alpha(A)/a$, and $a = \alpha(\mathbb{R}^m)$, the total mass. Also, observe that $\mathbb{V}ar(P(A)) = \alpha_0(A)/(a+1)$, so that the prior is more tightly concentrated around its mean when a is larger, that is, the prior is more precise. Hence the parameter a can be regarded as the precision parameter.

Now, let $(b_1, b_2, \ldots, b_n)$ be a sample from a Dirichlet process P, i.e., $b_1, b_2, \ldots, b_n | P \overset{i.i.d}{\sim} P$, $P \sim DP(a\alpha_0)$. The Dirichlet prior is conjugate on $\mathscr{P}(\mathbb{R}^m)$; in fact, the posterior distribution of P, given $b_1, b_2, \ldots, b_n$, is

$$P | b_1, b_2, \ldots, b_n \sim DP\left(a\alpha_0 + \sum_{i=1}^{n} \delta_{b_i}\right) \tag{5.7}$$

In this case, marginalizing with respect to P, the predictive distribution of b_{n+1} given $b_1, \ldots, b_n$ can be described as follows:

$$\begin{aligned} b_1 &\sim \alpha_0 \\ b_{n+1} | b_1, \ldots, b_n &\sim \frac{a}{a+n}\alpha_0 + \frac{n}{a+n}\left(\frac{\sum_{i=1}^{n} \delta_{b_i}}{n}\right) \end{aligned} \tag{5.8}$$

The predictive distribution in (5.8) is called *Blackwell-MacQueen Urn Scheme* [17], since it is a mixture of the base-line measure α_0 and the previous observations. This means that there is a positive probability of coincident values for any finite positive a. Moreover if α_0 is an absolutely continuous probability measure, then b_{n+1} will assume a different, distinct value with probability $\frac{a}{a+n}$. Formula (5.8) allows us to sample (marginally) from P without simulating any trajectory of the Dirichlet process.

Again, equation (5.8) highlights the roles of scaling parameter a and base distribution α_0. The unique values contained in $(b_1, \ldots, b_n)$ are drawn independently from α_0, and the parameter a determines how likely b_{n+1} is to be a newly drawn value from α_0 rather than take on one of the values from $(b_1, \ldots, b_n)$. This equation also reveals the clustering property of the joint distribution of $(b_1, \ldots, b_n)$: there is a positive probability that each b_i will take on the value of another b_j, leading some of the variables to share values.

Sethuraman [131] provided a useful representation of the Dirichlet process. Its construction gives an insight on the structure of the process and provides an easy way to simulate its trajectories. Let consider two independent sequences of random variables $\{v_i\}_{i\geq 1}$ and $\{\xi_i\}_{i\geq 1}$ such that $v_i \overset{i.i.d.}{\sim} Beta(1, a)$ and $\xi_i \overset{i.i.d.}{\sim} \alpha_0$ (defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$), and define the following weights

$$\begin{cases} p_1 = v_1 \\ p_n = v_n \prod_{i=1}^{n-1}(1 - v_i) \quad n \geq 2. \end{cases}$$

It is straightforward to see that $0 \leq p_n \leq 1$, $n = 1, 2, \ldots$ and that $\sum_{n=1}^{\infty} p_n = 1$ a.s.. This construction is called *stick-breaking*. In fact p_1 represents a piece of a unit-length stick, p_2 represents a piece of the remainder of the stick and so on, where each piece is independently modeled as a $Beta(1, a)$ random variable scaled down to the length of the remainder of the stick. Now we can define a random variable P on $\mathscr{P}(\mathbb{R}^m)$.

$$P(A) = \sum_{n=1}^{\infty} p_n \delta_{\xi_n}(A), \qquad A \in \mathscr{B}(\mathbb{R}^m)$$

Sethuraman [131] proved that P has Dirichlet prior distribution, i.e., P is a Dirichlet Process with parameter $\alpha = a\alpha_0$. From this construction it is clear that a Dirichlet Process has discrete trajectories, i.e., if $P \sim DP(\alpha)$, then $\mathbb{P}(\{\omega : P(\omega) \text{ is discrete }\}) = 1$.

Moreover, let $(b_1, b_2, \ldots, b_n)$ be a sample from P, where $P \sim DP(\alpha)$. If K_n denotes the random variable representing the number of distinct values among $(b_1, b_2, \ldots, b_n)$, it can be proved [8] that the distribution of K_n is the following

$$\mathbb{P}(K_n = k) = c_n(k) n! a^k \frac{\Gamma(a)}{\Gamma(a+n)} \qquad k = 1, 2, \ldots, n, \tag{5.9}$$

where $c_n(k)$ is the absolute value of Stirling number of the first kind, for instance tabulated or computed by a software. From (5.9) it is clear that the mass parameter a has a great influence on the prior of the number of clusters. In particular, the larger a, the higher the prior number of components.

Now, coming back to the main focus of the section, the idea is to fit a GLME model for binary responses with a DP prior on the random effects parameters (here only considered on intercept, i.e., as additive grouping factor) in order to take advantage of the in-built classification the DP induce on them. This is the case we considered in [67]. So, let assume the base measure for the b_js is Normal, i.e., $\alpha_0 \equiv \mathscr{N}(0, \Sigma)$. Denote by $f(y_{ij}|\theta, b_j)$ the distribution of the binary outcome y_{ij} for group j at observation i in the logistic regression case, as given in (5.5). The prior specifications for the parameters of the DP-GLME are

$$\begin{aligned} \theta &\sim \mathscr{N}(\mu_0, \Sigma_0) \\ \mathbf{b}_j | P &\overset{i.i.d}{\sim} P \\ P &\sim DP(a\alpha_0) \end{aligned}$$

where θ and P are assumed independent a priori. If P has the form reported above, it is impossible to write down the joint posterior density of parameters, because there is not a common dominating measure. Anyway, it is possible to obtain the full conditional distributions needed for Gibbs sampling, as detailed in [92]. In fact, it hs been shown that

$$f(\theta|\mathbf{b}, \mathbf{y}) \propto exp\left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} \log f(y_{ij}|\theta, b_j) - \frac{1}{2}(\theta - \mu_0)^T \Sigma_0^{-1} (\theta - \mu_0) \right).$$

as far as the full conditional of θ is concerned. Unless y_{ij} has the Normal distribution, sampling from this distribution is not straightforward, but it can still be accomplished, using for example a Metropolis step. Moreover, concerning the full conditionals of the random effects, we have

$$\begin{aligned} f(b_j|\theta, \mathbf{y}, \mathbf{b}_{-j}) &\propto \sum_{i \neq j}^{J} exp\left\{ \sum_{i=1}^{n_j} \log f(y_{ij}|\theta, b_j) \right\} \delta_{b_j} + \left[\alpha \int exp\left\{ \sum_{i=1}^{n_j} \log f(y_{ij}|\theta, b_j) \right\} \right. \\ &\times \left. \phi(b_j) db_j \right] \phi(b_j) \end{aligned}$$

where $\mathbf{b}_{-j}$ denotes the random effects for the groups excluding group j, δ_s a degenerate distribution with point mass at s and ϕ indicates the Gaussian density of α_0. An important subtlety in the formula above is that the terms $f(y_{ij}|\theta, b_j)$ in the first summation use data from group j and the random effects for each of other groups. That is, it evaluates the likelihood for group j using the other groups random effects. The better the fit of group j random effect, the greater the likelihood then the more likely it is that δ_{b_j} is the distribution from which b_j is drawn.

From the full conditional above, it can be evinced that the density structure for each random effect is similar to the one already seen in (5.8), i.e., a mixture of a new value for the parameter with the already exsisting ones. This leads to a natural clustering among random effects, and consequently among what they represent in th model, achieved using the sample's bias of b_js.

In the analyses carried out in Part III and in particular in Section 7.4, we will deal with a DP-GLME model where the binary oucome y_{ij} will be the survival of patient i admitted in j-th provider with STEMI diagnosis, and a nonparametric distribution (specifically DP prior) will be assumed for the additive random effect on grouping factor, i.e., for the hospital of patient admission.

5.3.2 Dependent Dirichlet Process for classification

Dependent Dirichlet Processes (DDPs) generalize the DP to allow for a collection of non parametric distributions, the realization of which are dependent, with the level of covariates governing the degree of dependence. The idea that drives the dependence for one dimensional DP is that, in the presence of a covariate, the location of the DP can be replaced by the sample path of the corresponding stochastic process. This sample path provides the location and the weights in Sethurman's representation at each value of the covariate.

Starting from the setting proposed in the previous Section, let consider a family of random measures $\mathscr{F} = \{F_z, z \in \mathscr{Z}\}$ indexed by a covariate z. MacEachern in [108] defined a probability model for $\mathscr{F}$ such that, marginally, for each z, $F_z = \sum w_h \delta_{b_{zh}}$ follows a DP. In the basic DDP model, the weights w_h are common to all F_z and are defined as in the stick-breaking procedure. The DDP induces dependence across z by assuming that $b_h = (b_{zh}, z \in \mathscr{Z})$ are i.i.d. realizations of a stochastic process (as a function of z). Independence across h, together with the stick-breaking prior for the weights w_h, guarantees that F_z marginally follows a DP.

In what follows and in Section 7.4 we will use this DDP structure to develop an ANOVA-like probability model over an array of random distributions.

Therefore, as in [33], assume $\mathscr{F} = (F_z, z \in \mathscr{Z})$ is an array of random distributions, indexed by a categorical covariate $\mathbf{z}$. For semplicity of explanation, assume that $\mathbf{z} = (u, v)$ is bivariate with $u \in \{1, \ldots, U\}$ and $v \in \{1, \ldots, V\}$. The covariates (u, v) could be, for example, the levels of two treatments in a clinical trial, and the distributions $F_{\mathbf{z}}$ may be sampling distributions for random effects. In this context we wish to develop a probability model for the random distributions $F_{\mathbf{z}}$ that will enable us to build an ANOVA-type dependence structure. For example, we want the random distributions $F_{\mathbf{z}}$ and $F_{\mathbf{z}'}$ for $\mathbf{z} = (u_1, v_1)$ and $\mathbf{z}' = (u_1, v_2)$ to share a common main effect due to the common factor u_1. The model should allow us to incorporate prior information about the presence of interaction between the covariates. If interactions are present, the effect of $u = u_1$ should be allowed to depend on the level of the other covariate v. The following model gives a formal definition to notions like "main effect" and "interaction". Briefly, instead of a nonzero additive effect on the mean of the response variable in an ANOVA model, an effect is recast as a difference in distribution of some quantity that has, in turn, an impact on the distribution of the final response (see [33] fur further details). Thus, the models we create allow us to transfer both the interpretation

and the structure used for unknown normal means in the traditional ANOVA model to unknown random distribution functions. Like standard ANOVA models the proposed model can be justified by a judgment of partial exchangeability for observed data. Assume y_{zi}, $i = 1,2,3,\ldots$, are observed data, indicating the i-th observation under condition z. If, for each z, the subsequence $(y_{z1}, y_{z2}, \ldots)$ is judged exchangeable, then by de Finetti's representation theorem y_{zi} can be assumed to be an i.i.d. sample from distribution F_z. More technical details on ANOVA-DDP models and their properties can be found in [33] or in the references therein.

The model we consider for our application can be derived as follows: let $F_{\mathbf{z}} = \sum w_h \delta_{b_{\mathbf{z}h}}$ for $\mathbf{z} = (u,v)$. We assume Sethuraman's stick breaking prior for the common weights as in the Dirichlet case:

$$\frac{w_h}{\prod_{i=1}^{h-1}(1-w_i)} \sim Beta(1,a) \qquad h = 2,3,\ldots,$$

where $w_1 \sim Be(1,a)$ We impose an additional structure on the locations $b_{\mathbf{z}h}$:

$$b_{\mathbf{z}h} = m_h + A_{uh} + B_{vh} \tag{5.10}$$

As in standard ANOVA models, we need to introduce an identifiability constraint for interpretability. We may impose any of the standard constraints, for example $A_{1h} = B_{1h} = 0$. For the remaining parameters we assume $m_h \sim \alpha_m^0(m_h)$, $A_{uh} \overset{i.i.d.}{\sim} \alpha_{A_u}^0(A_{uh})$ and $B_{vh} \overset{i.i.d.}{\sim} \alpha_{B_v}^0(A_{vh})$ with independence being across h, u and v. We refer to the joint probability model on $\mathscr{F}$ as $(F_z, z \in \mathscr{Z}) \sim$ ANOVA-DDP(a, α^0). Marginally, for each $(\mathbf{z} = (u,v))$, the random distribution F_z follows a DP with mass a and base measure $\alpha_{\mathbf{z}}^0$, given by the convolution of α_m^0, α_{Au}^0 and α_{Bv}^0. Model (5.10) defines dependence across $\mathbf{z}$ by defining the covariance structure of the point masses b_{zh} across z. As in standard ANOVA the structural relationships are defined by the additive structure (5.10) and the level of the dependence is determined by the variances in α_m^0, α_{Au}^0 and α_{Bv}^0.

Note that the ANOVA DDP model introduces an additional level of uncertainty by defining the random measures $(F_z, F_{z'})$. The resulting covariance structure remains unchanged eccept for the attenuation factor $1/(a+1)$, corresponding to the additional uncertainty about F_z. The same result remains true for arbitrary ANOVA structure, including more factors and possibly interactions. Moreover, model (5.10) is not constrained to univariate distributions F_z. The point masses b_{zh} and the ANOVA effects m_h, A_{uh}, and B_{vh} can be q-dimensional vectors. This is important, for example, if the random distributions F_z will represent the random effects in a hierarchical model.

5.4 Optimal decision rules for hospital ranking and classification

Investigations on surgical performance have always adopted routinely collected clinical data to highlight unusual provider outcomes. In addition, there are a number of regular reports using routinely collected data to produce indicators for hospitals. As well as highlighting possible high- and low-performers, such reports help in understanding the reasons behind variation in health outcomes, and provide a measure of performance which may be compared with benchmarks or targets, or with previous results to examine trends over time. Statistical methodology for provider comparisons has been developed in the context of both education and health.

The statistical components of such an analysis generally comprise a model for the provider effects that adjusts for differences between patient risks either through standardisation methods or incorporating covariates. In these cases, "provider" is used in a very general sense and might be a hospital, a health-care authority, or even an individual surgeon. It is known that pursuing the issue of adjustment for patient severity (case-mix) is a challenging task, since it requires a deep

knowledge of the phenomenon from a clinical, organizal, logistic and epidemiological point of view. However, this is the reason why it is always expected to be inadequate and therefore unavoidable residual variability (over-dispersion) will generally exist between providers. It is then crucial that a statistical procedure is able to assess whether a provider may be considered "unusual". In particular, note that although hierarchical models are recommended since they account for the nested structure in describing hospital performance, it is not straightforward how assessing unusual performance.

Studies of variations in health care utilization and outcomes involve the analysis of multilevel clustered data. Those studies quantify the role of contributing factors (patients and providers) and assess the relationship between health-care processes and outcomes. In [65], we develop Bayes rules for several families of loss functions for hospital report cards when Bayesian Semiparametric Hierarchical models are used, and discuss the impact of assuming different loss functions on the number of hospitals identified as "non acceptably performing". The analysis is carried out on a case study dataset arising from MOMI^2 survey (see Paragraph 3.2.2 and Section 8.3 for details on data description and analysis respectively) on patients admitted with STEMI to one of the hospitals of the Milan Cardiological Network. The major aim consists of the comparison among different loss functions to discriminate among health care providers' performances, together with the assessment of the role of patients' and providers' characteristics on survival outcome.

The model we assume here is a DP-GLME model where, for unit (patient) $i = 1,\dots,n_j$ in group (hospital) $j = 1,\dots,J$, Y_{ij} is a Bernoulli random variable with mean p_{ij}, i.e.,

$$Y_{ij}|p_{ij} \overset{ind}{\sim} Be(p_{ij}).$$

According to the DP-GLME model, the p_{ij}s are modelled through a logit regression of the form

$$logit(p_{ij}) = \log\frac{p_{ij}}{1-p_{ij}} = \theta_0 + \sum_{h=1}^{p}\theta_h x_{ijh} + \sum_{l=1}^{J} b_l z_{jl} \tag{5.11}$$

where $z_{il} = 1$ if $i = l$ and 0 otherwise. In this model, $\boldsymbol{\theta} = (\theta_0,\dots,\theta_p)$ represents the $(p+1)$-dimensional vector of the fixed effects, $\mathbf{x}_{ij}$ is the vector of patient covariates and $\mathbf{b} = (b_1,\dots,b_J)$ is the vector of the additive random-effects parameters of the grouping factor. Assuming a suitable prior for $b_1,\dots,b_J$, the b_js are i.i.d. according to this; the prior for $\boldsymbol{\theta}$ is parametric. If we assume a setting like the one proposed in (5.11), we saw how to take advantage of the bias of random effect distributions to cluster the random effects themselves. We show now how such a framework can be encompassed within a Bayesian decision analysis framework. In fact, once the posterior distribution of the random effect has been obtained, suitable loss functions can be defined in order to *a posteriori* weigh the decision of wrongly classifying the hospital as having acceptable or unacceptable performances.

The random intercepts of model (5.11), i.e., $\theta_0 + b_1, \theta_0 + b_2, \dots, \theta_0 + b_J$ represent the hospital performances quantifying the contribution to the model after patients' covariates adjustment. Let us denote by β_j the sum of θ_0 and b_j. We consider the class of loss functions

$$L(\beta_j, d) = c_I \cdot f_1(\beta_j)\cdot d\cdot \mathbb{I}(\beta_j > \beta_t) + c_{II}\cdot f_2(\beta_j)\cdot(1-d)\cdot\mathbb{I}(\beta_j < \beta_t), \tag{5.12}$$

where d is the decision to take ($d = 1$ means that the hospital has "unacceptable performances", $d = 0$ stand for "acceptable performances"), c_I is the weight assigned to the cost $f_1(\beta_j)$, occurring for a false positive, c_{II} is the weight assigned to cost $f_2(\beta_j)$, occurring for a false negative and β_t is defined as $\log(p_t/(1-p_t))$, being p_t a reference value for survival probabilities.

Without loss of generality, we can assume a proportional penalization, i.e., $f_2(\beta_j) = k \cdot f_1(\beta_j)$, taking k as the ratio c_{II}/c_I. In this sense, the parameter k quantifies our beliefs on cost, being greater than 1 if we credit that accepting a *false negative* should cost more than rejecting a true negative and less than 1 otherwise. An acceptable performance is then defined comparing the posterior expected losses associated with the decision that the hospital has *acceptable performances*

$$R(\mathbf{y}, d = 0) = E_\pi\left(L(\beta_j, d = 0)|\mathbf{y}\right) = \int f_2(\beta_j)\mathbb{I}(\beta_j < \beta_t)\Pi(\beta_j|\mathbf{y})d\beta_j$$

and the decision that the hospital has *unacceptable performances*

$$R(\mathbf{y}, d = 1) = E_\pi\left(L(\beta_j, d = 1)|\mathbf{y}\right) = \int f_1(\beta_j)\mathbb{I}(\beta_j > \beta_t)\Pi(\beta_j|\mathbf{y})d\beta_j.$$

$\Pi(\beta_j|\mathbf{y})$ denoting here the posterior distribution of β_js. In short, we classify an hospital as being *acceptable* (or with *acceptable performances*) if the risk associated with the decision $d = 0$ is less than the risk associated with the decision $d = 1$, i.e., if $R(\mathbf{y}, d = 0) < R(\mathbf{y}, d = 1)$. Within this setting, in [65] the comparison of four different loss functions is proposed, to address the decision problem of judging hospitals performances. Results of the comparison and of the sensitivity analysis carried out on parameter k are shown in Section 8.3.

The application of this theoretical setting to the problem of managing a Cardiological Network as presented in Section 8.3 is an example of how Bayesian decision theory could be employed within the context of clinical governance of Regione Lombardia. It may point out where investiments are more likely to be needed, and could help in not to loose opportunities of quality improvement.

5.5 Problems due to unbalanced share: a Bayesian solution

In Section 4.7 we discussed problems arising from unbalanced shares in logistic regression frameworks within the frequentist setting, and how they reflect on the predictive power of the models we considered and used for the analyses. The same problem arises also adopting a Bayesian approach. Anyway in this case, taking advantage of posterior predictive distributions of estimated responses, a new method for carrying out classification and prediction is proposed (see [64]), based on posterior predictive credibility intervals (CIs). As we will see in Section 7.4, this method seems to be able to increase the predictive power of generalized linear mixed effects models.

So, let consider the problem of predictive performances evaluation for a generalized linear mixed effects model. Once such a model is fitted to grouped data, estimated success probabilites p_{ij} for each given unit i belonging to the group j can be computed.

So let Y_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be a binary outcome assuming value 1 if a success is observed, 0 otherwise. In this case, i denotes the statistical units that belong to any of the J groups. When a generalized linear mixed effect model with an additive random effect on grouping factor is fitted to such data, the usual way to make prediction is then based on point estimates summarizing the posterior predictive distributions of success probabilities, i.e., on the comparison among these estimates with repect to a given threshold. In particular we could classify a unit i belonging to group j as "success" if $\mathbb{E}(p_{ij}|\boldsymbol{Y})$ is bigger than a given cut-off point. Since the classification is typically sensitive to the cut-off point, there are several criteria to choose the cut-off point. In [45] a review and comparison of the most popular criteria is proposed. In our application, since our dataset is particularly unbalanced, if we consider the standard cut-off point equal to 0.5, we would obtain a very low overall misclassification rate, but also a bad negative predictive power. A first remedy is

proposed in [28], where a cut-off point equal to the survival sample proportion $\bar{p}$ is adopted (see Paragraph 4.7.2). The false positive and false negative rates are more balanced than using a cut-off point of 0.5, but this does not guarantee an improvement in the overall misclassification rate. The overall misclassification rate as in [28] can be considered as a goodness of fit index since it is less dependent on the unequal sample proportion. On the other hand, instead of choosing a given cut-off point, we could plot the ROC curve. However, we believe that results coming out from comparison of pointwise estimates with any thresholds cannot be considered completely satisfactory. In fact, on one hand those cut-off criteria are not robust in case of a very unbalanced data-set. On the other hand, even if the ROC curve is indipendent of a given threshold, it can be used as a comparison tool among models rather than as a predictive tool itself.

To this aim, the idea we propose for improving the predictive power of logistic mixed effects models when shares are strongly unbalanced takes advantage of Bayesian approach: in fact, interval estimate is richer than point estimate which does not provide any information on the prediction uncertainty. The new approch adopts the entirely interval estimates, computed starting from posterior distributions. We then classify a statistical unit as "success" if the entirely α% interval estimate of the posterior predictive probability p_{ij} is over a given cut-off point, as "dead" if the entirely interval estimate is below the cut-off point and we do not classify it if the cut-off point lies in the CI. The higher is the credible level, the more patients will belong to the Uncertainty Class (UC).

In the application that motivated the development of this new method (recall that in STEMI we deal with a share of in-hospital mortality rates nearly equal to 4%), this solution came out to be much more effective than the corresponding pointwise one (see Section 7.4). This is really interesting from both statistical and clinical points of view, since the negative predictive power in the application of interest is the model ability in predicting when patients are likely to die. This ability is clinically much more relevant than the opposite one, i.e., the ability of predicting which patients are likely to survive. We are actually planning to set a battery of simulations for testing the method in order to prove and quantify its effectiveness, especially in increasing negative predictive power of models in the case of strongly unbalanced shares.

Chapter 6

Statistical models for healthcare: more complex data

In this chapter, statistical methods for dealing with multivariate functional data and longitudinal event-dependent data arising from clinical diagnostic devices and from the integration of clinical registries and administrative database are proposed. These methods have been applied to data coming from PROMETEO datawarehouse (presented in Section 3.2.4) and to the integrated database arising from the linkage between the STEMI Archive and Regione Lombardia Administrative datawarehouse (discussed in Paragraph 3.3.2). In the first case, two different targets have been pursued: the unsupervised classification of multivariate functional data and outlier detection, as reported in [80], [81] and [82]. In the second case, the idea is to point out suitable models for predicting binary outcomes also through functional covariates. An example of application of such techniques can be found in [14] and [15].

ECG signals can be considered as multivariate functional data with dependent components (see Paragraph 3.2.4). In this context, some issues of interest are, for example, classification of groups of curves with similar morphological patterns, multivariate functional outliers detection within a homogeneous group and classical inference on mean and quantiles of subpopulations. From a clinical point of view, the first issue concerns how to carry out a semi automatic diagnosis based only on the morphological deviations from physiological patterns induced by the presence of the disease of interest; the second one leads to profile "typical" curve expression for each pathology; finally the third one allows for the investigation of the presence of statistically significant differences in the subpopulations of pathological units with respect to physiological ones. On the other hand, the hospitalization process of each patient may be considered itself, for implementation of time to event models, or in connection with the output measured by a clinical survey, where it is modeled as counting process and considered as functional predictors for outcomes of interest.

Making inference on the underlying process generating any kind of curve (multivariate functional curves like ECGs, or counting processes like hospitalization processes and so on) has the following meanings: in the real world, it means discriminating among effect that pathologies or phenomena under examination induce in the observed data; in the statistical world, it means constructing a model able to captures the real phenomenon as well and reliably as possible.

6.1 Unsupervised classification of multivariate functional data

In [80] a method for the analysis and the classification of ECG curves starting from their sole morphology is proposed. The main goal is to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria.

In what follows, an overview of smoothing and registration techniques for functional data is given. The exhaustive tractation of these topics are beyond the scope of this thesis, so we just set the context where techniques we used come from and motivations that guided our choices in analyzing data as in [80].

6.1.1 An overview of smoothing and registration techniques for functional data

In statistics and image processing, to smooth a dataset is to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. A wide choice of different algorithms and methods are available for smoothing, among others Smoothing spline, Local regression also known as “loess”, Convolution, Wavelets, Laplacian smoothing, Kalman filter etc.. Among these, Wavelet analysis is an accurate and reliable tool for studying signals with sudden changes of phase and frequency. It is useful for audio/image/video analyzing and processing, data compression, signal smoothing and denoising, speech recognition and biomedical imaging. Concerning functional data like ECG signals can thought to be, Wavelet bases seem suitable to be used because every basis function is localized both in time and in frequency, being therefore able to capture ECG strong localized features (peaks, oscillations...). A systematic introduction to wavelets can be found in [110].

Wavelet bases have been so far mainly applied in problems where there was no interest in derivatives, because of the absence of close analytical forms for smooth wavelet bases. This issue has restricted their application to a small part of the Functional Data Anlaysis (FDA) field. To overcome this limitation, in [120] a numerical method that allows to obtain derivatives of wavelet estimated data is presented. This method holds in general for multivariate functional signals, allowing for a joint smoothing of all the components of the multivariate functional signal. Since the eight leads of an ECG jointly describe the same phenomenon (i.e., the heart dynamic), when smoothing these data it is appropriate to use a technique which takes into account all the eight leads simultaneously. This helps in detecting significant features, which reflect on more then one leads.

When the functional response evolves with respect to time, the subject may experience events at different times with the consequence that the sample curves are not aligned in some sense. In these cases, any estimator will fail in the attempt to be satisfactory. Curve registration is a way to solve this problem. Moreover, functional observations usually show both phase and amplitude variation, i.e., each curve has its own biological time so that same features can appear at different times among the statistical units. It is well known that a correct separation between these two kind of variability is necessary for a successful analysis [126].

Example of registration techniques are, among others, Landmarks registration, Continuous Monotone registration, Dynamic time warping, Semi-parametric registration, Shape invariant models. As shown in Section 9.1 and in [80], we address the problem of correctly separating the different kind of variability through a registration procedure based on landmarks, since in the ECG signals case they are provided and have a specific biological meaning (start/end points of each ECG wave).

The smoothed and registered curves may be clustered in groups solving a suitable optimization

problem that takes into account the distance in H^1 space between curves. So doing, we are able to implement a semi automatic diagnosis of ECGs based on statistical techniques, as detailed in Section 9.1 and in [80].

6.2 Depth measure for multivariate functional data and outlier detection

A challenging task in functional data analysis is to provide an ordering within a sample of curves that allows the definition of order statistics, such as ranks and L-statistics [43]. A natural tool to analyze these functional data features is the idea of statistical depth, which provides a measure of centrality or outlyingness of an observation with respect to a given dataset or a population distribution. Several definition of depth measures for multivariate data have been proposed and analyzed in literature (see [102], [109], [141], [149] and [150] among others). A generalization to functional data is given in [106], starting from depth measures for multivariate data. They also provide the extension of robust statistics to a functional framework, generalizing properties of depth measures which are proved to hold in multivariate case (see [102], [130] and [149] for further details on multivariate setting). Finally a specific focus on trimmed means for functional data can be found in [44], where a generalization of some issues treated in [43] about multivariate L-estimation is proposed. Once a depth measure is associated with each univariate or multivariate functional data within a sample, it is possible to rank them as well as to visualize graphically the result of ranking through functional boxplots, as proposed in [137] and [81], for univariate and multivariate functional data respectively.

There are lots of different aims which lead to rank curves according to suitable depth indexes. Indeed, several applications focus on classification of functions arising from different population and make inference about the latent differences among them analyzing the morphological effects they induce on the curves shape. This is usually carried out without parametric assumptions on the model which the sample of curves is associated with, like in [29] and [106]. On the other hand, sometimes the interest is in making inference on specific summary statistics, as proposed in [99] for the multivariate setting. In [81] and [82] we deal with multivariate functional observations, i.e., statistical units where each component is a curve. Firstly, a generalization of the concept of depth for functional data to the multivariate functional case is provided, then suitable generalizations of nonparametric statistics for ranking and classifying multivariate curves are defined, in order to make inference on them. So a new concept of a multivariate index of depth, derived from averaging univariate centrality measures for functional data in a suitable multivariate index is proposed, analyzed, and applied. The employment of the functional boxplots is widen, adopting this graphical tool also in the more complex case of samples of multivariate functions. Then the Wilcoxon rank test based on the order induced by the multivariate functional depth is proposed to test differences between groups of multivariate curves. In fact, two are the main goals of the analysis: the first one is to point out a suitable method for performing outliers detection in a multivariate functional setting, within a sample of curves arising from the same population (temptative analyses in this sense are proposed for the univariate functional case in [38]); the second one is to carry out non parametric test for comparing samples of multivariate curves and making inference on the corresponding populations.

A natural and motivating application of this theoretical framework comes from the biomedical context, and in particular from applications that deal with cardiovascular diseases diagnoses carried out using Electrocardiographic (ECG) devices.

6.2.1 Band depth and inference for multivariate functional data

As mentioned in the previous section, a natural tool to analyze and rank functional data is the idea of statistical depth, which measures the centrality of a given curve within a group of trajectories providing center-outward orderings of the set of curves itself. In general, several different definitions of depth can be given [149]. In our case, we refer to the band depth measure for functional data proposed by [105] and [106].

Let X a stochastic process with law P taking values on the space $\mathscr{C}(I)$ of real continuous functions on the compact interval I. The graph of a function $f \in \mathscr{C}(I)$ is the subset of the plane $G(f) = \{(t, f(t)) : t \in I\}$. The random band depth, of order $J \geq 2$, for a function $f \in \mathscr{C}(I)$ is then

$$BD^J_{P_X}(f) = \sum_{j=2}^{J} P_X\{G(f) \subset B(X_1, X_2, ..., X_j)\},$$

where $B(X_1, X_2, ..., X_j)$, for $j = 2, ..., J$ is the random band in $\mathbb{R}^2$ delimited by $X_1, ..., X_j$, independent copies of the stochastic process X, defined as

$$B(X_1, ..., X_j) = \{(t, y(t)) : t \in I, \min_{r=1,...,j} X_r(t) \leq y(t) \leq \max_{r=1,...,j} X_r(t)\}$$

We propose a new definition of a band depth measure for multivariate functional data, i.e., data generated by a stochastic process $\mathbf{X}$ taking values in the space $\mathscr{C}(I;\mathbb{R}^s)$ of continuous functions $\mathbf{f} = (f_1, ..., f_s) : I \to \mathbb{R}^s$.

Definition 6.2.1 *Let* $\mathbf{f}$ *be a function on* I *taking values in* $\mathbb{R}^s$*. The multivariate band depth measure is then defined as*

$$BD^J_{P_{\mathbf{X}}}(\mathbf{f}) = \sum_{k=1}^{s} p_k BD^J_{P_{X_k}}(f_k), \qquad p_k > 0 \text{ for } k = 1, ..., s, \qquad \sum_{k=1}^{s} p_k = 1. \tag{6.1}$$

Let $\mathbf{X}$ a multivariate random process such that $P(\min_{k=1,...,s} \|X_k\|_\infty > M) \to 0$ as $M \to \infty$, then it is easy to prove, using the properties of the functional depth measure summarized in [106], the following results on the basic properties of the multivariate band depth measure defined in (6.1).

Proposition 6.2.1

(a) Let $T(\mathbf{f}) = \mathbf{A}(t)\mathbf{f}(t) + \mathbf{b}(t)$*, where* $\forall t \in I$ $\mathbf{A}(t)$ *is a* $s \times s$ *diagonal matrix such that* $\mathbf{A_{kk}}(t)$ *are continuous functions in* I*, with* $\mathbf{A_{kk}}(t) \neq 0$*, for each* $t \in I$*, and* $\mathbf{b}(t) \in \mathscr{C}(I;\mathbb{R}^s)$*. Then* $BD^J_{P_{T(\mathbf{X})}}(T(\mathbf{f})) = BD^J_{P_{\mathbf{X}}}(\mathbf{f})$*.*

(b) $BD^J_{P_{\mathbf{X}(g(t))}}(\mathbf{f}(g(t))) = BD^J_{P_{\mathbf{X}(t)}}(\mathbf{f}(t)) =$ *when* g *is a one-to-one transformation of the interval* I*.*

(c) $\sup_{\min_{k=1,...,s} \|f_k\|_\infty > M} BD^J_{P_{\mathbf{X}}}(\mathbf{f}) \to 0$ *as* $M \to \infty$*.*

(d) If $\forall k = 1, ..., s$ *the probability distribution* P_{X_k} *on* $\mathscr{C}(I)$ *has absolutely continuous marginal distributions, then* $BD^J_{P_{\mathbf{X}}}$ *is a continuous functional on* $\mathscr{C}(I;\mathbb{R}^s)$*.*

Proof. (a) using Definition 6.2.1 and the property (1) of Theorem 3 in [106] we have

$$BD^J_{P_{T(\mathbf{X})}}(T(\mathbf{f})) = \sum_{k=1}^{s} p_k BD^J_{P_{A_{kk}X_k + b_k}}(A_{kk} f_k + b_k) = \sum_{k=1}^{s} p_k BD_{X_k}(f_k) = BD^J_{P_{\mathbf{X}}}(\mathbf{f}).$$

The diagonality requirement on matrix A means that the multivariate functional depth measure $BD^J_{P_{\mathbf{X}}}(\mathbf{f})$ is invariant as regards affine transformations of each component taken one by one, without combining different elements of the multivariate function.

(b) follows directly from property (2) of Theorem 3 in [106].

(c)

$$\sup_{\min_{k=1,\dots,s}\|f_k\|_\infty > M} BD^J_{P_{\mathbf{X}}}(\mathbf{f}) = \sup_{\min_{k=1,\dots,s}\|f_k\|_\infty > M} \sum_{k=1}^{s} p_k BD_{X_k}(f_k)$$

and each term in the sum over components goes to zero when M goes to infinity.

(d) also this point follows directly from property (4) of Theorem 3 in [106].

□

If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent copies of the stochastic process $\mathbf{X}$, the sample version of (6.1) can be introduced in order to conduct descriptive and inferential statistical analyses on a set of multivariate functional data $\mathbf{f}_1, \dots, \mathbf{f}_n$ generated by the process $\mathbf{X}$. For any $\mathbf{f}$ in the sample $\mathbf{f}_1, \dots, \mathbf{f}_n$ we can compute the depth as

$$BD^J_n(\mathbf{f}) = \sum_{k=1}^{s} p_k BD^J_{n,k}(f_k),$$

where for the function $f_k \in \mathscr{C}(I)$

$$BD^J_{n,k}(f_k) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} \mathbb{I}\{G(f_k) \subset B(f_{i_1;k}, \dots f_{i_j;k})\}$$

and $\mathbb{I}\{G(f_k) \subset B(f_{i_1;k}, \dots, f_{i_j;k})\}$ indicates if the band determined by $(f_{i_1;k}, \dots, f_{i_j;k})$ contains the whole graph of f. The k component of the vector $\mathbf{f_i}$ is denoted by $f_{i;k}$.

Proposition 6.2.2 *The sample version of multivariate functional depth is consistent, in fact*

$$|BD^J_n(\mathbf{f}) - BD^J_{P_{\mathbf{X}}}(\mathbf{f})| \to 0, \ a.s. \ \ if \ \ n \to \infty \tag{6.2}$$

Proof.

$$|BD^J_n(\mathbf{f}) - BD^J_{P_{\mathbf{X}}}(\mathbf{f})| = |\sum_{k=1}^{s} p_k BD^J_{n,k}(f_k) - \sum_{k=1}^{s} p_k BD_{X_k}(f_k)| \le \sum_{k=1}^{s} p_k |BD^J_{n,k}(f_k) - BD_{X_k}(f_k)| \tag{6.3}$$

and each term of the sum in the last term of (6.3) goes to zero as stated in Theorem 4 of [106]. □

As proposed in [106] also in this multivariate functional setting we can move to the analogous of the modified band depth:

$$MBD^J_n(\mathbf{f}) = \sum_{k=1}^{s} p_k MBD^J_{n,k}(f_k), \tag{6.4}$$

where for the function $f_k \in \mathscr{C}(I)$ the modified band depth measures the proportion of time that the curve f_k is in the band, i.e.,

$$MBD^J_{n,k}(f_k) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \le i_1 < i_2 < \cdots < i_j \le n} \tilde{\lambda}\{E(f_k; f_{i_1;k}, \dots, f_{i_j;k})\},$$

where $E(f_k) =: E(f_k; f_{i_1;k}, \dots, f_{i_j;k}) = \{t \in I, \min_{r=i_1,\dots,i_j} f_{r;k}(t) \le f_k(t) \le \max_{r=i_1,\dots,i_j} f_{r;k}(t)\}$ and $\tilde{\lambda}(f_k) = \lambda(E(f_k))/\lambda(I)$ and λ is the Lesbegue measure on I. As stated in [106] the values of

the modified band depth measure are stable with respect to the choice of J, and in order to be computationally faster we set $J = 2$ and we denote $MBD_n^J(\mathbf{f})$ as $MBD(\mathbf{f})$ in the following. The use of the modified band depth measure avoids also having too many depth ties.

Given the multivariate band depth measure defined in (6.4), a sample of multivariate functional data $\mathbf{f}_1, ..., \mathbf{f}_n$ can be ranked. In the following we denote $\mathbf{f}_{[i]}$ the sample curve associated with the i-th largest depth value, so $\mathbf{f}_{[1]} = \text{argmax}_{\mathbf{f} \in \{\mathbf{f}_1, ..., \mathbf{f}_n\}} MBD(\mathbf{f})$ is the *median* (deepest and more central) curve, and $\mathbf{f}_{[n]} = \text{argmin}_{\mathbf{f} \in \{\mathbf{f}_1, ..., \mathbf{f}_n\}} MBD(\mathbf{f})$ the most outlying one.

The idea of generalizing the concept of functional boxplot to multivariate functional data is based on the new definition of multivariate functional depth measure given in (6.4) which takes into account simultaneously the behaviour of all the s components of $\mathbf{f}$ weighting in a suitable way the components in order to take into account correlations among them. The same holds when the goal is to carry out multivariate functional outliers detection, to be used for example to robustify training set adopted in unsupervised classification algorithms. In such cases, the following steps should be implemented on multivariate curves sample $\mathbf{f}_1, ..., \mathbf{f}_n$:

1. For each statistical unit j, compute the value of measure depth $MBD(\mathbf{f_j})$;
2. Rank the multivariate functions $\mathbf{f}_j(t)$ according to the value of multivariate depth measure and define outliers those curves that, for at least one t, are outside the fences obtained inflating the envelope of the α% central region by h times the range of the α% central region. In particular the α% central region for the component f_k determined by a sample of curves is defined as

$$\mathscr{C}_\alpha = \left\{ (t, y(t)) : \min_{r=1,...,\lceil \alpha n \rceil} f_{[r];k}(t) \leq y(t) \leq \max_{r=1,...,\lceil \alpha n \rceil} f_{[r];k}(t) \right\}$$

   where $\lceil \alpha n \rceil$ is the smallest integer greater than or equal to αn. In the following we set $\alpha\% = 50\%$ and $h = 1.5$.

3. Visualize the functional boxplot of each component, building the envelope of the 50% deepest functions and then the functional boxplot according to the ranking arising from the multivariate index previously pointed out.

Notice that this algorithm defines outliers according to a multivariate index of depth, which takes into account simultaneously the depth of all components of the multivariate function. This implies that the envelope of the central region is composed of the same α% most central curves, with respect the multivariate index of depth, in each component.

Given the order in the sample of curves induced by the multivariate functional depth measure, the definition of *trimmed mean* following, for example, [44] can be extended to multivariate functional data straightforwardly. We can also widen to this framework a non parametric rank test to compare two samples of multivariate functions. In particular consider a sample $\mathbf{f}_1, ..., \mathbf{f}_n$ generated according to a distribution $P_{\mathbf{X}}$ and another sample $\mathbf{g}_1, ..., \mathbf{g}_m$ generated according to a distribution $P_{\mathbf{Y}}$. We want to test differences between the two populations; combine the two samples, that is, let $W = \mathbf{w}_1, ..., \mathbf{w}_{n+m} \equiv \mathbf{f}_1, ..., \mathbf{f}_n, \mathbf{g}_1, ..., \mathbf{g}_m$. We can assign to each element of the combined set a rank according to values of the multivariate functional depth, and in particular the higher the depth the lower the rank. The proposed test statistics R is the sum of the ranks of the second sample $\mathbf{g}_1, ..., \mathbf{g}_m$ with respect to the combined set W ($R = \sum_{j=1}^m \text{Rank}_W(\mathbf{g}_j) \equiv \sum_{j=1}^m r(\mathbf{g}_j)$). If according to the null hypothesis (H_0) there is no differences between the distributions generating the data $(r(\mathbf{g}_1), ..., r(\mathbf{g}_m))$

can be viewed as a random sample size m drawn without replacement from the set $(1,...,n+m)$, and we reject H_0 for values of R too small or too high. For large values of n and m it is possible to use a Normal approximation (see [99]).

Such test represents a quantitative method for carrying out inference in a supervised multivariate functional clustering framework. On the other hand, for the unsupervised clustering case, it can be also seen as a way to test if the process generating the outliers pointed out by the functional boxplot can be considered as different from the process generating the curves of the α% most central region.

In summary, in this section we generalize the notion of depth for functional data presented in [106] to the multivariate functional case and define also a new multivariate functional index of depth which is able to take into account jointly the depth of the multivariate functional data on each component. This provides a center-outward ordering criterion for a sample of multivariate functions. Extensions and proofs of the properties of the new index are also provided, as well as for its modified version. A generalization of the non parametric test to this framework has been adopted to carry out inference in a supervised clustering context. The application of the new index to a real case of ECG signals proposed and discussed in [81] and in [82] and detailed in Section 9.2, highlights how the methodology works effectively both in detecting outliers and in distinguishing between samples arising from different underlying processes.

6.3 Generalized linear models with functional predictors

In this section we present statistical methods that may be used for the analysis of complex data like those described in Paragraph 3.3.2. We are actually working on the implementation of these technique on integrated data arising from the linkage between STEMI Archive and Public Health Database (PHD) of Regione Lombardia. A preliminar example of application that testify for feasibility of such a statistical technique is provided in [15].

In what follows, only statistical methods for generalized linear models with functional covariates (FGLM) are presented [113]. The idea is to apply these models to binary outcomes such as in-hospital, long term mortality or time to the next admission, using, among others, functional predictors like the process of previous hospitalization of each patient.

6.3.1 Model for recurrent events

Let $(\mathscr{F}_t)_{t\in I}$ be a filtration associated to the probability space $(\Omega,\mathscr{F},\mathbb{P})$, with $I=[0,\tau]$. We define the counting process $(N(t))_{t\in I}$ adapted to $(\mathscr{F}_t)_{t\in I}$ as follows:

$$N(t)=\sum_{j=0}^{\infty} I\{S_j\le t, S_j\le\tau\}, \tag{6.5}$$

where S_j represents the calendar time of the j-th occurrence of the observed event and τ represents a random censoring time for the process. N is a submartingale such that, for every stopping time T, $N(T)$ is uniformly integrable, then the Doob-Meyer decomposition theorem states that there exists a unique predictable, non decreasing, cadlag and integrable compensator (or *cumulative hazard*) process $(\Lambda(t))_{t\in I}$ such that

$$M=N-\Lambda \tag{6.6}$$

is a zero-mean, uniformly integrable martingale [6]. Hence the distribution of event times is completely characterized by the knowledge of process Λ, on which modelling efforts should then be

focused. We assume that

$$\Lambda(t) = \int_0^t C(s)\lambda(s)ds, \tag{6.7}$$

where $C(s) = I\{s \leq \tau\}$ is the *at-risk process*, and $(\lambda(s))_{s\in I}$ is called *hazard function*, or *intensity process*.
A wide variety of models for the intensity process can be found in the literature on counting processes, ranging from Poisson processes to the Cox model [27], additive models, frailty and dynamic models (see for instance [1] and [6] for a presentation and discussion of various possibilities). Our choice for the target problem is the following: for $i = 1,\dots,n$, the i-th subject has covariate vector $\boldsymbol{X}_i(t) = (X_{i1}(t),\dots,X_{iq}(t))^T$ (eventually time dependent), and the intensity is

$$\lambda(t|\boldsymbol{X}_i) = \lambda_0(t)\alpha^{N_i(t^-)}e^{\gamma^T\boldsymbol{X}_i(t)}, \tag{6.8}$$

where $\lambda_0(t)$ is an unknown baseline hazard function, α is a real parameter and $\gamma = (\beta_1,\dots,\beta_q)^T$ a q-dimensional vector of real coefficients.

The model assumed in equation (6.8) is a Cox model with the process state $N_i(t)$ as a dynamic covariate, but the proposed methodology, as will be clear in the following, can be applied to a wide range of models for intensity. We choose to account for unobserved heterogeneity by using the dynamic component $\alpha^{N_i(t^-)}$ instead of a frailty variable, i.e., a multiplicative random effect. Dynamic and frailty modelling can be seen as two related methods for describing subject heterogeneity, but the former is more general and flexible (see [1] for a discussion on frailty and dynamic models).The dependence of intensity on process state, here representing the hospitalization process of each patient, is modeled by the term $\alpha^{N_i(t^-)}$ because of its clear interpretation: values of α higher than 1 indicate that a new event implies a worsening of the patient's condition, increasing future rehospitalization risk, vice versa for α values lower than 1. We assume the baseline intensity λ_0 to be dependent on total time t, but more general choices can be made within the same framework; see for example the concept of *effective age* introduced in [118].

Adding a censoring variable to account for different observation times, the model for cumulative hazard can be written as follows, for patients $i = 1,\dots,n$

$$\Lambda_i(t|\boldsymbol{X}_i) = \int_0^t C_i(s)\lambda_0(s)\alpha^{N_i(s^-)}\exp[\gamma^T\boldsymbol{X}_i(s)]ds, \tag{6.9}$$

where $C_i(s) = I\{s \leq \tau_i\}$ (i.e., subjects have different censoring times τ_i, assumed to be mutually independent). Independent censorship as defined in [89] can be reasonably assumed for the considered problem, as we will deepen in the following.

6.3.2 Cumulative hazard smoothing and reconstruction

Semiparametric estimation of cumulative hazard, as proposed in [113], produces a step function estimate $\widehat{\Lambda}_0$ of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ that has the following expression: defining t_j as the j-th observed jump time of the aggregated process $N_\cdot(t) = \sum_{i=1}^n N_i(t)$ and $\tau = \max_{i=1,\dots,n} \tau_i$

$$\widehat{\Lambda}_0(t) = \sum_{t_j\leq t} \frac{1}{\sum_{i=1}^n C_i(t_j)\hat{\alpha}^{N_i(t_j^-)}e^{\hat{\gamma}^T\boldsymbol{X}_i(t_j)}}, \qquad t \in (0,\tau],$$

where $\hat{\alpha}$ and $\hat{\gamma}$ are maximum likelihood estimates of α and γ. Assuming the real Λ_0 function to be absolutely continuous, we deal with the issue of smoothing its estimate $\widehat{\Lambda}_0$, successively moving on to the reconstruction of cumulative hazard process realizations for each patient.

The function $\Lambda_0(t)$ has two a priori characteristics that we want to be preserved by the smoothing procedure: increasing monotonicity and $\Lambda_0(0) = 0$. A fast and efficient way of smoothing functional data while enforcing desired constraints has been proposed in [70]. The method consists in a minimum absolute deviation estimate of coefficients for a B-spline basis expansion: given a set of observations $\{(x_i, y_i)\}_{i=1,\dots,m}$ from a function $y = f(x)$ to be smoothed, a set of knots $\{u_0 = 0, u_1, \dots, u_{k-1}, u_k = \tau\}$ and a fixed polynomial degree d, find $\boldsymbol{a}^* = (a_0^*, \dots, a_{k+d-1}^*)^T$ such that

$$\boldsymbol{a}^* = \underset{\boldsymbol{a}\in\mathbb{R}^{k+d}}{\operatorname{argmin}} \sum_{i=0}^{m} \left| y_i - \sum_{j=0}^{k+d-1} a_j B_j^{(d)}(x_i) \right|, \tag{6.10}$$

$B_0^{(d)}(x), \dots, B_{k+d-1}^{(d)}(x)$ being the B-spline basis of degree d on the chosen set of knots. If basis functions of polynomial degree $d = 1, 2$ are used, then monotonicity, convexity and pointwise constraints can be written as linear constraints. Since the quantity to be minimized can also be written as a linear objective function, the problem can be solved with linear programming techniques, whose efficiency and reliability are ascertained. Using $\{(0, \widehat{\Lambda}_0(0)), (t_1, \widehat{\Lambda}_0(t_1)), (t_2, \widehat{\Lambda}_0(t_2)), \dots\}$ as observations, the application of this method provides the desired smooth estimate $\widetilde{\Lambda}_0$.

We then need to reconstruct the realizations of processes $\Lambda_i(t)$ for every patient $i = 1, \dots, n$, under the chosen model, since in the following cumulative hazard functions are treated as functional data. Given the particular formulation of our model for cumulative hazard, we can rewrite it in a form that allows us to use directly the smoothed estimate $\widetilde{\Lambda}_0$ instead of an estimate of λ_0. For $i = 1, \dots, n$, we set $0 = t_0^{(i)}$ and let $(t_1^{(i)}, \dots, t_{N_i(t)}^{(i)})$ be the jump times for patient i; then

$$\begin{aligned} \Lambda_i(t) &= \int_0^t \lambda_0(s) e^{N_i(s^-)\log\alpha + \gamma^T \mathbf{X}_i(s)} ds \\ &= \sum_{k=0}^{N_i(t)} \int_{t_k^{(i)}}^{t_{k+1}^{(i)}} \lambda_0(s) e^{k\log\alpha + \gamma^T \mathbf{X}_i(s)} ds. \end{aligned} \tag{6.11}$$

Here we consider the case of a covariate vector $\mathbf{X}_i^T = (\mathbf{X}_i^{dT}, \mathbf{X}_i^{cT})$, $i = 1, \dots, n$, where $\boldsymbol{X}_i^d(t) = (X_{i1}(t), \dots, X_{in_d}(t))^T$ is a vector of differentiable functions, while $\mathbf{X}_i^c(t) = (X_{i(n_d+1)}(t), \dots, X_{i(n_c+n_d)}(t))^T$ is a vector of stepwise constant functions with discontinuities corresponding to the jumps of $N_i(t)$; hence we split also the parameter vector γ using $\gamma_d = (\gamma_1, \dots, \gamma_{n_d})^T$ and $\gamma_c = (\gamma_{n_d+1}, \dots, \gamma_{n_d+n_c})^T$, so that $\gamma = (\gamma_d^T, \gamma_c^T)^T$. Defining $P_{\mathbf{X}}(t) = \int_0^t \lambda_0(s) e^{\gamma_d{}^T \mathbf{X}_i^d(s)} ds$ and integrating by parts we obtain

$$P_{\mathbf{X}}(t) = \Lambda_0(t) e^{\gamma_d^T \mathbf{X}_i^d(s)} - \int_0^t \Lambda_0(s) \gamma_d^T [\mathbf{X}_i^d(s)]' e^{\gamma_d^T \mathbf{X}_i^d(s)} ds, \tag{6.12}$$

where $[\mathbf{X}_i^d(s)]' = \left(\frac{dX_{i1}(s)}{ds}, \dots, \frac{dX_{in_d}(s)}{ds} \right)^T$. Plugging $P_{\mathbf{X}}(t)$ into (6.11) leads to the expression

$$\Lambda_i(t) = \sum_{k=0}^{N_i(t)} e^{k\log\alpha + \gamma_c^T \mathbf{X}_i^c(t_k^{(i)})} \left[P_{\mathbf{X}}(t_{k+1}^{(i)}) - P_{\mathbf{X}}(t_k^{(i)}) \right]. \tag{6.13}$$

This form allows us to perform only one integration to obtain (6.12), which is computed substituting $\Lambda_0(t)$ with its smoothed estimate $\widetilde{\Lambda}_0(t)$, and to reconstruct the realizations $\widetilde{\Lambda}_i(t)$ by adding process jumps information.

6.3.3 Functional principal component analysis

We shall now use the reconstructed realizations $\tilde{\Lambda}_i(t)$ as functional covariates in a generalized linear model, to predict outcome. Since these data are high-dimensional, a common strategy is to perform

a suitable dimensional reduction. In the case of functional data, this can be done by expanding them on a functional basis, and choosing only relevant components of the expansion.

Consider a functional ANOVA decomposition of data, as suggested in [119]

$$\widetilde{\Lambda}_i(t) = \mu(t) + D_i(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n \tag{6.14}$$

where $\mu(t) = \mathbb{E}[\widetilde{\Lambda}(t)]$, $D_i(t)$ is the residual for subject i and $\varepsilon_i(t)$ a noise term. One of the possibilities for representing $\widetilde{\Lambda}_i(t)$ is to use Karhunen-Loève decomposition, which states that functional principal components of a set of functions defined on domain T form a complete orthonormal basis of $L^2(T)$ (see [41] for some theoretical results and [126] for details on the implementation of functional principal component analysis, briefly FPCA). At this point we assume that functional data are known on a common support T, thus enabling us to estimate a common Karhunen-Loève basis. Given the covariance operator

$$G(t,s) = \mathbb{E}\left[\left\{\widetilde{\Lambda}(t) - \mathbb{E}\big[\widetilde{\Lambda}(t)\big]\right\}\left\{\widetilde{\Lambda}(s) - \mathbb{E}\big[\widetilde{\Lambda}(s)\big]\right\}\right] \quad \text{for } (t,s) \in I \times I,$$

the eigenvalue problem to be solved in order to obtain principal components is to find the couples $\{(\psi_k, \nu_k)\}_{k\in\mathbb{N}}$, with $\psi_k \in L^2(T)$ and $\nu_k \in \mathbb{R}$, such that

$$\int_T G(t,s)\psi_k(s)ds = \nu_k \psi_k(t). \tag{6.15}$$

Once eigenfunctions $\{\psi_k\}_{k\in\mathbb{N}}$ and eigenvalues $\{\nu_k\}_{k\in\mathbb{N}}$ have been found, we can express the functional ANOVA decomposition (6.14) through the following representation

$$\widetilde{\Lambda}_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n,$$

where $\xi_{ik} = \int_T D_i(s)\psi_k(s)ds$ is the k-th score for subject i.

Eigenfunction-eigenvalue couples $\{(\psi_k, \nu_k)\}_{k\in\mathbb{N}}$ completely explain modes of variation in the data, in the sense that eigenfunctions represent orthonormal directions of decreasing variability with respect to the explained variances expressed by the corresponding eigenvalues. Thanks to the basis expansion given by principal components, it is possible to represent data using just the first K elements of $\{\psi_k\}_{k\in\mathbb{N}}$, the linear combination of which is, by construction, a good approximation for the original curves. The interpretation of eigenvalues as variances is useful also to determine a criterion to choose the most relevant modes. Since $\sum_{k=1}^{K} \nu_k$ represents variance captured by the first K components, we can choose K so that the proportion of variance described by these components is higher than a threshold c, i.e.,

$$\frac{\sum_{k=1}^{K} \nu_k}{\sum_{k=1}^{m} \nu_k} \geq c,$$

where m is the number of abscissa values on which functional data are known, which is an upper bound to the number of components that can be estimated. We then use the following approximation

$$\widetilde{\Lambda}_i^K(t) = \mu(t) + \sum_{k=1}^{K} \xi_{ik}\psi_k(t) + \varepsilon_i(t), \qquad i = 1, \ldots, n.$$

For the sake of notation simplicity, from now on we will write $\widetilde{\Lambda}_i(t)$ even when its truncated basis expansion $\widetilde{\Lambda}_i^K(t)$ is used.

6.3.4 Generalized linear models with functional covariates

Let us consider now a logistic regression model taking into account $\tilde{\Lambda}(t)$ as a functional covariate. In particular, let us call η_i the linear predictor composed by both functional and traitional covariate related to subject i.

$$\begin{aligned}\eta_i &= \int_T D_i(t)\delta(t)dt + \mathbf{x}_i^T\beta \\ &\approx \int_T \delta(t)\sum_{k=1}^{K}\zeta_{ik}\psi_k(t)dt + \mathbf{x}_i^T\beta,\end{aligned}$$

where $\delta : T \mapsto \mathbb{R}$ is a functional parameter, β is a vector of time independent parameters to be estimated and $\mathbf{x}_i$ is a vector of time independent covariates. Notice that we used the K most relevant principal components to represent $D_i(t)$. If $\delta(\cdot)$ is also represented with respect to the principal components basis, i.e., $\delta(t) = \sum_{j=1}^{K}\delta_j\psi_j(t)$, for the orthonormality of $\{\psi_k\}_{k\in\mathbb{N}}$ we obtain

$$\eta_i = \sum_{k=1}^{K}\zeta_{ik}\delta_k + \mathbf{x}_i^T\beta$$

In this formulation the first K FPC scores can be used to summarize the features of hazard functions with a finite dimensional vector, thus providing a powerful methodology to use functional data in many different classical models for multivariate data: the functional estimation problem is reduced to the multivariate estimation of parameter vectors β and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^T$. We are actually working to the generalization of such a framework to mixed-effects models. Temptatives in this direction are suggested in [134] and [135].

Part III

Data Analysis and Applications

In God we trust; all others must bring data.
...
Uncertainty makes research predictable,
but you still need proof to satisfy everyone else.
...
The most important things cannot be measured. The issues that are most important, long term, cannot be measured in advance. However, they might be among the factors that an organization is measuring, just not understood as most important at the time.

Edwards Deming

Keywords: Data Mining; Providers' profiling; League Tables; Model selection; Survival Prediction; Bundle Brunch Block; Functional Boxplots.

Chapter 7

Statistical analysis of STEMI Archive data

In this chapter, the analyses carried out on data arising from STEMI Archive (Paragraph 3.2.3) are presented. Data refer to data collection performed from November 2010 to July 2011. Main goals of the analyses are:

- to depict the epidemiological reality of STEMI characterizing Regione Lombardia cardiological context, through STEMI Archive and its linkage to administrative datawarehouse. This real time data collection, performed simultaneously by all providers of Regione Lombardia using the same technical support, enhances the use of already existing data resources without any further economic effort, improves the quality of data and can be considered an effective method for monitoring and giving feedback to the involved structures and players;
- to monitor hospital performances in terms of both outcomes and process indicators, in order to point out atypical situations to invest into in terms of quality improvement and to support decision on management and rationalization of resources and the Cardiological Network itself;
- to classify healthcare providers according to evidence based criteria and to effect they have on outcomes of interest, using suitable statistical methods and adjusting for different case mix.

Within each of the following sections will be specified the dataset analyzed among those cited in Section 3.2 and the method adopted among those presented in Part II.

7.1 Descriptive analysis of data

Information contained in the dataset we describe here and we will analyze in the following sections concern the fields mentioned in Paragraph 3.2.3, and can be divided in the following areas: Overall description of patients; Way of admission and symptoms; Clinical evaluation at admittance; Reperfusion Therapy; Process indicators; Outcomes. In what follows we will provide an overall descriptive analysis of STEMI Archive, with some final benchmarks of providers' performances concerning the main process indicators, focusing only on those elements that will be of interest for inferential analyses in the following sections. Detailed descriptive analysis of all the contents of the Archive can be found in [85].

Overall description

The dataset consists of 1889 statistical units whose Hospital Discharge Form (*Scheda di Dimissione Ospedaliera*, briefly SDO) has been closed and sent to Regione Lombardia within the time period from November 2010 to July 2011. These patients were eligible to be inserted in the Archive since admitted with STEMI diagnosis in one of the hospital belonging to the Cardiological Network (Section 2.1) of Regione Lombardia. In particular, 35 providers result to have inserted cases within the period of interest. The minimum number of inserted cases is 3, the maximum is 154.

We are working also on tests for assessing if the sample of cases inserted in STEMI Archive is effectively representative of the population treated by each hospital, thanks to the integration with *Ricoveri* database of Regione Lombardia.

Gender

Stratifying the population by gender, we observe 1355 (71.73%) men and 534 (28.27%) women, according to data that can be found in literature concerning STEMI patients.

Age

The overall mean age ($\pm$ standard deviation) is equal to 66.29 ($\pm$13.24). Inspecting Figure 7.1, it can be evinced that women are significantly elder than men (mean age of 73.42 ± 12.87 vs mean age of 63.48 ± 12.30 years), as confirmed by the Wilcoxon nonparametric test (p-value $< 2.2 * 10^{-16}$).

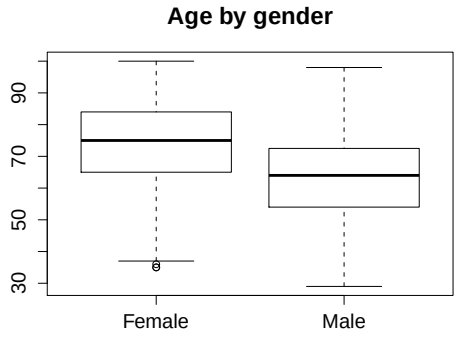


Figure 7.1: Flanked Boxplots of patients' age, stratified by gender.

Moreover, partitioning the overall age in classes as reported in Table 7.1, it can be noticed that more than a quarter of the population is over 75 years.

$(35,50]$	$(50,65]$	$(65,70]$	$(70,75]$	$(75,80]$	over 80
263	636	232	248	195	315
14%	34%	12%	13%	10%	17%

Table 7.1: Stratified overall age.

Admission to hospital and symptoms

Mode of admission

Stratifying patients acording to their mode of admission to hospital (there are 58 missing data in this field), we obtain what is reported in Table 7.2. In particular, 53.12% of the population is managed by 118 rescue units, whereas 46.88% is self-presented. Among patients managed by 118, 61.71% are rescued by Advanced Rescue Units (ARUs), i.e., rescue units with doctors onboard, 35.06% by Basic Rescue Units (BRUs), i.e., the common ambulances, and the remaining 3.23% by rescue units with nurse onboard (IRUs). For all patients, but especially for those managed by 118, it is possible to check that for 44.7% of patients, the Fast-Track have been activated (i.e., patterns of care connecting directly patients with hemodynamic or cath-lab without passing through Emergency Room - ER), in order to monitor and evaluate the efficiency of Network for the providers admitting these patients.

ARU	BRU	IRU	Self-presented
535 (32.78%)	304 (18.62%)	28 (1.72%)	765 (46.88%)

Table 7.2: Stratified way of admission, without patients transferred from one hospital to another one.

Moreover, 199 patients who were transferred from one hospital to another one have been excluded; they will be analyzed apart later, since represent a different population with respect to the one we are interested in, both in terms of treatment type and process indicators.

Symptoms

The stratification of declared symptoms at *call* time (see Paragraph 2.2.2) is reported in Figure 7.2. Thoracic pain is, as expected, the most common declared symptom, followed by epigastric pain and dyspnea, which are also typical of infarction, then more atypical symptoms and the cardiac arrest.

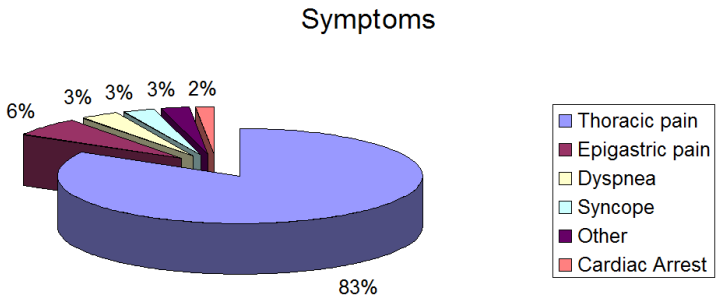


Figure 7.2: Declared symptoms stratification.

Admission/discharge department and drugs

Patients of this collection of STEMI Archive are admitted in one of the following departments: UTIC (the Italian acronym for Intensive Care Coronary Unit, 93.22%), *Cardiologia* (2.01%), *Rianimazione* (3.02%), *Medicina Generale* (1.38%), *Medicina d'urgenza* (0.37%). On the other hand, they all result to be discharged by on of the two following: UTIC (87.80%) and Cardiology

(12.20%). Most of them result treated with almost one of the following drugs: *Aspirina* (83.22%), *Tienopiridine* (15.30%), *Eparina* (60.14%), *Bivaluridina* (7.31%), *Prasugrel* (4.50%), *Clopidogrel* (33.62%). Moreover, 24.14% of them are treated with antiplatelet, 19.45% with beta-blockers, and 3.65% take insulin. These fields are useful to check for the compliance to prescribed protocols of care for STEMI patients.

Evaluation at admittance

Killip

The Killip class quantify the severity of the infarction on a discrete scale from I (less severe infarction) to IV (the most severe infarction). According to this classification, the patients of STEMI Archive are divided as follows: 81% in Killip class I, 11.06% in Killip class II, 3.92% in Killip class III and 4.02% in Killip class IV. The latter class is the one containing the patients presenting with highest risk and then more likely to die, as proved by the number of death within each class, reported in Table 7.3.

	I	II	III	IV
Alive pts.	1503	193	58	37
Dead pts.	27	16	16	39

Table 7.3: Number of dead patients within each Killip class.

Among the 76 patients in Killip class IV, 52 present Cardiogenic shock among Major Adverse Cardiovascular Events (MACE), i.e., complications due to the clinical conditions of patients and to the clinical practice they are treated with.

Risk Factors

Risk Factors can be used to depict the inner case mix of each provider. This information can not be obtained from an administrative database, but only from a clinical registry. The survey of risk factors in STEMI Archive is the following: Diabetes (17.79%), Smoke (39.54%), Hypertension (59.82%), Cholesterol (50.24%), Vasculopathy (10.85%), Chronic Kidney Disease (CKD) (9.11%), previous AMI (12.55%). In summary, only 5.61% of patients has no risk factors, 94.39% present at least one risk factor, 11.69% present more than 4 risk factors. The presence of high number of risk factors could be considered as prognostic of in-hospital mortality, whereas CKD is prognostic of long term mortality, as we will see in the following.

Other clinical feautres of interest

For each patient inserted in the STEMI Archive, we saw in §3.2.3 that several clinical data are recorded. Among these, we report in the following the most interesting for carrying out a preliminary clinical evaluation of the patient.

- ▷ **Systolic Blood Pressure** - The distribution of the Blood Pressure in patients of STEMI Archive has mean equal to 136.9 mmHg and standard deviation equal to 30 mmHg.
- ▷ **Creatinine peak** - The adverse prognostic significance of biomarker elevations (i.e., creatine kinase among others) during AMI has been object of many studies in the past decades. In fact,

peak levels of biomarker have been shown to be reliable predictors of infarct size and prognosis in patients with AMI. In our population, the mean of Creatinine peak is equal to 1.212 mg/dl, with standard deviation equal to 0.816 mg/dl. Creatinine peak and its relationship with the creatinine value at admittance are of great importance for decision on optimal treatment, and are related with consequences on kidneys health. In STEMI Archive, 139 patients have a Creatinine peak greater than 1.5 times the value at admittance. Among these 10 present CKD among MACE.

- ▷ **Site of infarction** - Among others criteria, different type of infarction can be distinguished according to which part of the myocardium wall is interested by the necrosis event. According to the location, the prognosis and the treatment may be different. In particular an anterior infarction is an infarction affecting the anterior surface of the heart, i.e., the portion facing forward just beneath the chest wall. In STEMI Archive, 44.68% of patients are affected by anterior infarction, whereas 55.32% presents it in other sites.
- ▷ **Bundle Branch Blocks and Atrial Fibrillation** - 3.12% of infarctions are classified as Left Bundle Brunch Blocks (LBBB) (see Paragraph 3.2.4) and 6.03% of patients present also Atrial Fibrillation.

Reperfusion therapy

Two are the main categories of treatment that STEMI patients may undergo: primary Percutaneous Coronary Intervention (PCI) and pharmacological treatment, namely Thrombolysis, which can be further divided in pre-hospital (preH) and in-hospital (intraH), as reported in Table 7.4:

Therapy	Primary PCI	preH Thrombolysis	intraH Thrombolysis	No therapy
pts. (%)	77.92%	0.16%	4.98%	16.94%

Table 7.4: Reperfusion therapy of patients of STEMI Archive

As a whole, 83.06% of patients result to undergo a reperfusion therapy. In particular primary PCI is the most common procedure, as suggested by protocols and guidelines whenever times of intervention make it possible.

Thrombolysis and no-treatment

320 patients come out as not treated. For the first time in a clinical registry, with STEMI Archive it is possible to investigate the reasons of missing treatment (see Table 7.5). This could be of great interest for people in charge with healthcare government as well as physicians of each provider, since it enables them to understand if missing treatments are to be ascribed to errors, logistics, delays or other factors. Noticed that this is the first time that such an enquire can be addressed with a clinical registry.

Patients transferred from spokes to hubs

Of the 199 patients transferred from a first to a second hospital, we observed the received treatment (14 thrombolysis, 152 primary PCI and 33 no treatment) and measured times from Door to Door and Door to Balloon (DB) time in the receiving hospital. The boxplots relating to the latter time indexes are reported in Figure 7.3. The median time of Door 1 to Door 2 time is equal to 111 minutes,

Other causes	110 (34.70%)
Delay in diagnosis	105 (33.12%)
Spontaneous ST resolution	47 (14.83%)
Spontaneous Reperfusion.	20 (6.31%)
Comorbidities	15 (4.73%)
Hemodynamic not available	10 (3.15%)
Pt. refuses treatment	7 (2.21%)
Pt. dies before	3 (0.95%)

Table 7.5: Causes of missed treatment.

whereas the median time of DB in the receiving hospital is equal to 31 minutes. This information may induce considerations on treatment of these patients, and in general to the treatment to be given in such a situation, in order to get the best reperfusion.

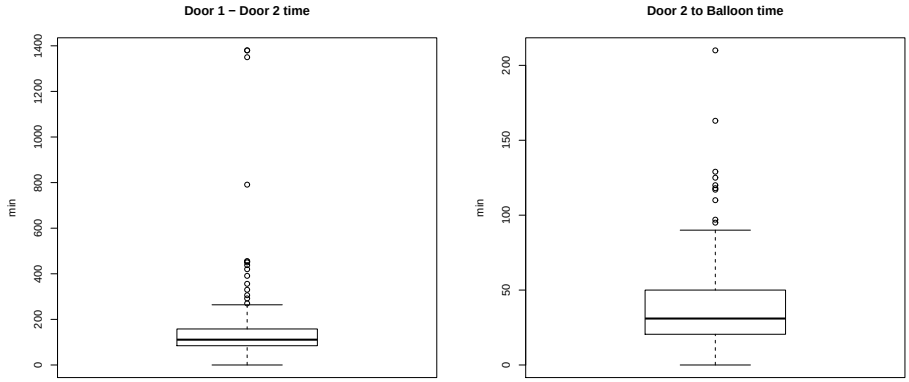


Figure 7.3: Boxplots of distributions of times from admittance to the transferring hospital to admittance to the receiving hospital (left panel) and from admittance to the receiving hospital to Balloon (right panel), for the 199 transferred patients.

Process Indicators

According to times of call, treatment and intervention as defined in Paragraph 2.2.2 and in [36], it is possible to compute process indicators for each provider, in order to evaluate and monitor their efficiency in activating and managing suitable patterns of care for STEMI patients. In the following analysis, we focus on the subpopulation of patients undergone primary PCI and not transferred from one hospital to another one, i.e., 1286 of the original 1889 units.

In Table 7.6, a summary of the main process indicators is reported. Median time of each distribution are presented, together with guidelines (if any) and proportion of patients treated according to guidelines.

Index name		Median	Guidelines (if any)	% of pts. treated according to guidelines
Onset to Balloon	(OB)	193 min		
Onset to Door	(OD)	97 min	180 min	81.81%
Door to Balloon	(DB)	72 min	90 min	64.76%
Onset to First Contact	(OFC)	75 min		
First Contact to Balloon	(FCB)	90 min		
Onset to fisrt ECG	(OfECG)	85 min		
first ECG to Balloon	(EB)	78 min	80 min	50.64%
ECG to Needle	(EN)	26 min	30 min	52.50%

Table 7.6: Summary indexes concerning distributions of the main process indicators.

Onset to Balloon (OB)

The total ischaemic time is a global measure of efficiency, which jointly accounts for delays due to patients (call delay), to 118 service (pre-hospital delay) and to providers logistics and STEMI care protocols (in-hospital delay). It provides a snapshot of the whole process timing, measuring the fluency of the STEMI care pattern from symptoms onset to care delivery.

The distribution of Onset to Balloon (OB) time has a median equal to 193 minutes and mean equal to 273 minutes (patients whose OB time is greater than 24 hours have been excluded by the analysis, since probably these times represent an error in inputing data or anyway, for them, primary PCI should not have been performed, according to guidelines that suggest PCI to be effective within at most 6 hours from symptoms onset). It can be observed that 75% of patients has OB less than 310 minutes, and 90% of them has OB less than 572 minutes. The boxplot and the histogram of the distribution are reported in Figure 7.4. The red lines indicate the threshold of 6 hours. 12 missing data are present.

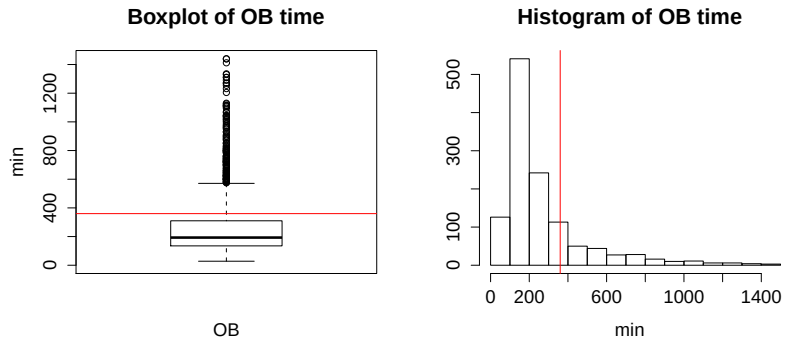


Figure 7.4: Boxplot and histogram of OB time for patients with OB less than 24 hours.

Onset to Door (OD)

The distribution of Onset to Door (OD) time measures the delay to treatment due to pre-hospital events. Unfortunately, it does not allow for distinguishing if the delay is due to the patient fault (delay in calling 118 or in presenting at ER) or to the rescue system (delay in rescue units arrival). This can be done through the joint use of this indicator with the one measuring the time between

symptoms onset and time of first contact with healthcare system (OFC).

The distribution of OD time has median equal to 97 minutes and mean equal to 116 minutes. 75% of patients has OD less than 153 minutes, 90% less than 229 minutes. Patients with OD greater than 6 hours (185) have not been considered for the previously mentioned reasons. Figure 7.5 shows the boxplot and the histogram of OD distribution for patients whose OD is less than 6 hours. The red lines indicate the gold standard threshold of 3 hours. 23 missing data are present.

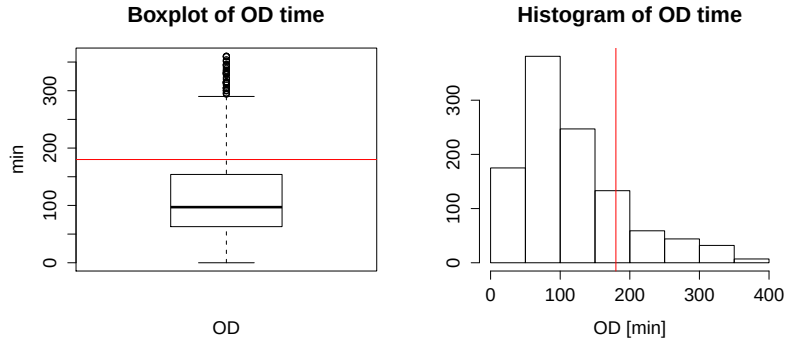


Figure 7.5: Boxplot and histogram of OD time for patients with OD less than 6 hours.

Door to Balloon (DB)

The Door to Balloon is the time interval starting with the patient's arrival in the emergency department and ending when a catheter guidewire crosses the culprit lesion in the cardiac cath lab. It is an index of internal organizational efficiency of hospitals, and it is one of the most accepted process indicator to be monitored in literature (see [156], [158], [189] and [197] among others).

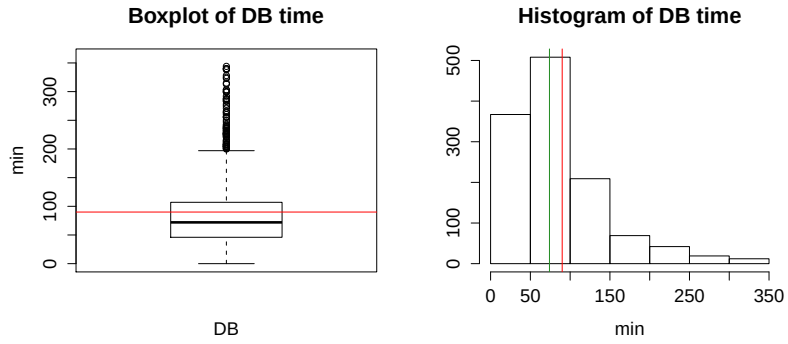


Figure 7.6: Boxplot and histogram of DB time for patients with DB less than 6 hours.

The distribution of DB time, evaluated for all patients whose DB is less than 6 hours, has median equal to 72 minutes and mean equal to 86 minutes. 75% of patients has DB time less than 107 minutes, 90% of them less than 163 minutes. On the whole, 64.76% of patients has DB time less than prescribed 90 minutes. Figure 7.6 shows the boxplot and the histogram of DB distribution for patients whose DB is less than 6 hours. The red lines indicate the gold standard threshold of 90 minutes. 14 missing data are present.

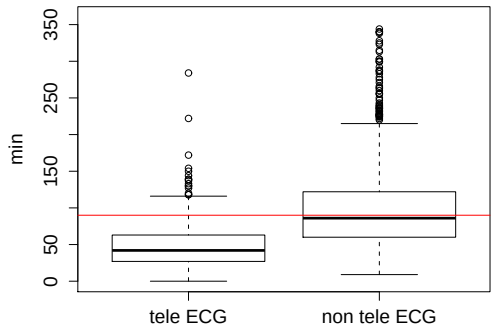


Figure 7.7: Flanked boxplots of DB time stratified according to ECG tele-transmission.

In Figure 7.7 the distributions of DB time for patients with and without tele-transmission of pre-hospital ECG from the rescue units which delivered them to the hospitals are compared. It can be immediately appreciated that the first is stochastically lower than the second. This is confirmed by Wilcoxon nonparametric comparison test (p-value $< 2 * 10^{-16}$). In the subpopulation with tele-transmitted ECG, 88.76% of patients has DB time less than 90 minutes, whereas in the subpopulation without tele-transmitted ECG, only 53.31% of patients results to be treated according to guidelines.

Onset to First Contact (OFC)

The time from symptoms onset to the First Contact with the National Health Service is the Door time for self-presented people, whereas it is the arrival time of the rescue units for people delivered by 118. The distribution of OFC time for patients whose OFC is less than 6 hours (161 patients are then excluded), has median equal to 75 minutes and mean equal to 101 minutes. 75% of patients has OFC less than 130 minutes, and 90% less than 224 minutes. 13 missing data are present.

Onset to First ECG (OfECG)

The first ECG time is defined as the time of ECG tele-transmission by the rescue units (if any) for patients delivered by 118, and as the time of first ECG at ER for all the others. The time from symptoms onset to first ECG (OfECG) enable us to monitor how fast the diagnosis is carried out. The distribution of OfECG time, excluding patients whose OfECG is greater than 6 hours (33) or negative (15), has median equal to 85 minutes and mean equal to 108 minutes. 75% of patients has OfECG less than 145 minutes, and 90% less than 224 minutes.

First ECG to Balloon (EB)

The time from first ECG to primary PCI (EB) enables us to monitor how fast the therapy is executed for those patients who received a STEMI diagnosis through ECG.
The distribution of EB time, excluding patients with EB time greater than 6 hours or negative (9), has median equal to 78 minutes and mean equal to 90.65 minutes. 75% of patients has EB time less than 109 minutes, and 90% less than 151 minutes. Figure 7.8 shows the boxplot and the histogram

of EB distribution for patients whose EB is less than 6 hours. The red lines indicate the gold standard threshold of 80 minutes. 69 missing data are present.

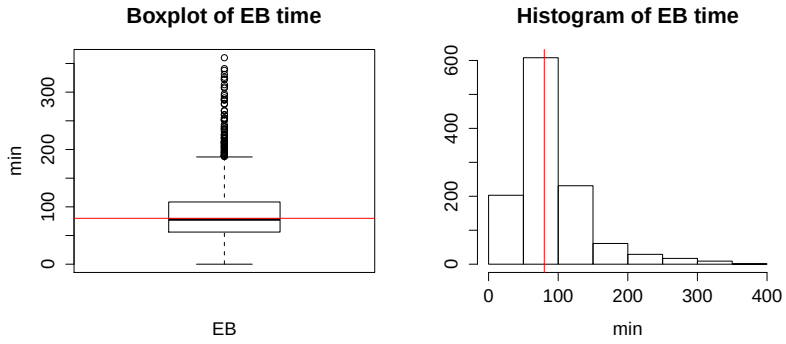


Figure 7.8: Boxplot and histogram of EB time for patients with EB less than 6 hours.

Overall benchmarks of process indicators

In this paragraph, we show some benchmarks of the most important process indicators, relating the behaviour of each provider (all the information are available to us in anonymous form) with respect to the overall behaviour of the network. These benchmarks are a simple and effective instrument that people in charge with healthcare governance could be take advantage of in describing the real condition of the network in term of efficacy and efficiency.

DB vs Exposure

In Figure 7.9 the proportion of patients with DB less than 90 minutes against Exposure, i.e., the proportion of cases inserted in the STEMI Archive by each hospital over the total of 1889 units. The coloured lines show respectively: the threshold of 25 cases (yellow vertical line), the threshold of 50% of patients treated according to guidelines (red horizontal line) and the threshold of 75% of patients treated according to guidelines (green horizontal line). The best-operating hospitals are those being in to right-up corner. It seems to be no evidence for the hypothesis “the greater number of patients treated, the better the performances achieved”.

DB stratified by hospital

Figure 7.10 shows the boxplots of DB distribution for each hospital of the STEMI Archive (only patients with DB less than 6 hours are considered). The red line indicates the gold standard threshold of 90 minutes. From this graph can it be better understood the good result already indicated by the median of the overall DB time. In fact, most of providers maintain medians and sometimes even 3rd quartile under the guidelines. This is probably the effect of the last years campaigns, carried out especially on Milan urban area, for improving in-hospital organizations.

EB vs first ECG

In Figure 7.11 is reported, for each hospital of STEMI Archive, the median EB time against the median of the time of first ECG. The circle dimension is proportional to the exposure of each structure. The best performing hospitals are those in the lower-left corner. Again it seems to be no

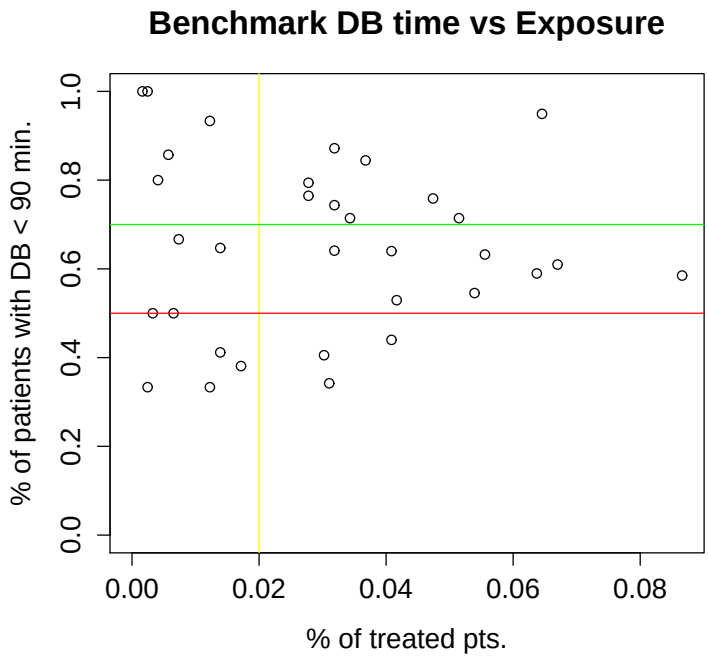


Figure 7.9: Benchmark of DB time against Expoure for each hospital of STEMI Archive. The yellow line is the threshold of 25 cases inserted, the red one is the threshold of 50% of patients treated according to guidelines, and the green line the threshold of 50% of patients treated according to guidelines.

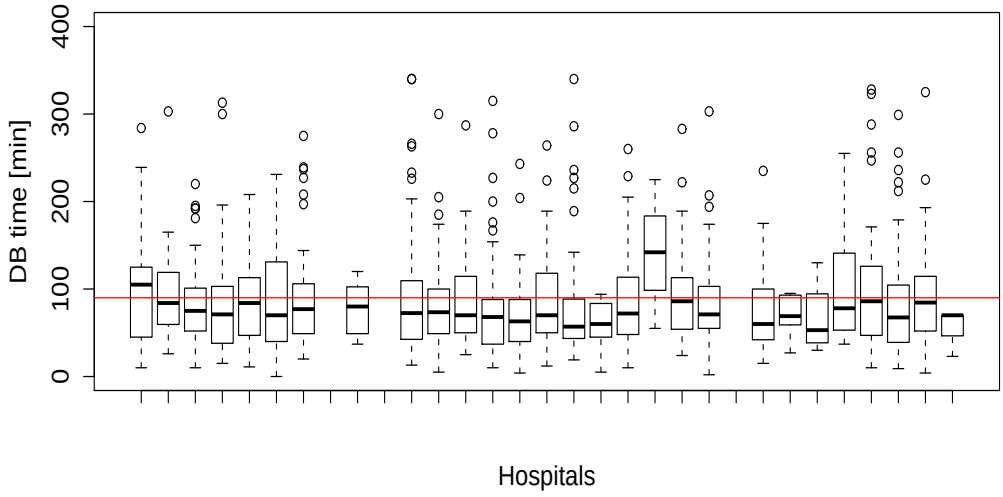


Figure 7.10: Flanked Boxplots of DB time for each hospital of STEMI Archive (only for patients whose DB time is less than 6 hours).

evidence for correlating the number of treated patients with good performances. The green square is the global median of all hospitals. Coloured lines highlight the guidelines of 80 and 10 minutes respectively.

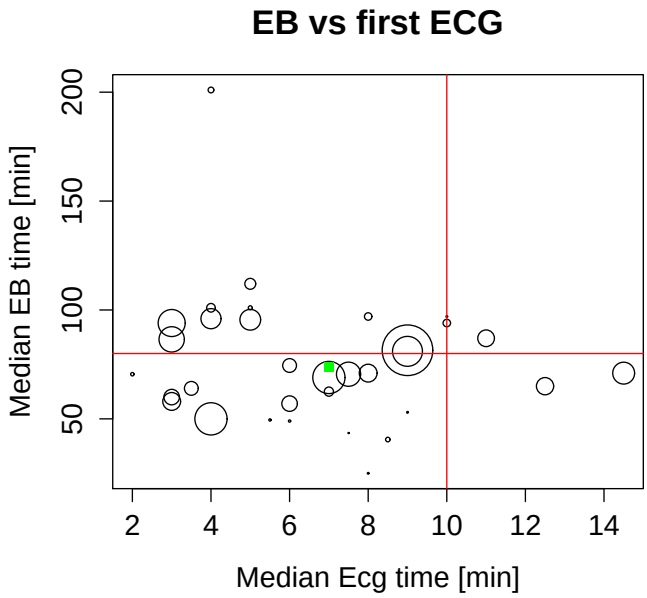


Figure 7.11: Benchmark of median DB times against median times of first ECG for each hospital of STEMI Archive. The diameter of each circle is proportional to the number of cases inserted in the registry by each hospital. Red lines indicate the thresholds of acceptability (80 minutes and 10 minutes respectively) suggested by international guidelines [151]. The green square shows the overall median of all hospitals.

Outcomes

Major Adverse Cardiovascular Events (MACEs)

As we said before, MACEs are adverse events arising after treatment and/or induced by the critical profile of patients, which have a great influence not only on the mortality outcome, but also on the quality of life of the patient. Consequently, cost-effectiveness considerations could be facilitate by the knowledge of their presence. In particular in STEMI Archive 44.79% of patients present at least one MACE. In particular, there are 157 (8.31%) patients presenting Shock, 23 (1.22%) with Re AMI, 37 (1.96%) with Mechanical complications, 190 (10.06%) with Mitral insufficiency, 111 (5.88%) with Pulmonary Oedema, 63 (3.34%) with Major Bleedings, 29 (1.54%) with Ischemia, 628 (33.24%) with Arrhythmias.

In-hospital Mortality

The in-hospital mortality of STEMI Archive patients is equal to 5.19% (98 deaths). Concerning death patients, 63.27% of them have been treated with primary PCI, 6.12% with intraH thrombolysis, and 30.61% of them received no treatment. Deaths are distributed in each hospital as reported in Figure 7.12. Even at first sight, the great variability among structures is evident, then it is reasonable that this grouping factor may induce an overdispersion effect on the mortality outcome.

Long term survival

The linkage between STEMI Archive and the administrative database of *Anagrafica*, enable us to observe for each patient of STEMI Archive, the long term survival, i.e., the censored data on life

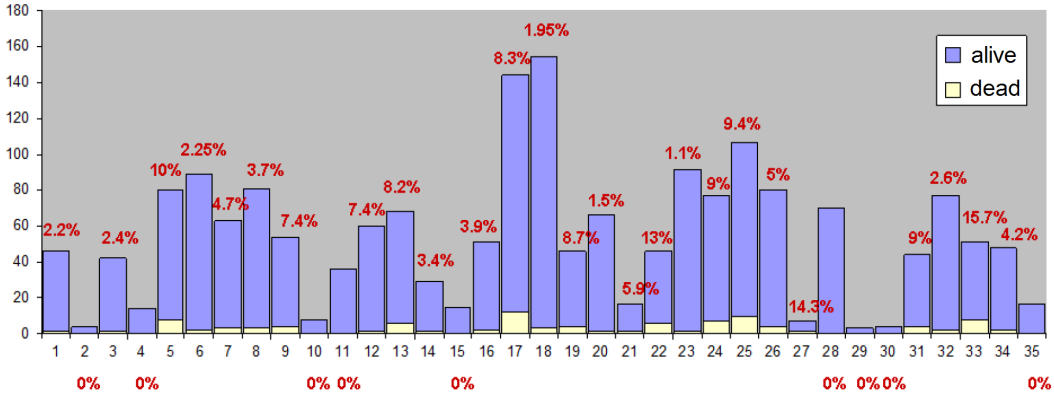


Figure 7.12: Number of patients and in-hospital mortality (%) for each hospital of STEMI Archive.

status of patients at the time of linkage between STEMI Archive and administrative databases. In *Anagrafica* database we observe the mortality due not only to cardiovascular causes, anyway within a month (or few months) from discharge, it is likely that the death is connected with the cardiovascular event happened not far ago. In particular in our case, having performed the integration at the end of September 2011, it is possible to point out for all patients mortality after 30 and 60 days from discharge. Moreover, according to each date of discharge, we have long term mortality up to September 30th, 2011. In what follows we will refer to the analysis of mortality after 60 days as *long term* mortality. For patients of STEMI Archive, the long term mortality comes out to be equal to 7.9%.

ST-resolution after 60 minutes

Analysis of ST-segment resolution on ECG, after any treatment for ST elevation Myocardial Infarction, offers an attractive and cost effective solution to assess coronary reperfusion, since it measures the degree of microvascular reperfusion, which is strongly correlated with the outcome. ST segment is therefore a good indicator of prognosis. We have this information for patients undergone to primary PCI (1436 units). Among these, 76.1% have a successful reperfusion, in term of ST-segment resolution of 70% after 60 minutes from PCI.

As we will see later, ST-segment can be thought both as outcome itself (models will be provided for it) and as prognostic factor for in-hospital and long term mortality, as known also from literature (see [162] and [203]). It is in fact strongly correlated with in-hospital mortality outcome (p-value of independance Chi-squared test $= 1.437 * 10^{-11}$), as well as influenced by pre-hospital and in-hospital times.

Integrated databases: an overview of complexity

Integrating the STEMI Archive with the PHD of Regione Lombardia, a complex longitudinal data can be obtained for each patient. We saw in Paragraph 3.3.2 the main features of such data both in terms of complexity and potential interest for a wide range of analyses. Here we describe some preliminary information arising from integrated data.

Number of admissions to hospitals

The first and most simple information arising from integrated data is the longitudinal overview of patients' clinical history. For example, with respect to the index event of the STEMI Archive, we can assess how many (if any) previous admissions each patient has as well as how many (if any) re admission he/she experiences in the 60 days after discharge.

In particular, we are actually working on previous admissions (and all the features related to them) with models like those described in Section 6.3, as well as focusing on the 772 patients who have re-admissions after STEMI Archive event.

Further analyses on integrated data

The integration between STEMI Archive and administrative databases of *Farmaci* enables us to address further epidemiological enquires like the identification of subpopulation of patients "already known" to the therapeutic iter, and allows for checking the compliance to the prescribed therapy after STEMI Archive event of infarction. In general, we verified also the compliance of each hospital concerning the number of cases inserted in the STEMI Archive. Since it is known to Regione Lombardia how many SDO are dued during data collection periods, we observed that, on average, 70% of cases have been inserted by structures involved in data collection. Anyway, since the percentage of inserted cases ranges from 12% to 98%, we are actually checking if samples of cases inserted in STEMI Archive by eaxh hospital are really representative of the population of STEMI patients they treat.

Results of analyses on STEMI Archive concerning these last two paragraphs are actually under inspection of Regione Lombardia healthcare district, and then cannot be reported here, yet. See [14], [15] and [77] for an example of application of such statistical techniques to a similar databases.

Quality control

Finally, the integration between STEMI Archive and PHD of Regione Lombardia allows for checking of data quality, in terms of completeness, accuracy and reliability. No more self-referenced data are admitted, since all information can be compared with those contained in the administrative database (whose data collection is continuous and compulsory, since it is used by Regione Lombardia to refund hospitals for services delivered). For example, we checked for the correspondence of admission date of STEMI Archive. For 1743 of 1889 patients, it has been possible to achieve the match. Some of the missing matches differed by one day, some were completely missing.

7.2 Frequentist approach to outcomes modelling

In this section, we present the statistical models fitted for the main outcomes of interest of STEMI Archive, i.e., in-hospital mortality (Paragraph 7.2.1), Long term mortality (Paragraph 7.2.2) and MACE (Paragraph 7.2.3). We will focus on subpopulation of STEMI patients undergone primary PCI and not transferred, i.e., on 1286 of the original 1889 statistical units. The main aims of the modelling effort are firstly to verify that literature assessments about factors to be considered as prognostic for outcomes of interest hold also in our dataset, then to include all the significant factors useful to make predictions at patient's level. The first step of the analysis is then the selection of the most significative features to be considered, both from clinical and statistical point of view. Then the selected models are fitted to STEMI Archive data, providing estimates for parameters of interest.

Plots of the predictive surfaces for some benchmark cases are presented in order to quantifying the gain/loss in responses for each setting of interest.

7.2.1 In-hospital survival

Concerning the in-hospital survival (measured through the binary variable *Survival*, assuming value 1 if patient is discharged alive, 0 otherwise), we observe in the subpopulation of patients undergone primary PCI and not transferred a share of success (96.11%) even higher than the in-hospital survival of the overall population. In this case, according to clinical best practice criteria, we considered the following covariates as eligible to be inserted in the model:

- ▷ *Age*: a continuous variable indicating the age of each patient at hospital admission;
- ▷ *Sex*: a categorical variable indicating the sex of each patient;
- ▷ *Killip*: a binary variable for categorized Killip class, assuming value 0 for the less severe class of infarction (Killip I) and 1 for the most severe ones (Killip II, III and IV);
- ▷ *FE*: a continuous variable indicating the percentage of ejection fraction of each patient at admittance;
- ▷ *Risk*: a binary variable indicating the presence ($Risk = 1$) of almost 4 risk factors among those registered in STEMI Archive (see Section 7.1);
- ▷ *STresolution*: a binary variable indicating if an effecacy reperfusion has been reached (almost 70% of ST elevation reduction after 60 minutes from intervention);
- ▷ *Mezzo*: a binary variable indicating the type of rescuing (1 if 118 delivering is observed, 0 if the patient is self-presented);
- ▷ *logOB*: a continuous variable indicating the total ischaemic time (in logarithmic scale) for each patient.

We do not include *Sex* in the model, since it is strongly correlated with (and then masked by) the patient's age (p-value of Pearson Correlation test $< 2.2 * 10^{-16}$); in fact, as we saw in Section 7.1, women are much elder than men. Also *Mezzo* is not included, since it is confounded by *Killip* (p-value of Fisher test for independence $= 2.363 * 10^{-06}$), in the sense that patients in worse conditions are often delivered by 118 rescue units. Finally, the total ischaemic time results to be not significant in explaining mortality, but anyway it is correlated with ST resolution (p-value of Wilcoxon test for comparison $= 0.002123$). Nonlinear growth like those presented in Section 4.5 or Section 4.6 for modelling the behaviour of DB time as function of suitable process indicators are actually under analysis. Figure 7.13 shows the distribution of OB time in logarithmic scale for subpopulations of patients with negative (left) and positive (right) ST resolution respectively.

After these considerations and the stepwise variables selection based on AIC index for variable selection, we fit a GLM model for a Bernoulli binary outcome with canonical logistic link function, obtaining the following output:

```
Call:
glm(formula = Survival ~ Age + Killip + FE + Risk + STresolution,
family = binomial())
```

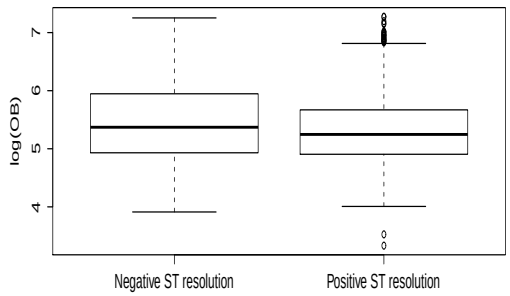


Figure 7.13: Flanked boxplots of OB time in logarithmic scale for patients with positive (left) and negative (right) outcome of ST resolution.

```
Coefficients:
               Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    4.72247    1.73793    2.717   0.006582 **
Age           -0.04931    0.01760   -2.801   0.005096 **
Killip        -1.74335    0.46070   -3.784   0.000154 ***
FE             0.09987    0.02145    4.657   3.21e-06 ***
Risk          -1.73056    0.72437   -2.389   0.016892 *
STresolution   0.97143    0.42839    2.268   0.023351 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 ‘ ‘ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 272.90  on 945  degrees of freedom
Residual deviance: 171.89  on 940  degrees of freedom
AIC: 183.89

Number of Fisher Scoring iterations: 8
```

The model for in-hospital survival of patient *i* is then

$$\begin{aligned} logit(\mathbb{E}[Survival = 1|Age, Killip, FE, Risk, STresolution]) = \\ 4.72247 - 0.04931 \cdot Age_i - 1.74335 \cdot Killip_i + 0.09987 \cdot FE_i \\ - 1.73056 \cdot Risk_i + 0.97143 \cdot STresolution_i \end{aligned} \tag{7.1}$$

As expected, for increasing age, risk and Killip class, the survival probability decreases. On the other hand, the higher the ejection fraction and the reperfusion efficacy, the better the survival. Figure 7.14 shows the estimated survival surfaces (as functions of age and ejection fraction) for 8 different scenarios: from left to right and from top to bottom, we pass from the best case scenario to the worst case one.

Respectively:

(a) Less severe infarction (i.e., Killip class I), low risk (i.e., less than 4 risk factors are present),

reperfusion accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has been achieved);

(b) Less severe infarction, high risk (i.e., almost 4 risk factors are present), reperfusion accomplished;

(c) More severe infarction (i.e., Killip class II, III or IV), low risk, reperfusion accomplished;

(d) More severe infarction, high risk, reperfusion accomplished;

(e) Less severe infarction, low risk, reperfusion not accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has not been achieved);

(f) Less severe infarction, high risk, reperfusion not accomplished;

(g) More severe infarction, low risk, reperfusion not accomplished;

(h) More severe infarction, high risk, reperfusion not accomplished;

The green points on each survival surface mark the estimated survival probability (resumed in Table 7.7) for a reference patient, aged 75 and with 50% of ejection fraction at admittance. This is a simple way of quantifying the impact of each component of the model on outcome. The different shape of each surface is speaking about different case-mix.

Scenario	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
Killip	0	0	1	1	0	0	1	1
Risk	0	1	0	1	0	1	0	1
STresolution	0	0	0	0	1	1	1	1
Survival prob.	99.47%	97.11%	97.07%	85.47%	98.62%	92.71%	92.62%	69.01%

Table 7.7: Estimated survival probability of a reference patient ($Age = 75$ years, $FE = 50\%$ in different case-mix scenarios.

The estimated survival surfaces and related probabilities reported in Table 7.7 are computed not taking into account the grouped nature of data. It is reasonable that they change from hospital to hospital; in fact, the overall mean (std. dev.) of survival probability estimated on GLM model fitted values is 0.9672 (0.0884), whereas the same qauntity computed hospital by hospital ranges from 0.8789 to 0.9955. In other words, there may be latent factors we are actually not accounting for with a simple GLM model, and it is possible that these latent factors influence the estimation procedure. So the next step is not only to adjust our estimates for case-mix, but also to estimate the influence of grouping factor on outcomes. This can be accomplished adopting a mixed-effects approach, which indeed not only enables us to account for the overdispersion induced on data by the presence of a grouping factor (the hospital of admission), but also to quantify the "provider effect" on survival. Moreover, we will take advantage of random effects estimates for clustering together providers characterized by similar influences.

7.2.2 Long term survival

Following the same standards adopted for the analysis of in-hospital survival, in this paragraph we present the model pointed out for long term survival in the subpopulation of patients undergone primary PCI and not transferred. Also in this case, the outcome is represented by the binary variable *Survival*, assuming value 1 if patient is alive after 60 days from discharge, 0 otherwise (we observe

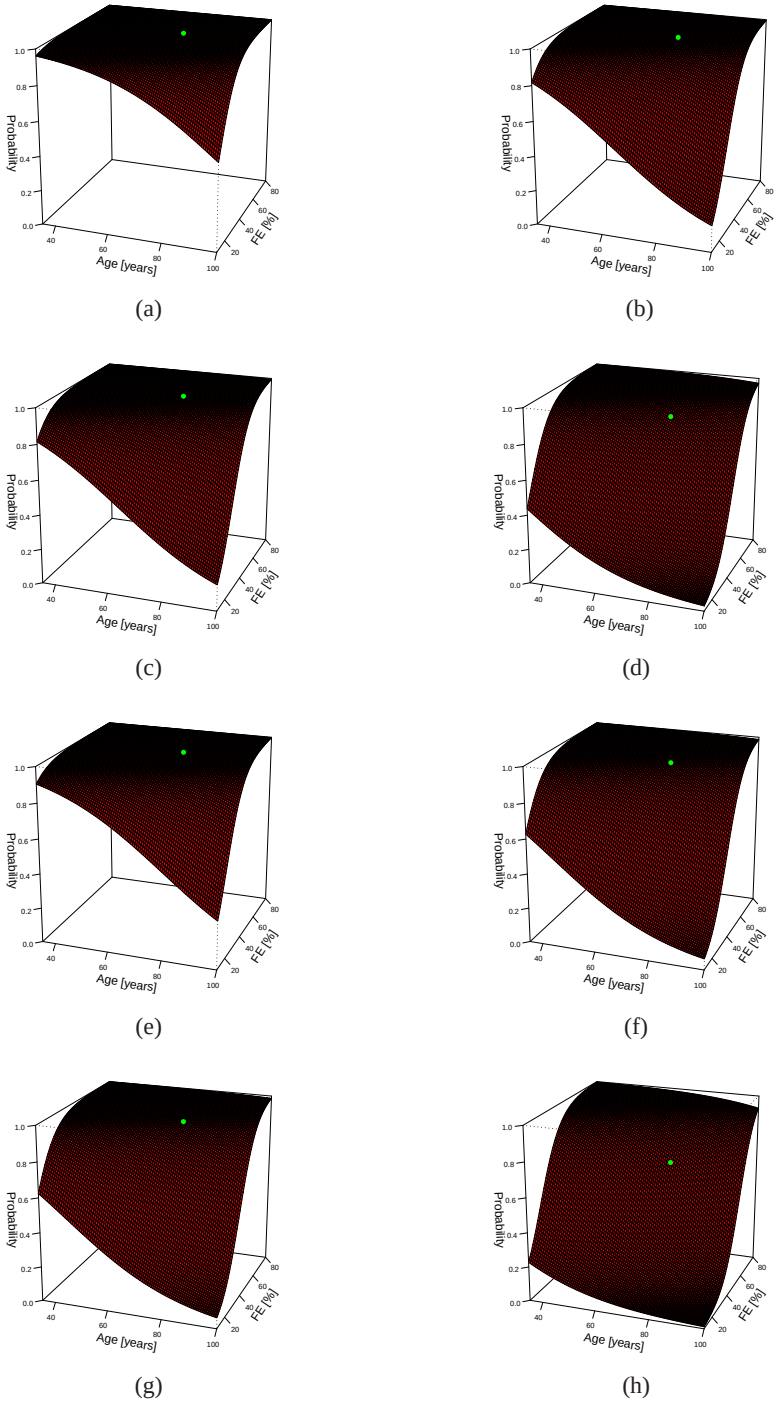


Figure 7.14: Estimated survival surfaces in different case-mix scenarios, obtained fitting a GLM model for survival outcome. Green points indicate the survival probability for a patient aged 75 and with 50% of ejection fraction at the entrance.

a share of success equal to 93%). We recall that it is the first time that it is possible to achieve this information, together with patients' clinical follow up, and that it is due to the integration of STEMI Archive with administrative database *Anagrafica*, where births and deaths of each citizen of

Regione Lombardia are registered. As we said in Paragraph 3.3.2, this is the overall mortality, i.e., not only the mortaliy due to cardiovascular events. Anyway, up to 60 days from a cardiovascular event, the death it is likely to be connected to hearth failure.

After the stepwise variable selection based on AIC index for variable selection, again we fit a GLM model for a Bernoulli binary outcome with canonical logistic lin function, obtaining the following output:

```
Call:
glm(formula = Survival ~ Age + Killip + FE + CKD + STresolution,
family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2811   0.1219   0.1954   0.3138   1.3245

Coefficients:
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     2.17916    0.96190    2.265    0.0235  *
Age            -0.05226    0.01115   -4.687   2.77e-06 ***
Killip         -1.19473    0.27873   -4.286   1.82e-05 ***
FE              0.08712    0.01342    6.492   8.45e-11 ***
CKD            -0.66008    0.37081   -1.780    0.0751  .
STresolution    0.34597    0.13787    2.509    0.0121  *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 ‘ ‘ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 587.84  on 1162  degrees of freedom
Residual deviance: 417.31  on 1157  degrees of freedom
  (121 observations deleted due to missingness)
AIC: 429.31

Number of Fisher Scoring iterations: 6
```

where CKD stays for Chronic Kidney Disease and assumes value 1 if the *i*-th patient is affected by CKD, 0 otherwise. As it has been told us by physicians, it is relevant that such factor is included in the model, since often the drugs given for primary PCI treatment are difficult to be drained by kidneys, then if a patient undergoing primary angioplasty is also affected by chronic kidney disease, the choice of drugs and treatments could strongly affect his/her quality of life (and then his/her long term survival) after discharge. The model for long term survival of patient *i* is then

$$\begin{aligned} logit(\mathbb{E}[Survival = 1|Age, Killip, FE, CKD, STresolution]) = \\ 2.17916 - 0.05226 \cdot Age_i - 1.19473 \cdot Killip_i + 0.08712 \cdot FE_i \\ - 0.66008 \cdot CKD_i + 0.34597 \cdot STresolution_i \end{aligned} \tag{7.2}$$

As expected, for increasing age, CKD and Killip class, the long term survival probability decreases. On the other hand, the higher the ejection fraction and the reperfusion efficacy, the better the long term survival.

7.2.3 MACE

In this paragraph we present the analysis pointed out for Major Adverse Cardiovascular Events (MACE) in the subpopulation of patients undergone primary PCI and not transferred. The outcome is represented by the binary variable *MACE*, assuming value 1 if patient present at least 1 Major Adverse Cardiovascular Events among re-AMI, major bleedings, ischaemy and death, and 0 otherwise (we observe a share of success equal to 20.2%). The idea is to consider a less unbalanced outcome than survival one. At the same time, we would like to catch a more general index of quality of life after STEMI event, i.e., to consider as good prognosis not only the absence of death, but also the absence of events that worsen the patient's health.

Of course, such an outcome depends on covariates and previous clinical history of patient in a complex way. A preliminar stepwise variable selection based on AIC criterion detected only a strong correlation between *MACE* and ejection fraction (p-value of Wilcoxon test $= 6.22 * 10^{-7}$), as shown in Figure 7.15. Anyway, several confounding are likely to be present. That's why we are actually trying to consider a wider set of covariates, led by clinical best practice. Further modelling analyses connecting MACE with patients previous clinical history are actually under inspection of physicians and Regione Lombardia healthcare district.

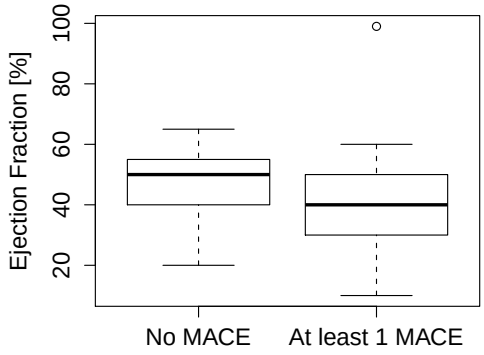


Figure 7.15: Flanked boxplots of FE for patients presenting no MACE (left) and at least one MACE (right).

7.3 Frequentist approach to hospitals clustering

Models adopted in the previous section are now enriched in order to focus on clustering healthcare providers with similar effects on patients survival. Mixed effect models are suitable candidate to enable us to reach such a target. In what follows, we will focus on the in-hospital survival outcome alone, but the approach can be straightforwardly applied to any of the other outcomes of interest.

Performance indicators for assessing quality in healthcare contexts have drawn more and more attention over the last few years because they enable the research workers to measure several components of the healthcare process, clinical outcomes and disease incidence. At the same time, questions about the right use of such indicators as a measure of care quality have emerged. Here we propose the use of performance indicators in modelling the outcomes of clinical structures in order to identify "similar behaviours" among clinical structures. These models include variability between institutions (not forgetting case-mix) and performance indicators are computed starting from

data collected through clinical registries. The purpose of this section is, in fact, to highlight how advanced statistical methods can be used to identify suitable models for complex data coming from clinical registries in order to classify and evaluate health-care providers. In clinical literature (see [160], [177] and [200] among others), several examples make use of clinical registries to evaluate performances of medical institutions.

We want to capture the real standards of performances of the Cardiological Network of Regione Lombardia, a very heterogeneous area in terms healthcare offer. So we identify an effective and robust statistical technique to find similar behaviours or clusters among clinical structures. In general, procedures for analyzing and comparing the effects of the healthcare providers on health services delivery and outcomes are known as *provider profiling*. In a typical profiling procedure, patient-level responses are measured by clusters of patients treated by different providers.

So three different methodologies to evaluate hospital's performance in this provider profiling perspective are proposed. In the first one, we estimate the in-hospital survival rates after fitting a GLM on outcome of interest and then we use the estimated survival probability for computing a score called Stadewide Survival Rate (SSR), relating the actual survival at the $j-$hospital to the expected survival in the same hospital, adjusted for different patient severity resumed in the covariates of the GLM. In the second one, we fit a GLME model to explain in-hospital survival outcome, with a parametric random effect due to the hospital grouping factor, then we perform an explorative classification and ranking analysis on the point estimates of hospital effects. In the third one, we classify the hospitals on the basis of the variance components analysis explained by a GLME model on outcome with a nonparametric random effect. We then use all the three methods to discriminate between different behaviours: we compare classification structures obtained starting from these models and quantify the effect of making part of different groups on outcomes of interest. Results of similar analyses, tested on MOMI^2 dataset, can be found in [62].

7.3.1 Stadewide Survival Rate (SSR)

Starting from the GLM model in (7.1), we compute the Statewide Survival Rate (SSR) for hospital j, defined as

$$SSR_j = \frac{\sum_{i=1}^{n_j} y_{ij}^{obs}}{\sum_{i=1}^{n_j} \hat{p}_{ij}},$$

where y_{ij}^{obs} is the resulting outcome for patient i treated in the hospital j, and $\hat{p}_{ij}$ is the corresponding survival probability estimated by using the GLM in (7.1). An elementary assessment of hospital j can be obtained by comparing SSR_j with 1. Once computed the SSR_j for every hospital, we are able to separate the hospitals in two groups, namely "A" and "B", according to the comparison of *SSR* with the threshold of 1: if greater, then hospital belongs to the first group, if less to the second. According to this criterion, we obtain the classification reported in Table 7.8.

h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11
A	A	A	A	A	B	B	B	B	A	A
h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22
A	A	A	A	B	A	B	A	A	B	A
h23	h24	h35	h26	h27	h28	h29	h30	h31	h32	h33
B	A	B	B	A	A	A	B	A	A	A

Table 7.8: Providers' clustering according to SSR criterion.

7.3.2 Analysis of random effect estimates of a parametric GLME model

We now consider a GLME model like the one presented in Pargraph 4.3.1 for the survival outcome, considering the i index as associated to statistical units grouped in-hospital j. In this case we consider univariate Gaussian random effect additive on the intercept of the GLME model. We denote by σ^2 the related variance. Estimates for fixed effects coefficients (β) and standard deviation of Normal random effect (σ) can then be obtained through maximization of Likelihood function

$$L(\beta,\sigma) = \prod_j \int \prod_i f(y_{ij}|\beta,\sigma,b_j) f(b_j) db_j$$

where $f(b_j)$ is the Normal density function. This integral does not have a closed form except for Normal outcomes, then approximations have to be computed. We fitted GLME models using `lme4` package [16], which makes use of Laplace approximation (Paragraph 4.3.2) for computing high-dimensional integrals.

Considering the database composed by survival response (*Survival*), the categorical variable indicating hospital of admission (*hospital*) and the covariates included in the GLM model (*Age*, *Killip*, *Risk*, *STresolution*) omitting all the NA (missing data), it consists of 1065 statistical units, grouped in 33 hospitals. The output for this model, including the covariates previously explained in the analogous model with fixed effects only is:

```
Formula:
Survival ~ Age + Killip + FE + Risk + STresolution + (1 | hospital)

    AIC    BIC   logLik  deviance
   217.5  252.3  -101.8   203.5

Random effects:
  Groups      Name      Variance   Std.Dev.
   hospital  (Intercept)   0.24187    0.4918

Number of obs: 1065, groups: ospedale, 33

Fixed effects:
                Estimate  Std. Error z value Pr(>|z|)
(Intercept)    2.25645    1.47949   1.525   0.12722
Age           -0.04711    0.01652  -2.852   0.00435  **
Killip        -1.36351    0.43236  -3.154   0.00161  **
FE             0.10598    0.02121   4.996   5.84e-07 ***
Risk          -1.41011    0.68168  -2.069   0.03859  *
STresolution   1.25275    0.41570   3.014   0.00258  **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 ‘ ‘ 1
```

The model for in-hospital survival of patient i treated in-hospital $j[i]$ is then:

$$\begin{aligned} logit(\mathbb{E}[Survival = 1|Age,Killip,FE,Risk,STresolution,hospital)) = \\ 2.25645 - 0.04711 \cdot age_i - 1.36351 \cdot killip_i + 0.10598 \cdot FE_i \\ - 1.41011 \cdot Risk_i + 1.25275 \cdot STresolution_i + hospital_{j[i]} \end{aligned} \tag{7.3}$$

where with $hospital_{j[i]}$ we mean the effect of $j-$th hospital, the one where $i-$th patient has been admitted. Again, as expected, for increasing age, risk and Killip class, the survival probability

decreases, whereas the higher the ejection fraction and the reperfusion outcome, the better the survival. Moreover, it is to be noticed that the residual variability of the random effect is high, attesting the significative contribution coming from the inclusion of the random effect among the model parameters. Moreover, it is known that mixed effects models are suitable for those situation where unbalanced units per group are present, since they "borrow strength" in carrying out estimation for single unit level by upper level modelling. In summary, fitting a GLME model is the more suitable way to account for overdispersion of our data: this will turn in a better estimation and prediction of in-hospital survival where suitable adjustment for case mix is done, but also in a innovative method for clustering hospitals, according to the effect they have on survival itself. In fact it would be of interest to quantify the effect of each hospital on in-hospital survival, but, it would be even more useful to detect if groups of "similar hospitals" are present, and then summarizing the effect of each class on the outcome.

Fitting a GLME model, we obtain also the point estimates of the random effects, i.e., the additive contributions of each hospital to the intercept of the linear predictor for survival probability. Indicating them by $\hat{b}_j$, $j = 1,\ldots,33$, we apply to this set of points a k-means clustering algorithm [69]. A robustness analysis for the number of clusters using the average silhouette width (see [136] and Section 8.1) indicates $k = 3$ as the optimum choice for the number of cluster. Anyway, since no strong evidence exist for discriminating among $k = 2$ or $k = 3$ (as can be evinced observing the average silhouette indexes in Figure 7.16, which indicates that, in both cases of k equal to 2 and 3, a reasonable clustering structure has been found), we will consider both the choices of 2 and 3 groups, in order to be able to compare results with the clustering carried out by the method previously presented.

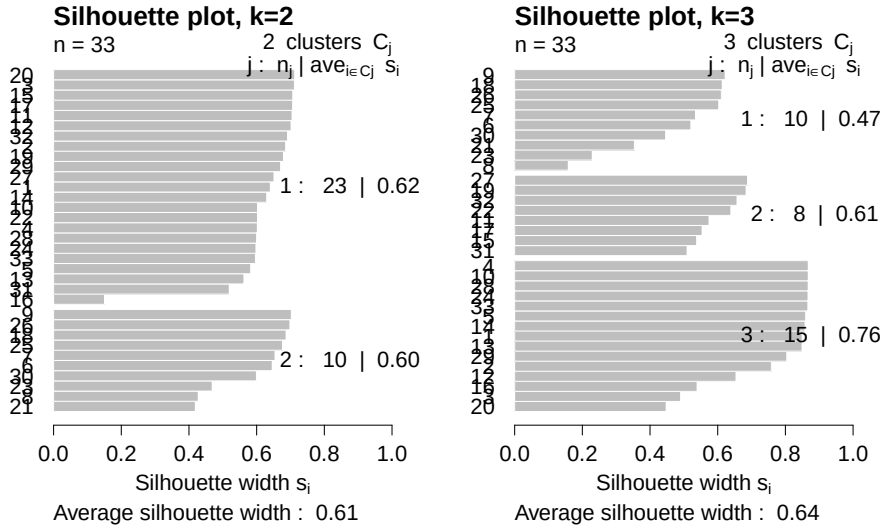


Figure 7.16: Silhouette plots for the choices of the number of cluster, with $k = 2$ (left) $k = 3$ (right) respectively.

The means of the two clusters are 0.08863 and -0.2312 in the case of $k = 2$, and 0.2151, 0.0211 and -0.2312 for the case of $k = 3$. Again, we label the groups as "A" and "B" in the case of $k = 2$, respectively for groups with higher and lower means. In the case of $k = 3$, we add a third name ("C" as central). According to this criterion, we obtain the classification reported in Table 7.9.

	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11
GLME $k=2$	A	A	A	A	A	B	B	B	B	A	A
GLME $k=3$	C	C	C	C	C	B	B	B	B	C	A
	h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22
GLME $k=2$	A	A	A	A	A	A	B	A	A	B	A
GLME $k=3$	C	C	C	A	C	A	B	A	C	B	A
	h23	h24	h35	h26	h27	h28	h29	h30	h31	h32	h33
GLME $k=2$	B	A	B	B	A	A	A	B	A	A	A
GLME $k=3$	B	C	B	B	A	C	C	B	A	A	C

Table 7.9: Providers clustering according to GLME model random effect estimates criterion.

We can observe, that passing from 3 to 2 groups, the hospitals belonging to classes “A” and “C” are simply joined together. The reason is clear observing plots in Figure 7.17.

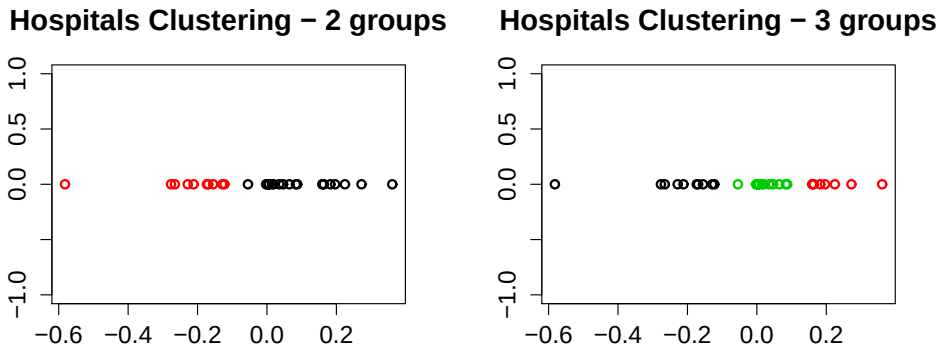


Figure 7.17: Clustering of estimated random effects of the GLMM model in (7.3) into 2 (left panel) and 3 (right panel) groups respectively.

Figure 7.3.2 shows estimated survival surfaces, functions of age and ejection fraction, for 6 different benchmark settings. In particular, in the first row, from left to right “best case” scenario is considered (i.e., patient affected by less severe infarction, with low risk and accomplished ST resolution after 60 minutes), respectively admitted in the centroid hospital of group A (left), C (center) and B (right). Then in the second row, the “worst case” scenario (patient affected by more severe infarction, with high risk and not accomplished ST resolution after 60 minutes) for the same groups is depicted. More specifically:

(a) Less severe infarction (i.e., Killip class I), low risk (i.e., less than 4 risk factors are present), reperfusion accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has been achieved) in the centroid of hospitals belonging to group A;

(b) Less severe infarction, low risk, reperfusion accomplished in the centroid of hospitals belonging to group C;

(c) Less severe infarction, low risk, reperfusion accomplished in the centroid of hospitals belonging to group B;

(d) More severe infarction (i.e., Killip class II, III or IV), high risk (i.e., almost 4 risk factors are present), reperfusion not accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has not been achieved) in the centroid of hospitals belonging to group A;

(e) More severe infarction, high risk, reperfusion not accomplished in the centroid of hospitals belonging to group C;

(f) More severe infarction, high risk, reperfusion not accomplished in the centroid of hospitals belonging to group B.

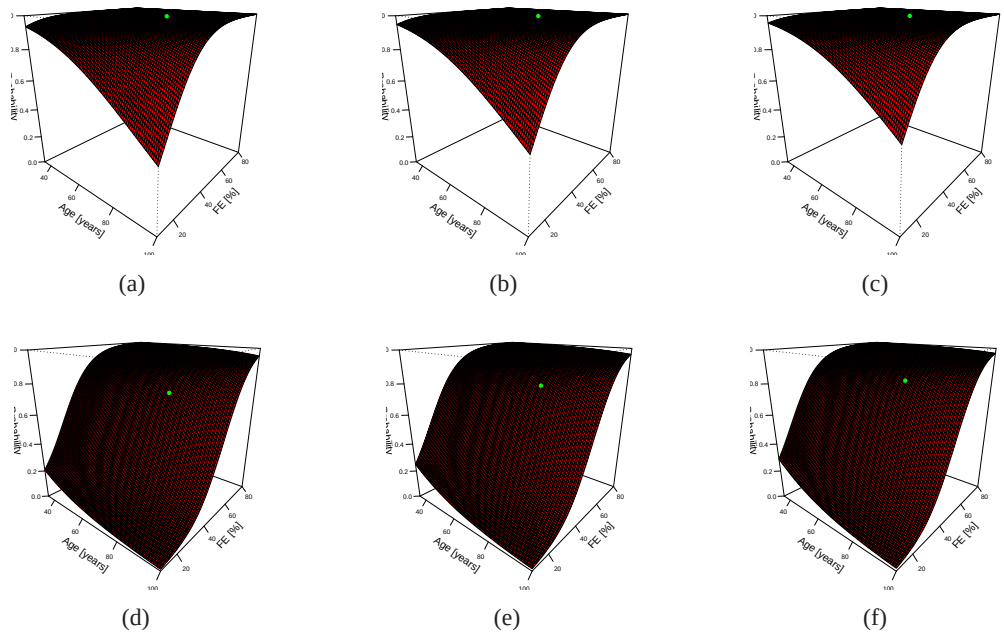


Figure 7.18: Survival surfaces in different case-mix scenarios (best case, first row vs worst case, second row), in the hospital centroid of group "A" (left panel), "C" (central panels) and "B" (right panels) respectively. Green points indicate the survival probability for a reference patient aged 75 and with 50% of ejection fraction at Admittance.

The green points indicate the estimated survival probability for a reference patient aged 75 and with 50% of ejection fraction at the entrance, which are equal, respectively, to 99.59% (case *a*), 99.50% (case *b*), 99.36% (case *c*), 81.21% (case *d*), 78.07% (case *e*), 73.45% (case *f*).

7.3.3 Analysis of random effect estimates of a nonparametric GLME model

In modelling overdispersed and grouped data, we saw that the use of a fully parametric model for random effects could be considered as too restrictive, and then the idea of Non Parametric Maximum Likelihood (NPML) estimation for distribution of random effect has been considered (see Paragraph 4.4.1). We computed non parametric maximum likelihood estimations on our data with `npmlreg` package [37].

Considering the database composed by survival response (*Survival*), the categorical variable indicating hospital of admission (*hospital*) and the covariates included in the GLM models (*Age*, *Killip*, *Risk* and *STresolution*) omitting all the NA (missing data), it consists of 1065 statistical units,

grouped in 33 hospitals. The output for this model, including the covariates previously explained in the analogous model with fixed effects only is:

```
Call:
allvc(formula = Survival ~ Age + Killip + FE + Risk + STresolution,
       random = ~1 | hospital, family = binomial(), k = 2, random
       distribution = "np")

Coefficients:
                Estimate  Std. Error   t value
Age           -0.04666938 0.01622140 -2.877025
Killip        -1.36574575 0.41456261 -3.294426
FE             0.10499803 0.02050766  5.119943
Risk          -1.39121735 0.67262080 -2.068353
STresolution   1.21984131 0.39789590  3.065730
MASS1          1.89308139 1.44684078  1.308424
MASS2          2.59791339 1.44789628  1.794268

Mixture proportions:
        MASS1      MASS2
  0.4838802  0.5161198

Random effect distribution - standard deviation:    0.3522328

-2 log L:     203.7     Convergence at iteration  10
```

where `MASS1` and `MASS2` are the mass points (c_1, c_2) of the nonparametric random effects model, whose corresponding mixing proportions are estimated to be $\omega_1 = 0.4838$ and $\omega_2 = 0.5162$. The model for in-hospital survival of patient i in-hospital j is then:

$$\begin{aligned} logit(\mathbb{E}[Survival = 1 | Age, Killip, FE, Risk, STresolution, hospital]) = \\ 2.25645 - 0.04711 \cdot Age_i - 1.36351 \cdot Killip_i + 0.10598 \cdot FE_i \\ - 1.41011 \cdot Risk_i + 1.25275 \cdot STresolution_i + c_{k[j]} \end{aligned} \tag{7.4}$$

where with $c_{k[i]}$ we mean the effect of $k-$th group ($k = 1, 2$), which is the one hospital j has been assigned to. In fact, the package provides the estimation of masses composing the mixture, as well as the probability distribution of belonging to any of the k groups for each hospital. Again, as expected, for increasing age, risk and killip class, the survival probability decreases, whereas the higher the ejection fraction and the reperfusion outcome, the better the survival. Moreover, also in this case we can see that the residual variability of the random effect is quite high, attesting the significative contribution coming from the inclusion of the random effect among the model parameters. Here we chose to set $k = 2$, i.e., to cluster hospitals in 2 groups, since resulted to be best. This is a further evidence for the real presence of the dichotomic structure among hospitals effects. In summary, fitting a nonparametric GLME model is a suitable way to account for overdispersion of our data with the flexibility due to the nonparametric modelling of the random effects.

In order to use `npmlreg` estimates for clustering, we assign the $j-$th hospital to the $k-$th group according to the arg-max of the probabilities of each structure estimated for the two masses (i.e., assigning each hospital to the group whose estimated probability is greater). Adopting this criterion, we obtain the clustering of hospitals reported in Table 7.10.

h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11
A	A	A	A	A	B	B	B	B	A	A
h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22
A	A	A	A	B	A	B	A	A	B	A
h23	h24	h35	h26	h27	h28	h29	h30	h31	h32	h33
B	A	B	B	A	A	A	B	A	A	A

Table 7.10: Providers clustering according to `npmlreg` random effect estimates criterion.

Figure 7.19 shows the estimated survival surfaces, functions of age and ejection fraction, for 4 different benchmark settings.

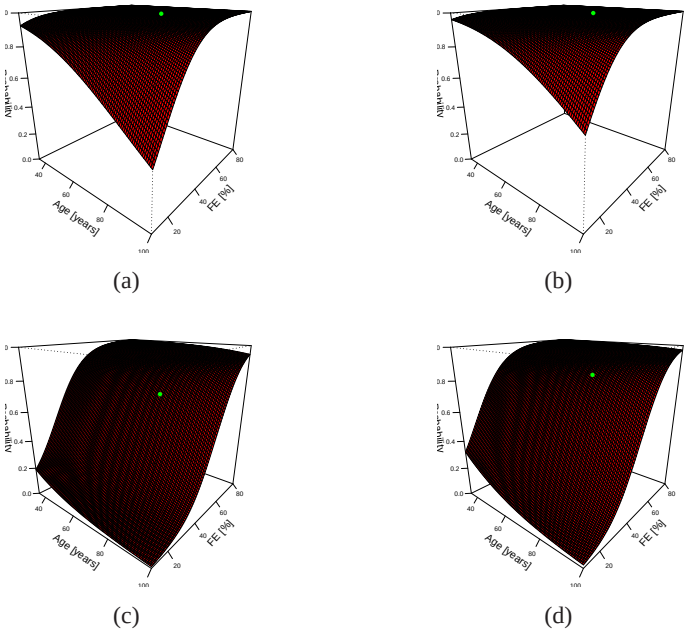


Figure 7.19: Estimted survival surfaces in different case-mix scenarios (best case, first row vs worst case, secondo row), in a hospital belonging to group “A” (left panel) and “B” (right panels) respectively. Green points indicate the survival probability for a reference patient aged 75 and with 50% of ejection fraction at admittance.

In particular, the first row from left to right the “best case” scenario is considered (i.e., a patient affected by less severe infarction, with low risk and accomplished ST resolution 60 minutes after PCI), in a hospital classified as belonging to the group A (left), and B (right) respectively. Then in the second row, the “worst case” scenario (i.e., a patient affected by more severe infarction, with high risk and not accomplished ST resolution 60 minutes after PCI) for the same groups is depicted. More specifically:

(a) Less severe infarction (i.e., Killip class I), low risk (i.e., less than 4 risk factors are present), reperfusion accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has been achieved) in a hospital belonging to group A;

(b) Less severe infarction, low risk, reperfusion accomplished in a hospital belonging to group B;

(c) More severe infarction (i.e., Killip class II, III or IV), high risk (i.e., almost 4 risk factors are present), reperfusion not accomplished (i.e., 70% STsegment elevation reduction 60 minutes after PCI has not been achieved) in a hospital belonging to group A;

(d) More severe infarction, high risk, reperfusion not accomplished in a hospital belonging to group B.

The green points indicate the survival probability for a patient aged 75 and with 50% of ejection fraction at the entrance, which are equal, rspectively, to 99.61% (case *a*), 99.23% (case *b*), 83.07% (case *c*), 70.80% (case *d*). All the values are similar to the corresponding ones computed starting from the parametric GLME model with 2 groups.

7.3.4 Comparison of different methods

In the previous paragraph, we focused on the influence of group effect on survival prediction at patient level. We will now compare (see Table 7.11) classifications of providers pointed out by the different methods proposed.

In conclusion, following three different clustering procedures, we obtain the same clustering structure except for hospital 16 which is classified as belonging to the group “B” of poorer-performing hospitals according to SSR and nonparametric GLME methods, as belonging to the group “A” of better-performing hospitals according to parametric GLME method with 2 groups, and as belonging to the group “C” of central-performing hospitals according to parametric GLME method with 3 groups. The nearly unanimous agreement in the classification of the three methods support the idea that a real clustering structure in two groups exists.

	h1	h2	h3	h4	h5	h6	h7	h8	h9	h10	h11
SSR	A	A	A	A	A	B	B	B	B	A	A
GLME param - 2 clusters	A	A	A	A	A	B	B	B	B	A	A
GLME param - 3 clusters	C	C	C	C	C	B	B	B	B	C	A
GLME nonparam - 2 clus.	A	A	A	A	A	B	B	B	B	A	A
	h12	h13	h14	h15	h16	h17	h18	h19	h20	h21	h22
SSR	A	A	A	A	B	A	B	A	A	B	A
GLME param - 2 clusters	A	A	A	A	A	A	B	A	A	B	A
GLME param - 3 clusters	C	C	C	A	C	A	B	A	C	B	A
GLME nonparam - 2 clus.	A	A	A	A	B	A	B	A	A	B	A
	h23	h24	h35	h26	h27	h28	h29	h30	h31	h32	h33
SSR	B	A	B	B	A	A	A	B	A	A	A
GLME param - 2 clusters	B	A	B	B	A	A	A	B	A	A	A
GLME param - 3 clusters	B	C	B	B	A	C	C	B	A	A	C
GLME nonparam - 2 clus.	B	A	B	B	A	A	A	B	A	A	A

Table 7.11: Providers clustering provided by the three different criteria described in Section 7.3.

7.4 Bayesian Hierarchical Models for Hospital Clustering

In this Section, we apply models and techniques described in Chapter 5 to STEMI Archive data, with the aim to model in-hospital survival in the treatment of STEMI. As we said in Paragraph 3.2.2, similar studies were conducted in the Milano Cardiological Network (see [56], [57], [58], [60], [62], [74], [75], [78] and [84]). Now we consider a larger and more heterogeneous set of hospitals, i.e., those belonging to the network of all hospitals of Regione Lombardia, with the aims of identifying differences between providers, pointing out factors that increase/decrease the probability of survival (both at patient and hospital level), and evaluating the efficiency of process indicators, using Bayesian methods.

The idea of the analysis, detailed in [64], is to compare results in modelling in-hospital survival arising from different model setting, from parametric to semi-parametric Bayesian models, and in particular to point out a new method for prediction and classification of new patients, within the context of strongly unbalanced shares. In fact, as we mentioned in Section 4.7, when the outcome is particularly unbalanced it is difficult to evaluate the fitting of the model and predict the within sample negative outcomes. From a frequentist perspective, Cramer [28] suggests a criterion to improve the predictive capacity of negative outcomes. Cramer's criterion is based on point estimate, whereas here we propose a new approach based on interval estimate of the posterior predictive distributions proposed in Section 5.5.

As we saw in Paragraph 3.2.3, there is a hierarchical structure in the data arising from STEMI Archive: the providers at a higher level and the patients at a lower one. Bayesian generalized linear mixed models (see Paragraph 5.2.2) provide a natural framework for such kind of data. Concerning the models considered in this section, in the first one we put a parametric prior on all factors, whereas in the second and the third ones lower level factors are treated parametrically, while higher level factors are treated in a nonparametric way, according to [92]. In particular, in the second model we considered a bivariate Dirichlet process (DP) prior for the random effects, while in the third model random effects with a Dependent Dirichlet process (DDP) prior are assumed, according to the setting proposed in [107]. The DDP prior will take into account specific hospital-covariates, yielding dependency among the distribution of parameters of different subpopulations; in particular we include a geographical binary covariate *Milano* (equal to 1 if the hospital is in Milano, equal to 0 otherwise) in the semiparametric prior for the providers' random effects. Both DP and DDP priors relax the parametric assumption and induce a grouping of the random effects. Relaxing the parametric assumption brings to more flexible priors and better estimates, while the random effects' clustering provides a starting point for providers' profiling.

The information provided by STEMI Archive we are most interested in are mode of admission of each patient (spontaneous or delivered by different types of 118 rescue units), personal data (age, sex), clinical appearance (Killip class), risk factors (diabetes, smoke, Chronic Kidney Disease CKD, ...), pre-hospital and in-hospital treatment times, and clinical outcomes (in-hospital survival, MACE, ST-resolution). Killip classification with values $\{1,2,3,4\}$ is used to risk stratify patients, being 1 the less severe class of infarction, and 4 the most severe one. In this study we focus on in-hospital survival probability. Moreover, we make use of information provided by the administrative databanks concerning the grouping factors (the hospitals). From the linkage with database *Ricoveri* it is possible to know if the hospital is in Milano or outside and the hospital's exposure. Here the exposure of a hospital is the number of patients treated there with primary PCI in a year. We consider a sub population of patients for which all covariates of interest are present and filled in the correct way, focusing on patients underwent primary PCI and not-transferred to another hospital.

Then the dataset consists of 697 patients from $J = 29$ hospitals. There is a large heterogeneity of the number of patients for each hospital. Moreover, in-hospital survival is particularly unbalanced in our sample (97% of patients were discharged alive), coherently with literature [158].

A first patient covariates' selection was done according to clinical know-how and frequentist selection procedures; the most significant factors which explain survival probabilities are age, the time from symptom Onset to Balloon in log-scale (logOB), Killip and CKD. In addition to the above mentioned patient covariates, in our models we considered the provider covariates (*Milano* and exposure). In fact, we are interested in evaluating if there are differences among the hospitals and, in case, if those differences are related to particular characteristics of the providers.

7.4.1 Parametric and semiparametric models

As mentioned before, in the first model we considered all factors parametrically, while in the second and the third models the random effects normality assumption is relaxed and treated semiparametrically. In particular, in the second model we use a bivariate DP prior for the hospitals' random intercepts and the hospitals' exposure random slopes, while in the last model we remove the hospitals' exposure random slopes and we use a DDP prior for the providers' random effects including the geographical binary covariate *Milano*. Finally, we provide the latent variable representation [5] of the logistic regression, which is useful in performing Bayesian inference and model checking.

The parametric model

For patient $i = 1,\dots,n_j$, in each group $j = 1,\dots,J$, let Y_{ij} be a Bernoulli random variable with mean p_{ij}, which represents the probability that the patient i treated in-hospital j survived after STEMI. The p_{ij}'s are modelled through a logit regression with covariates and group (random) effects:

$$Y_{ij}|p_{ij} \overset{\text{ind}}{\sim} Be(p_{ij}) \tag{7.5}$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{l=1}^{4} \alpha_l u_{ijl} + \sum_{k=1}^{5} \beta_k x_{ijk} + b_{0j} + b_{1j} z_j, \tag{7.6}$$

where $\boldsymbol{u}_{ij} = (u_{ij1},\dots,u_{ij4}) = (\text{killip1},\dots,\text{killip4})_{ij}$ is a dummy vector, $\boldsymbol{x}_{ij} = (x_{ij1},\dots,x_{ij5}) = (\text{age}, \text{logOB}, \text{CKD}, \text{exposure}, \text{Milano})_{ij}$ and $z_j = \text{exposure}_j$. The prior is

$$\begin{aligned} &\alpha_1,\dots,\alpha_4,\beta_1,\dots,\beta_5 \overset{\text{i.i.d.}}{\sim} N(0,100) \\ &\boldsymbol{b}_1,\dots,\boldsymbol{b}_J|\Sigma_b \overset{\text{i.i.d.}}{\sim} N_2(\boldsymbol{0},\Sigma) \end{aligned} \tag{7.7}$$

where

$$\boldsymbol{b}_j = (b_{0j}, b_{1j}) \text{ and } \Sigma = \begin{bmatrix} \sigma_0^2 & \rho\sigma_0\sigma_1 \\ \rho\sigma_0\sigma_1 & \sigma_1^2 \end{bmatrix}, \tag{7.8}$$

while

$$\sigma_0, \sigma_1 \overset{\text{i.i.d.}}{\sim} Unif(0,5) \text{ and } \rho \sim Unif(-1,1).$$

Dirichlet process model

As far as the second model is concerned, we relax the random effects parametric assumption and put a DP prior on them. The likelihood is as in (7.5) and the prior is

$$\begin{aligned}
&\alpha_1,\dots,\alpha_4,\beta_1,\dots,\beta_5 \overset{\text{i.i.d.}}{\sim} N(0,100)\\
&\boldsymbol{b}_1,\dots,\boldsymbol{b}_J|P \overset{\text{i.i.d.}}{\sim} P\\
&P|a,P_0 \sim DP(a,P_0) \text{ where } P_0|\sigma_0^2,\sigma_0^2 \text{ is } N(0,\sigma_0^2)\times N(0,\sigma_1^2)\\
&a \sim trunc-Exp(1) \text{ with support } (1,+\infty)\\
&\sigma_1,\sigma_2 \overset{\text{i.i.d.}}{\sim} Unif(0,5).
\end{aligned} \tag{7.9}$$

Dependent Dirichlet process model

Here we remove the exposure effect (both the fixed and the random slope) since we found it is not statistically significant (see the analyses below) and considered an ANOVA-DDP prior for the covariate *Milano*, as in the framework proposed in [33] and in Paragraph 5.3.2. The conditional distribution of the outcomes under (7.6) is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_{l=1}^{4}\alpha_l u_{ijl} + \sum_{k=1}^{3}\beta_k x_{ijk} + b_{jz_j}, \tag{7.10}$$

where $\boldsymbol{u}_{ij}$ is the killip dummy vector, $\boldsymbol{x}_{ij} = (x_{ij1},x_{ij2},x_{ij3}) = (\text{age},\text{logOB},\text{CKD})_{ij}$ and b_{jz_j} is the *Milano* ($z_j = 0/1$) hospital random intercept. Unlike (7.5) - (7.9), here we distinguish the random intercept parameter according to the geographical origin of the hospital: b_{j1} if the j-th hospital is in Milano, b_{j0} otherwise. We assume a prior dependency between b_{j0} and b_{j1} through a DDP. The prior is

$$\begin{aligned}
&\alpha_1,\dots,\alpha_4,\beta_1,\dots,\beta_3 \overset{\text{i.i.d.}}{\sim} N(0,100)\\
&b_{jz_j}|P,z_j \overset{\text{ind}}{\sim} P_{z_j}\\
&P_{z_j}|P_{0z_j},a \overset{\text{ind}}{\sim} ANOVA-DDP(a,P_{0z_j})\\
&P_0|\boldsymbol{\mu}_b,\Sigma_b \text{ is } N_2(\boldsymbol{\mu}_b,\Sigma_b)\\
&a \sim trunc-Exp(1) \text{ with support } (1,+\infty)
\end{aligned}$$

where

$$\Sigma_b = \begin{bmatrix} \sigma_0^2 & 0\\ 0 & \sigma_1^2 \end{bmatrix},\ \boldsymbol{\mu}_b = \begin{bmatrix} 0\\ \mu_1 \end{bmatrix} \text{ and } P_{0z_j} = \begin{cases} N(0,\sigma_0^2) & if \quad z_j = 0\\ N(\mu_1,\sigma_1^2) & if \quad z_j = 1. \end{cases} \tag{7.11}$$

Moreover, we assume

$$\sigma_0,\sigma_1 \overset{\text{i.i.d.}}{\sim} Unif(0,5) \text{ and } \mu_1 \sim N(0,10).$$

The marginal prior of $(b_1,\dots,b_J)$ is partially exchangeable; more precisely the prior distribution for (P_0,P_1), representing the distributions of the hospital random effects outside and in Milano, respectively, is a bivariate Dirichlet process

$$\begin{pmatrix} P_0\\ P_1 \end{pmatrix}\Bigg| a, \begin{pmatrix} P_{00}\\ P_{01} \end{pmatrix} \sim DP\left(a, \begin{pmatrix} P_{00}\\ P_{01} \end{pmatrix}\right),$$

i.e., for z equal to 0 or 1,

$$P_z = \sum_{h=1}^{+\infty} w_h \delta_{b_{zh}}, \quad \begin{pmatrix} b_{0h} \\ b_{1h} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N_2\left(\begin{pmatrix} 0 \\ \mu_1 \end{pmatrix}, \Sigma \right).$$

Observe that marginally P_z is $DP(\alpha, P_{0z})$, whereas the covariance structure of the locations b_{zh}'s across z yields dependency between P_0 and P_1. Since the dependance structure makes the prior richer and more flexible, we expect better estimates of the hospital random intercepts.

In (7.9) we put a DP prior on the random effects, while in the last one we use a generalization of the DP prior. It is well known that the DP selects discrete distribution almost surely [40]. Since there is a positive probability of coincident values, sampling from P induces a random partition on the positive integers. More specifically, let $X_1, \ldots, X_n$ be a sample from a DP P (on $\mathbb{R}^k$ for some positive integer k), i.e., $X_1, \ldots, X_n$, given P, are i.i.d. from P. Since P is almost surely discrete, two sampled random variables X_i and X_j could be equal with positive probability. We say that X_i and X_j share the same cluster if and only if $X_i = X_j$. In this case, the set of integers $\{1, 2, \ldots, n\}$ is partitioned into a finite number of sets $\{A_1, \ldots, A_{k(n)}\}$, where $k(n)$ is the number of different determinations among $(X_1, \ldots, X_n)$ and each A_j contains the labels of the random variables $(X_1, \ldots, X_n)$ which coincide but are different from the others. Note that since $(X_1, \ldots, X_n)$ is a random vector, the partition $\{A_1, \ldots, A_{k(n)}\}$ of $\{1, 2, \ldots, n\}$ is random as well. This is what is usually meant by random partition induced by the sampling from a DP (or from a random probability measure which is discrete with positive probability). The nonparametric models based on DP priors (or like) are very useful when the aim is clustering. For instance, in the DP model a priori the hospital random effects $\boldsymbol{b_1}, \ldots, \boldsymbol{b_J}$ are a sample from P on $\mathbb{R}^2$ (intercept and slope). Since the DP is conjugate, a posteriori the random effects are still a bivariate sample from a DP and hence we could have coincident values among them. Instead, in the DDP model we could observe coincident values within each subpopulations.

7.4.2 Posterior inferences and prediction

In this paragraph we present the posterior inferences obtained from the three models introduced so far. First we provide posterior estimates of the parameters for each model, focusing in particular on posterior interval estimates and clustering of the hospital random intercepts; then we evaluate their fit and classify the patients. All estimates were computed using the program JAGS [123] via Gibbs sampler algorithms. In the two nonparametric models we implemented the truncated DP approximation suggested by [87] to obtain a trajectory from P. We ran the three models for 200.000 iterations, the first 100.000 were discarded, we used a thinning of 20 to reduce autocorrelation and so the final sample size was 5.000. Traceplots, autocorrelations and Geweke diagnostics indicate that the Gibbs sampler algorithms could have converged.

A robustness analysis showed that inferences are not particularly sensitive to the choice of the fixed effects' hyperparameters, while they are quite sensitive to the prior choice on the variance components σ_0^2 and σ_1^2. We used two classes of priors: the conjugate inverse-gamma distribution on the variances or the uniform distribution on the standard deviations. The estimates of the random effects are particularly sensitive to the choice of the inverse-gamma hyperparameters, while they are more robust using the uniform prior. We refer to [50] for a discussion on priors of the variance components in hierarchical models. The lower bound of the support of the prior distribution for the total mass parameter was set equal to 1 to avoid computational problems. We assumed an exchangeable prior for the killip effect parameter vector, instead of indipendent priors for each vector component, but the inferences were quite the same under the two different choices.

In Table 7.12 we provide posterior 95% Credibility Intervals (CIs) of the fixed effects under the three models. Notice that all the estimates are similar. In particular, the Killip seems a good stratification parameter for all models, since the posteriors of the Killip 1 parameter concentrate on "high" values (i.e., it brings to high survival probability), those of Killip 2 and 3 concentrate on "average" values, while those of Killip 4 concentrate on "small" values. As we could expect, as long as age or logOB or CKD increases, it will have a negative effect on the survival probability. Finally, the *Milano* covariate has a weak negative effect in both parametric and DP models, while the exposure is not significant, i.e., it seems that the number of patients treated with primary PCI does not improve the survival probability. as suggested by preliminar descriptive analyses of benchmarks in Section 7.1. For this reason we decided to omit the exposure from the last model, but used *Milano* covariate to enrich the hospital random intercept prior distribution.

	Parametric model			DP model			DDP model		
Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
killip1	5.14	7.46	10.80	3.22	7.12	11.71	2.24	5.00	8.32
killip2	3.00	5.48	9.09	0.97	5.17	10.06	0.35	3.24	6.68
killip3	2.00	4.78	8.19	0.27	4.50	9.49	-0.72	2.41	5.98
killip4	-1.11	1.41	4.19	-2.49	1.47	5.96	-3.86	-0.85	2.43
age	-3.87	-2.00	-0.37	-3.81	-2.03	-0.42	-3.53	-1.73	-0.13
log(OB)	-4.43	-2.54	-0.41	-4.04	-2.28	-0.39	-4.32	-2.54	-0.62
CKD	-3.72	-2.03	-0.47	-3.56	-1.92	-0.43	-3.86	-2.34	-0.87
exposure	-3.06	0.48	3.96	-4.47	0.24	5.96			
Milano	-5.83	-2.68	0.06	-5.24	-2.38	-0.19			

Table 7.12: Posterior 95% CIs of the fixed effects

As mentioned before, one of the main aim of the analysis is to investigate if there are hospitals (or groups of hospitals) with a better/worse service than others. In other words, we want to evaluate differences among the provider random effects and cluster them in some way.

In Table 7.13 we provide posterior 95% CIs of the components of the covariance matrices of the hospital random effects $\boldsymbol{b}_j$'s of the parametric and the DP model. Notice that in the parametric model the outer diagonal term of the matrix Σ in (7.8) does not seem significantly different from zero and hence we omitted the correlation parameter ρ from the DP model (see P_0 in (7.9)). We recall that in the DP model the variance and the covariance of the j-th hospital random effect could be represented by the following random variables

$$Var[b_{kj}|P] = \mathbb{E}[b_{kj}^2|P] - (\mathbb{E}[b_{kj}|P])^2, \;\; k = 0,1,$$
$$Cov[b_{0j}, b_{1j}|P] = \mathbb{E}[b_{0j}b_{1j}|P] - \mathbb{E}[b_{0j}|P]\mathbb{E}[b_{1j}|P],$$

where

$$\mathbb{E}[b_{0j}^m b_{1j}^n|P] = \sum_{h\geq 1} p_h b_{0h}^m b_{1h}^n, \quad m,n = 0,1,2,\ldots.$$

The interval estimates of the variances of the random intercept are similar under the first two models (σ_0^2 and $Var[b_{0j}|P]$, respectively), while the interval estimate of the variance of the random slope seems shorter for the DP model than for the parametric one. Moreover, the covariance $Cov[b_{0j}, b_{1j}|P]$ of the DP model seems centered around zero.

When considering the DDP model, the random variables considered in Table 7.13 have a different interpretation: in fact they represent the variances of the two different subpopulations and the

covariance between them, respectively. In particular, we observe a greater heterogeneity among the hospitals in Milano then among those outside Milano, since the posterior 95% CIs of $Var[b_{0j}|P]$ and $Var[b_{1j}|P]$ we obtained are $(0.01, 8.92)$ and $(0.67, 30.80)$, respectively. The covariance among the two subpopulations is not significantly different from zero also in this model. We recall that the prior marginal covariance is zero, since $Cov[b_{0j}, b_{1j}] = Cov[P_{00}, P_{01}]/(a+1) = 0$ for any given a (since the outer diagonal elements of Σ in (7.8) are zero).

(a) Parametric model

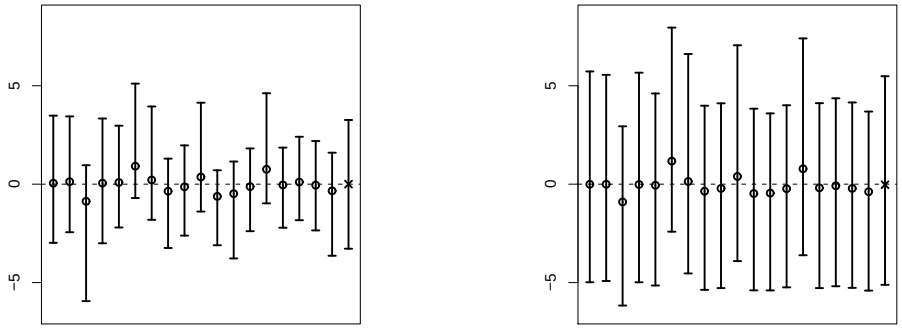
Parameter	2.5%	50%	97.5%
σ_0^2	0.04	2.25	13.91
σ_1^2	0.11	3.86	22.51
$\rho\sigma_0\sigma_1$	-4.94	0.36	9.47

(b) DP model

Parameter	2.5%	50%	97.5%
$Var[b_{0j}\|P]$	0.00	1.37	14.75
$Var[b_{1j}\|P]$	0.00	1.20	15.02
$Cov[b_{0j}, b_{1j}\|P]$	-3.58	0.03	5.61

Table 7.13: Posterior 95% CIs of the random effects' covariance matrices elements.

In Figure 7.20 we provide posterior 95% CIs of the hospital random intercepts with at least ten patients for the parametric model (Figure 7.20, left panel) and for the DP model (Figure 7.20, right panel). The posterior medians under the DP model are more shrunk towards zero and the widths of the intervals are larger. The widths of the DP model's new random intercepts are also larger than the parametric ones. The plots of the hospital slopes (exposure) for both models show CI's even more shrunk towards zero and for this reason we do not include them here.



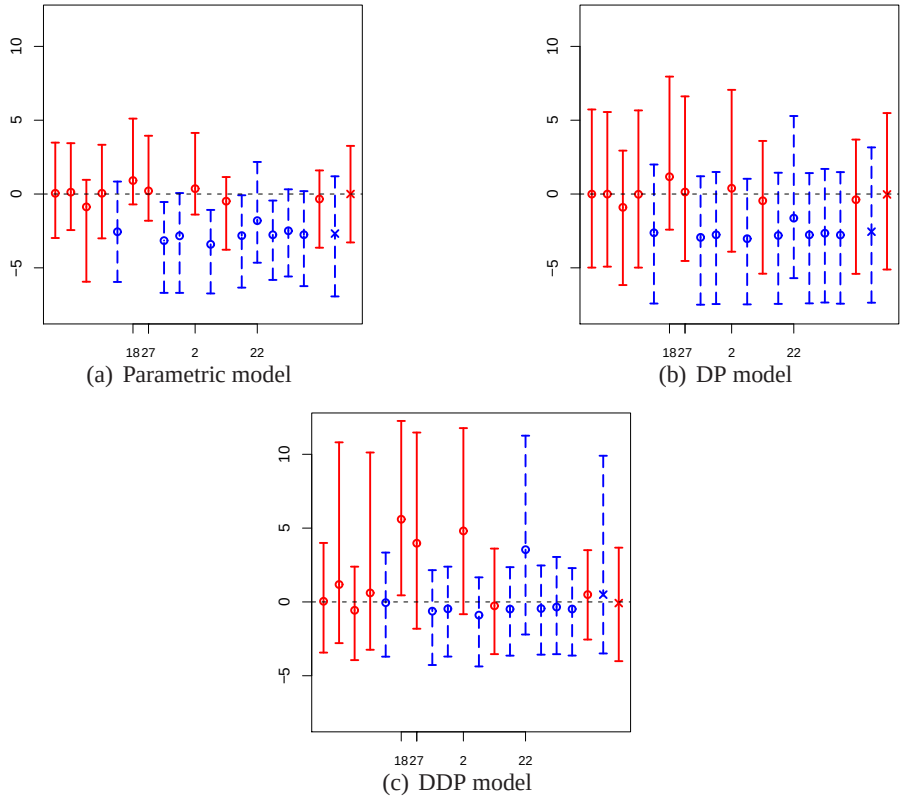
(a) Parametric model (b) DP model

Figure 7.20: Posterior 95% CIs of hospital random intercepts with at least ten patiens. The estimates are in increasing order of number of patients. The last estimates represent new random intercepts.

We cannot compare the hospital random intercepts of the DDP model directly to those of the previous models, since in the DDP model the *Milano* covariate is included in the DDP prior. The hospital random intercepts of the DDP model are equivalent to the hospital random intercept plus the *Milano* effect of the parametric and the DP models.

In Figure 7.21 we provide posterior 95% CIs of the hospital random intercepts with at least ten patients plus the *Milano* effect for the parametric and DP models on the top, and the hospital random intercepts for the DDP model on the bottom. In the parametric and the DP models all hospitals outside *Milano* have higher median than *Milano* ones, and in the parametric model intervals are shorter. Notice that in the DDP model there is larger variability within each of the two subpopulations. We can guess that this variability is due to the flexibility of DDP prior. However, we obtain a better clustering in the DDP model, since it is clear from Figure 7.21 (c) that we can identify two groups:

the first group with hospitals 18, 27, 2 and 22 and the second one with all the others. On the other hand in the parametric and the DP models the grouping is less clear and mainly due to the *Milano* effect.



(a) Parametric model (b) DP model

(c) DDP model

Figure 7.21: Posterior 95% CIs of hospital random intercepts plus *Milano* effect with at least ten patiens. The hospital effect in Milano are depicted in blue dashed, those ooutside Milano in red solid lines. The estimates are in increasing order of number of patients per hospital. The last two intervals represent new random intercepts for a hospital in and outside Milano, respectively.

As mentioned before, the DP induces a random partition on the grouping of the random effects and so we analyse the posterior of the process P to obtain an insight on the clustering among the hospitals. The mass parameter a is a posteriori concentrated around small values under both parametric models: mean 1.60 (std. dev. 0.62) in the DP model and 1.65 (std. dev. 0.66) in the DDP one. Consequently we observe a reduction of the expected number of groups, which from the prior mean of 5.8 becomes a posteriori 5.1 for the DP and 5.2 for the DDP. The a priori expected number of clusters is slightly higher than our actual *a priori* knowledge since we would expect less groups, indicating macro behavioural setting. We run the algorithm fixing the mass parameter a equal to one (doing so, the expected number of cluster is 4.0) and we obtained similar posterior estimates; hence we can conclude that the inference is quite robust to the prior specification of the mass parameter a. Since the DP induces a random partition, the posterior partition mode could provide an estimate of the clustering among the hospitals. However the posterior partition mode is the one with all hospitals in the same group, but it is reached only $135/5000$ and $4/5000$ times in the DP and DDP, respectively, and hence there is much uncertainty in both posteriors of the random partitions.

7.4.3 Model fit and patients classfication

Once we fitted our models and compared their estimates, we focus on the evaluation of their predictive performance by computing survival probabilites for each given patient. In particular, we compare two different predictive methods: the usual one based on point estimates summarizing the posterior predictive distributions and a new one based on interval estimates. As will result in the following, our method leads to a more accurate classification. The usual predictive method is based on point estimate of the posterior predictive distribution; in particular we could classify a patient i from hospital j as alive if $\mathbb{E}(p_{ij}|\boldsymbol{Y})$ is bigger than a given cut-off point. Since the classification is typically sensitive to the cut-off point, there are several criteria to choose the cut-off point; see [45] for a review and comparison of the most popular ones in the frequentist literature.

In our application, since our dataset is particularly unbalanced, if we consider the standard cut-off point equal to 0.5, we would obtain a very low overall misclassification rate (around 2% for all models), but a bad negative predictive power (more than 50% of the deaths are misclassified). Table 7.14 displays the results of the patient classification under the three models considered here, using a cut-off point equal to the survival sample proportion $\bar{p} = 0.97$, as suggested by Cramer [28]. The false positive and false negative rates are more balanced then using a cut-off point of 0.5, but we obtain a worse overall misclassification rate (around 10% for all models). The overall misclassification rate as in [28] can be considered as a goodness of fit index since it is less dependent on the unequal sample proportion. On the other hand, instead of choosing a given cut-off point, we could plot the ROC curve (see Figure 7.22). The overall misclassification rate (with cut-off point equal to the sample proportion) and the ROC curve shows a good and similar predictive fit of the three models.

(a) Parametric model

	$Y=1$	$Y=0$
$\hat{Y}=1$	605	3
$\hat{Y}=0$	69	20

(b) DP model

	$Y=1$	$Y=0$
$\hat{Y}=1$	599	3
$\hat{Y}=0$	75	20

(c) DDP model

	$Y=1$	$Y=0$
$\hat{Y}=1$	600	3
$\hat{Y}=0$	74	20

Table 7.14: Predictive tables using point estimates and cut-off point equal to $\bar{p} = 0.97$.

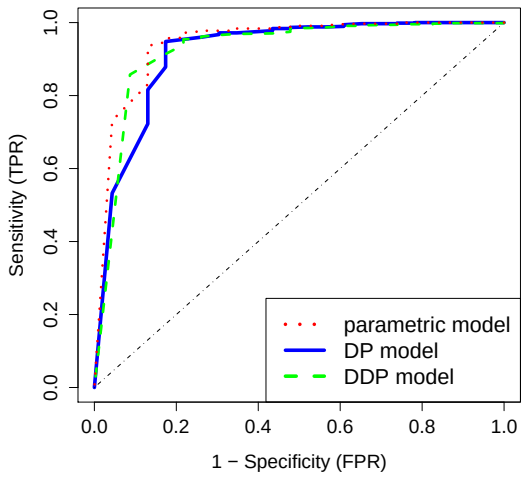


Figure 7.22: ROC curves for the three models.

However, we believe that the previous results are not completely satisfactory. On one hand those cut-off criteria are not robust in case of a very unbalanced data-set, as in our case, coherently with the analysis in [45]. On the other hand,the ROC curve is indipendent of a given threshold, it can be used as a comparison tool among models rather then as a predictive tool itself. To this aim, we may take advantage of Bayesian approach: in fact, interval estimate is richer than point estimate which does not provide any information on the prediction uncertainty. The new method we propose is based on interval estimate and is a straightforward generalization of the classical one.

We classify the patient as alive if the entirely interval estimate is over a given cut-off point, as dead if the entirely interval estimate is above the cut-off point and we do not classify it if the cut-off point lies in the CI. The higher is the credible level, the more patients will belong to the Uncertainty Class (UC). In Table 7.15 we report classification tables based on 90% posterior predictive CIs and assume equal misclassification costs, i.e., the cut-off point is set equal to 0.5. With our data-set, only around 4% of the patients belong to the UC and the total misclassification rate, based only on classified patients, is around 1% for all the models. If the number of patients in UC provides an index of the predictive performance of the model, then the three models have similar results.

(a) Parametric model

	$Y=1$	$Y=0$
$\hat{Y}=1$	659	7
$\hat{Y}=0$	0	3
UC	15	13

(b) DP model

	$Y=1$	$Y=0$
$\hat{Y}=1$	660	7
$\hat{Y}=0$	0	4
UC	14	12

(c) DDP model.

	$Y=1$	$Y=0$
$\hat{Y}=1$	657	7
$\hat{Y}=0$	0	3
UC	17	13

Table 7.15: Predictive tables using 90% CIs and cut-off point equal to 0.5.

In Figure 7.23 we provide the 90% posterior predictive CIs for all patients under the DDP model (the plots for the other two models are quite similar and we do not report them here). Notice that most of the interval widths of the survived patients are quite small, while there is more uncertainty on the negative outcomes.

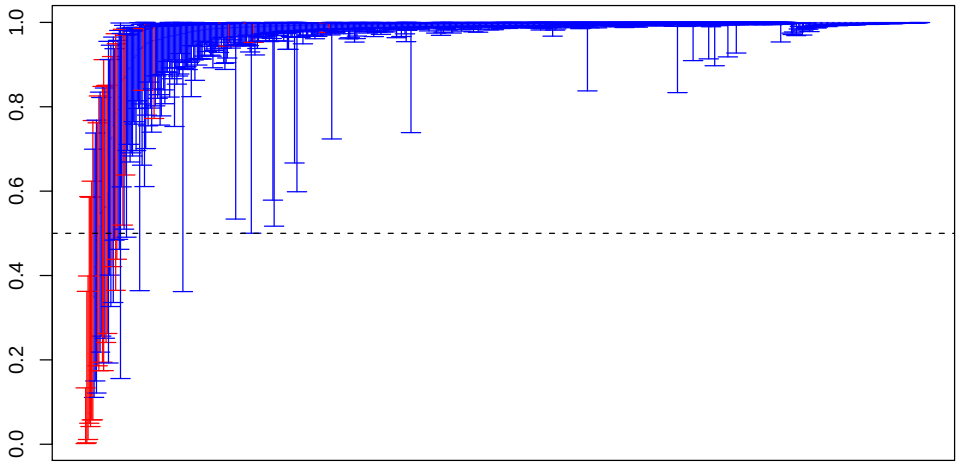


Figure 7.23: 90% posterior predictive CIs of all the patients (ordered by increasing median) under the DDP model. The positive outcomes are in blue and the negative ones in red.

As an example, in Figure 7.24 we focus on a smaller set of patients, those 29 treated in-hospital 19 (under the DDP model). Notice that predictive distributions with very large and very low mean

have small width, while those with mean around 0.5 have wider interval estimates. There are 6 unclassified patients (but in one case the cut-off point falls on the lower estreme of the inteval) and only one is misclassified

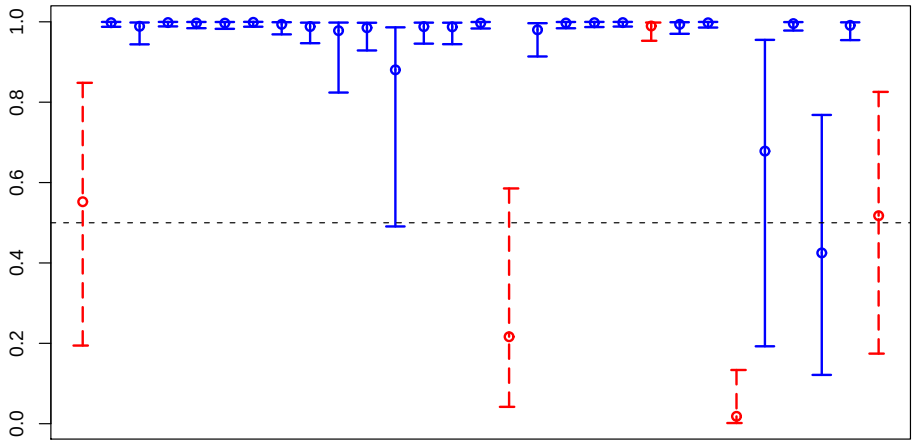


Figure 7.24: 90% posterior predictive CIs of all the patients from hospital 19, under the DDP model. The CIs corresponding to alive patients are in blue solid line, while those corresponding to dead patients are in red dasehd. There are six unclassified patients and only one misclassified.

Chapter 8

Statistical analysis of other clinical surveys

In this chapter, the analyses carried out on data described in Paragraph 3.2.2 and Section 3.3 are presented. These are examples of data mining on administrative databanks (Section 8.1 and Section 8.2) carried out with statistical techniques presented in Section 4.6, and of Bayesian decision analysis (Section 5.4) applied to the evaluation of acceptability of providers' performances.

8.1 Nonlinear parametric models for an epidemiologic enquire

In this section we illustrate a pilot data mining case study on hospital discharges data for patients with NON-STEMI diagnosis. The behaviour of NON-STEMI cases over the years is, in fact, a very interesting problem for epidemiologists, here faced for the first time using statistical methods applied to administrative data. In fact, data come from PHD of Regione Lombardia (Paragraph 3.1.1), and the study is part of the Strategic Program (Section 2.2). This study represents an example of unsupervised clustering carried out starting from the random effects estimates in a NLME model setting, according to method and purposes highlighted in Section 4.2.3.

The statistical analysis is conducted along different phases. The visual evidence for growth in the number of NON-STEMI diagnoses is firstly questioned by fitting a semiparametric mixed effect model, in order to capture the shape of growth curves and to test the significance of the grouping factor effect. The relevant features emerged with this first analysis are then modeled by means of parametric nonlinear models of decreasing complexity, which are easier to interpret and more suited to inferential purposes. We focused on the numbers of hospital discharges with a diagnosis of NON-STEMI, grouped by hospital and relative to the 30 largest clinical institutions of Lombardia Region, during years $2000 - 2007$. Cases detection is performed according to the AHQR guidelines [173]. Figure 8.1-left panel represents the number of Acute Myocardial Infarction without ST-elevation (NON-STEMI) diagnoses, along the time period $2000 - 2007$, for the 30 hospitals. The total number of diagnoses in the time period $2000 - 2007$ has a considerable variability between institutions: in fact it ranges from a minimum value of 715 to a maximum of 1872. This difference is due to the different exposure of different hospitals; indeed, exposure could be a confounding factor in a statistical analysis focused on the growth trend of the number NON-STEMI cases. Hence, in order to analyze comparable data, for each hospital the yearly number of diagnoses has been standardized by the hospital total number of diagnoses in the time period $2000 - 2007$, thus adjusting for hospital exposure (see Figure 8.1-rigth panel).

The high variability between hospitals and the structure of the data grouped by hospital, motivate

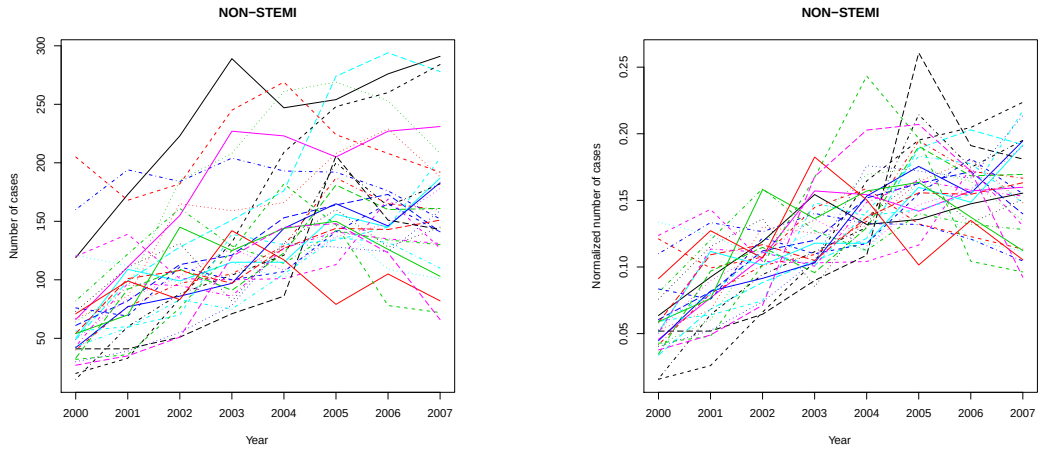


Figure 8.1: Left panel: Number of AMI without ST-elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. Right panel: Standardized number of AMI without ST-elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. For each hospital the yearly number of diagnoses has been divided by the hospital total number of diagnoses in the time period $2000-2007$.

the use of mixed effects models for the analysis of these longitudinal data. A first explorative analysis conducted by means of a linear mixed model, where the standardized number of NON-STEMI diagnoses appears as a linear function of time, with hospital as a grouping factor, shows a significant linear trend over time (the p-value of the test on the "year" fixed effect is less than 10^{-14}). Since the use of a linear parametric model can be quite binding, a further enquire into the growth trend has been conducted by fitting a semiparametric mixed effect model. Indeed, we set $\tilde{N}_{ij}$ to be the standardized number of NON-STEMI diagnoses for hospital $i = 1,...,30$ and year $j = 1,...,8$, where $j = 1$ is for year 2000 and $j = 8$ is for year 2007, and following [147], we fit the following mixed effects semiparametric model with respect to time

$$\tilde{N}_{ij} = s(t_j) + b_{0i} + b_{1i}t_j + \varepsilon_{ij} \qquad i = 1,...30, \quad j = 1,...,8, \tag{8.1}$$

where t_j is the centered time covariate (i.e. $t_0 = 2000 - 2003.5 = -3.5$, $t_1 = 2001 - 2003.5 = -2.5$ and so on), s is a common cubic regression spline, while b_{0i} and b_{1i} are i.i.d samples of the random variables $b_0 \sim \mathcal{N}(0, \sigma^2_{b_0})$ and $b_1 \sim \mathcal{N}(0, \sigma^2_{b_1})$ respectively, representing gaussian additive independent random effects, grouped by hospital. The quantities ε_{ij} are i.i.d. samples from the random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ representing residual error: ε, b_0 and b_1 are assumed to be independent. Estimates are obtained by maximization of restricted likelihood. Figure 8.2 shows the estimated growth curves together with the original data.

We fitted a semiparametric mixed effects model in order to catch a common behavior in the growth of normalized number of NON-STEMI diagnoses in the years 2000-2007, smoothing data and taking into account overdispersion due to the grouping factor. In fact, inspection of Figure 8.2 suggests a common "S-shaped" growing pattern. Concerning the random effects, the estimated parameters are: $\hat{\sigma}_{b_0} = 2.702 * 10^{-07}$, $\hat{\sigma}_{b_1} = 0.00765$ and $\hat{\sigma} = 0.02297$. The negligible effect of the random variable b_0 suggests that the curves are in fact different only with respect to their growth rate. The greater effect of the random variable b_1 is conducive to a further analysis of these data by means of a model that captures the common growth trend while taking into account overdispersion in the growth rates. Indeed, the following parametric logistic mixed effects model accommodates for the

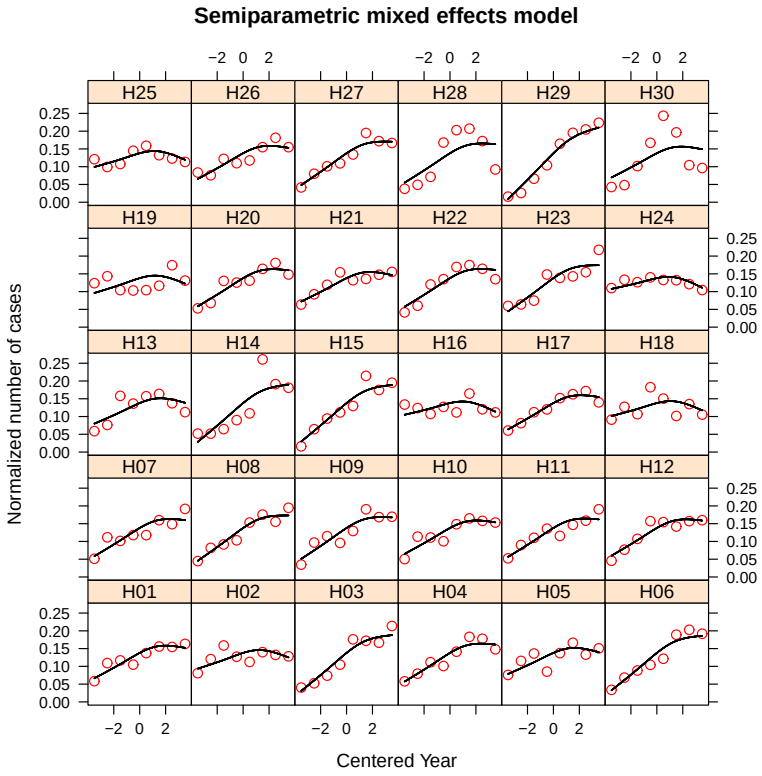


Figure 8.2: Estimated growth curves through model (8.1) together with the original data.

"S-shaped" common growing pattern, pointed out by the nonparametric analysis, while enabling the testing of its significance:

$$\tilde{N}_{ij} = \frac{\text{Asym} + \alpha_i}{(1 + exp(\text{Tmid} + \tau_i - t_j))} + \varepsilon_{ij}, \qquad i = 1, ...30, \quad j = 1, ..., 8, \tag{8.2}$$

where t_j is the centered time covariate, the fixed effects Asym and Tmid represent, respectively, the asymptote and the inflection point of the logistic curve, while α_i and τ_i are i.i.d samples of the random variables $\alpha \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\tau \sim \mathcal{N}(0, \sigma_\tau^2)$, respectively, representing gaussian additive random effects, grouped by hospital. The quantities ε_{ij} are i.i.d. samples from the random variable $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and they represent residual error. The two random effects α and τ are assumed to be independent, and independent of ε; all estimates are computed by restricted maximum likelihood. Table 8.1 shows that both fixed effects Asym and Tmid are significant.
Concerning the random effects, the estimated parameters are: $\hat{\sigma}_\alpha = 6.8183 * 10^{-07}$, $\hat{\sigma}_\tau = 0.4821$ and $\hat{\sigma} = 0.0287$. It is then confirmed that the variability of the additive random effect relative to the asymptote is negligible; thus α_i can be removed from model (8.2) without loss in model performance. On the contrary, the variability of the random effect relative to the inflection point is large and implies a very significant effect; this stimulates an interesting interpretation, since, in the logistic model, the inflection point indicates the time of maximum growth speed and this, in turn, is directly related to the timing of a growth speed significantly different from zero.

The inspection of the set of (estimated) random effects $\tau_i, i = 1, ..., 30$, related to the inflection point suggests a clustering structure that has been captured by partitioning the set in $k = 1, 2, ...,$

Fixed effects estimates:

	Value	Std. Error
Asym	0.1544	0.0026
Tmid	-2.7017	0.1368

Anova Table:

	numDF	denDF	F-value	p-value
Asym	1	209	5417.630	$< .0001$
Tmid	1	209	389.845	$< .0001$

Table 8.1: Fixed effects estimates and Anova table for model (8.2).

clusters by means of the Partitioning Around Medoids procedure (PAM, [90]), implemented with the euclidean distance, denoted by d. A critical point is the choice of k, the number of groups: an helpful method is the computation of the average silhouette width, and the inspection of the silhouette plot of PAM. For each estimated τ_i, let A be the cluster to which τ_i has been assigned and compute $a(\tau_i)$, the average dissimilarity of τ_i to all other objects in A,

$$a(\tau_i) = \frac{1}{|A|-1} \sum_{\tau_j \in A,\, \tau_j \neq \tau_i} d(\tau_j, \tau_i).$$

Now, if C is a cluster different from A, denote by

$$d(\tau_i, C) = \frac{1}{|C|-1} \sum_{\tau_j \in C} d(\tau_j, \tau_i)$$

the average dissimilarity of τ_i from all objects in C and set $c(\tau_i)$ to be the smallest value of $d(\tau_i, C)$ when C is let to range over the set of all clusters different from A. The *silhouette value* $s(\tau_i)$ of τ_i is defined as

$$s(\tau_i) = \frac{c(\tau_i) - a(\tau_i)}{\max\{a(\tau_i), c(\tau_i)\}}.$$

Clearly $s(\tau_i)$ lies between -1 and 1; large values of $s(\tau_i)$ support the fact that the element τ_i is well classified in A. The entire silhouette plot, i.e. the plot of all $s(\tau_i)$, and the Average Silhouette Width, i.e. the average of all silhouette values, are qualitative indexes helpful to judge and compare the results obtained by PAM for different values of k [136]. By inspecting the silhouette plot, represented in Figure 8.3, the presence of $k = 3$ clusters can be sustained. Indeed, for $k = 3$, the Average Silhouette Width is equal to 0.58 and, as a general rule, it can be asserted that a reasonable clustering structure has been found when the Average Silhouette Width is greater than 0.5. The medoids representative of the three clusters correspond to years $y_A = 2000, y_B = 2001$ and $y_C = 2002$. "Cluster A" denotes the institutions for which the estimated time of inflection point $\text{Tmid} + \tau_i$ in model (8.2) is closer to -3.1692, i.e. closer to year $y_A = 2000$. Analogously, "Cluster B" denotes the institutions for which the estimated time of inflection point is closer to -2.6839, i.e. closer to year $y_B = 2001$, and "Cluster C" denotes the institutions for which the estimated time of inflection point is closer to -2.3014, i.e. closer to year $y_C = 2002$.
In the left panel of Figure 8.4, the curves estimated by model (8.2) are represented, one curve for each hospital, together with the real data; the right panel shows the estimated logistic growth curves. The thick red, black and green curves represent the three benchmarks growth curves, i.e. medoids for cluster A, B and C, respectively.

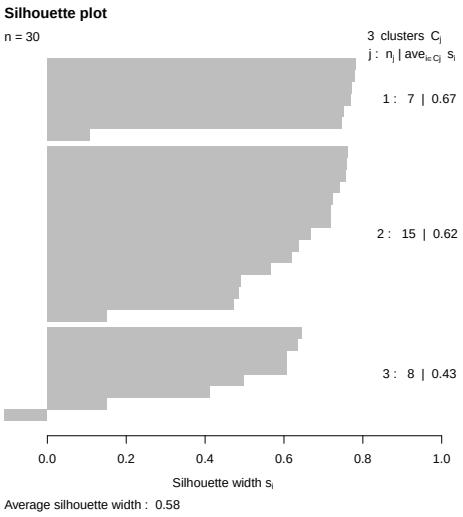


Figure 8.3: Silhouette plot of PAM procedure on the estimated inflection points with $k = 3$ clusters.

The particular interest in analyzing the clustering structure of the random effects related to the inflection points derives by the clinical surmise about their presence. Indeed, it is known that from the early 2000s the troponin exam has been introduced in hospital practices as a diagnostic device to better identify NON-STEMI events; hence, the presence of 3 clusters for the random effects τ_i could be a consequence of the different hospital timings in the introduction and adoption of this practice. This hypothesis cannot be validated directly since the timings of adoption of the troponin exam by the 30 different hospitals included in the analysis are not available.

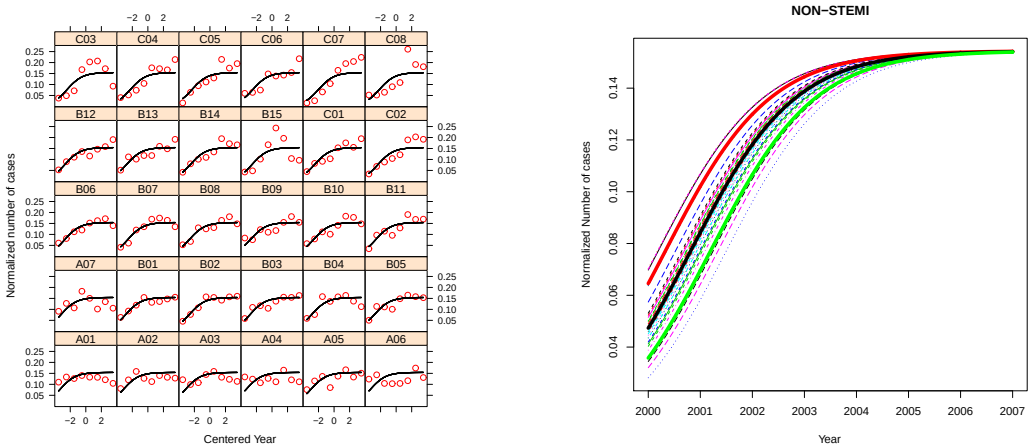


Figure 8.4: Estimated logistic growth curves for different medical institutions.

The previous analysis suggests a simpler model with fixed effects only, where dummy variables represent the identified cluster structure (clusters A, B, or C). This model is easier to interpret and communicate to clinicians; for instance, it quantifies the statistical evidence of the existence of groups in terms of p-values reported in Table 8.2. The model is:

$$\tilde{N}_{ij} = \frac{\text{Asym}}{(1+exp(\text{Tmid}_A \cdot 1_{i\in A} + \text{Tmid}_B \cdot 1_{i\in B} + \text{Tmid}_C \cdot 1_{i\in C} - t_j))} + \varepsilon_{ij}, \tag{8.3}$$

where $i = 1,...30$, is the institution index, $j = 1,...,8$, is the year index, and ε is defined as before. Estimates for the effects of model (8.3) appear in Table 8.2; they are all significant. Notice that the fixed effects estimates reported in Table 8.2 are close to the values identifying the inflection points of the three medoids y_A, y_B and y_C generated by the analysis of model (8.2).

	Value	Std. Error	p-value
Asym	0.1540	0.0025	$< .0001$
Tmid_A	-3.9434	0.2383	$< .0001$
Tmid_B	-2.6719	0.1294	$< .0001$
Tmid_C	-1.9108	0.1637	$< .0001$

Table 8.2: Fixed effects estimates for model (8.3).

Testing all possible contrasts between the three different fixed effects related to the inflection point, always generates a p-value less than 10^{-4}; there is a strong evidence of different inflection points in the three groups. Diagnostic checks show that normality assumption of residuals can be sustained.

In conclusion, the statistical analysis advocates the presence of three groups of hospitals, possibly distinguished by different timings of introduction and adoption of the troponin test and supports the clinical tenet that in the time period $2000-2007$ there has been an apparent increase in the normalized number of NON-STEMI diagnoses that is not due to a real increase in the disease incidence, but to a new diagnostic procedure adopted in hospitals along different timings.

This study represents an example of unsupervised clustering carried out starting from the random effects estimates in a NLME model setting, according to methods and purposes highlighted in Paragraph 4.2.3.

8.2 Nonlinear nonparametric models for an epidemiologic enquire

In this Section, the analysis related to the theoretical framework proposed in Paragraph 4.6.1 are presented. Simulated dataset, analyses and results are detailed in [12].

8.2.1 Linear growth model

Starting from the simulation study for linear models, let g in model (4.44) is linear. The general model, for $i = 1,\ldots,N$, include three different cases, that are:

$$\mathbf{y}_i = \begin{cases} \alpha + d_i\mathbf{t} + \boldsymbol{\varepsilon}_i & \text{(random-slope case)} \\ a_i + \delta\mathbf{t} + \boldsymbol{\varepsilon}_i & \text{(random-intercept case)} \\ a_i + d_i\mathbf{t} + \boldsymbol{\varepsilon}_i & \text{(fully random case)} \end{cases}$$

where $\boldsymbol{\varepsilon}_i$ are i.i.d. from $\mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}_n)$ and $\mathbf{t}$ is the vector of sampling times. Intercept and slope are treated as fixed or random effects according to the different cases. In the fully random case, both slope and intercept parameters are considered random, i.e. $\mathbf{b}_i = (d_i, a_i)$, whereas in the random-slope and random intercept case, $b_i = d_i$ and $b_i = a_i$ respectively. The interest is focused on random effects estimation, because our main goal is to test the performance of our algorithm in identifying the correct number of groups in simulated data and in estimating properly location and weights

of different groups. Testing the linear case enables us to compare results of our algorithm with those carried out by the R algorithm npmlreg, which implements [3] procedure of non parametric random effect estimation. To be noticed is that our method is not efficient in the linear case, since it doesn't take advantage of the linearity of the problem. However it doesn't need any a priori specification of the number of support points of the random effects. Even if we don't specify the exact number of groups beforehand, the proposed method is able of carrying out a good estimation of the random effects distribution.

Some examples of simulated data from a linear growth model are shown in left panels of Figure 8.5. We simulated 8 datasets of linear curves grouped in a number of clusters that vary form 2 to 10. Different values of the error variance σ^2 have been chosen for each test set, in order to obtain noisy observations for each curve. Datasets addressed with the name "S" contain groups in which only slopes is random, "I" datasets contain groups where only intercept is random and "SI" datasets contain curves where both slope and intercept are random. The simulated datasets are then:

- lin2S: 2 balanced groups, each one composed by 25 curves, with the same intercept (equal to 4), 2 different slopes ($\mathbf{c} = (c_1, c_2) = (1, 2)$) and $\sigma = 1$;
- lin2I: 2 balanced groups, each one composed by 25 curves with the same slope (equal to 1), 2 different intercept ($\mathbf{c} = (c_1, c_2) = (3, 10)$) and $\sigma = 0.65$;
- lin4SI: 4 balanced groups, each one composed by 25 curves, where location points $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4)$ are obtained from all possible combinations of 2 different slopes (equal to 1 and 3) and 2 different intercepts (equal to 40 and 60), i.e., $\mathbf{c}_1 = (1, 40)$, $\mathbf{c}_2 = (1, 60)$, $\mathbf{c}_3 = (3, 40)$ and $\mathbf{c}_4 = (3, 60)$ with $\sigma = 1$;
- lin3S: 3 unbalanced groups, composed by 24, 24 and 2 curves respectively, with the same intercept (equal to 4), 3 different slopes ($\mathbf{c} = (c_1, c_2, c_3) = (1, 2, 3.5)$) and $\sigma = 1$;
- lin3I: 3 unbalanced groups, composed by 24, 24 and 2 curves respectively, with the same slope (equal to 1), 3 different intercepts ($\mathbf{c} = (c_1, c_2, c_3) = (2, 7, 14)$) and $\sigma = 1$;
- lin9SI: 9 unbalanced groups, 6 of whom containing 24 curves and 3 containing 2 curves, where location points $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9)$ are obtained from all possible combinations of 3 different slopes (equal to 1, 4 and 7) and 3 different intercept (equal to 20, 35 and 60) with $\sigma = 1.5$;
- lin10S: 10 balanced groups, each one composed by 50 curves with the same intercept (equal to 1), 10 different slopes ($\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}) = (0.5, 2, 4, 5.5, 7.5, 10, 12, 13.5, 16, 20)$) and $\sigma = 1.5$;
- lin10I: 10 balanced groups, each one composed by 15 curves with the same slope (equal to 1), 10 different intercepts ($\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}) = (1, 5, 10, 15, 20, 25, 30, 35, 40, 45)$) and $\sigma = 1$.

All these datasets represent typical situations in which fitting a parametric mixed effects model could be wrong because random effects are not normally distributed. On these datasets, we fitted models with both the NLNPEM method and the nonparametric maximum likelihood approach introduced in [3].

The method introduced in [3] is a method for fitting overdispersed generalized linear models: the idea is to approximate the unknown and unspecified distribution of the random effects by a discrete

mixture of densities from exponential family. This approximation leads to a simple expression of the marginal likelihood that can be maximized using a standard EM algorithm. Once specified the model and the number of random effects groups k, the `R` package `npmlreg` fits a linear mixed effects model using nonparametric maximum likelihood. Since we are testing the proposed method in a simulation setting, when `npmlreg` method is used we provide the correct number of groups, whereas, when NLNPEM is used, we don't have to. The M starting points for random effects distribution are randomly chosen in a proper range and the starting fixed effects are estimated through linear least squares. Finally, the tolerance D is set equal to 0.05 and $\tilde{\omega} = 0.05$. According to the dimension of the random effect ($q = 1$ for random-slope or random-intercept case, $q = 2$ when both effects are random), we properly define the model in `npmlreg` and NLNPEM algorithms.

Notice that `npmlreg` does not allow to select one dimensional random effect for slope only but provides a random effects estimation for both intercept and slope parameters. In this case, in order to correctly compare the two methods, we have set also in the NLNPEM method both slope and intercept to be random in the random-slope case. Of course, in the NLNPEM method, random effects only for the slope may be selected by the user, if necessary.

In Tables 8.6, 8.7 and 8.8 of Paragraph 8.2.5, results of `npmlreg` and NLNPEM algorithms for three representative cases are compared, i.e the estimations of random effects in terms of points and weights are reported and compared with the corresponding true distributions. Observing estimated values reported in these Tables, it can be argued that both methods estimate well both the discrete random components of the model and the fixed effects when a small number of groups is considered. Increasing the number of groups, the two algorithms show different behaviors. In particular we notice that, for large number of groups, `npmlreg` misses some points of the nonparametric distribution, whereas NLNPEM performs better, even ignoring the true number of groups. The number of groups estimated by the NLNPEM algorithm depends in general on the tuning tolerances D and $\tilde{\omega}$, introduced in Paragraph 4.6.1. This algorithm tends to overestimate the number of points of the discrete distribution. However, even if the number of points is greater than the real number, the points tend to cluster near the true ones. Moreover, summing the weights of the points in each cluster, we obtain results similar to the exact weights. The hints concerning the number of groups provided by NLNPEM algorithm make this method a powerful tool in explorative analyses within an unsupervised framework. The NLNPEM method is also capable of detecting outlier groups, whereas the `npmlreg` method is able to detect them only in presence of small number of groups. In general, we notice that sometimes `npmlreg` method performs poorly in estimation or even misses convergence, whereas NLNPEM doesn't. These situations happen in particular when there are 9 different groups both for intercept and slope ("lin9SI" dataset) and when there are 10 groups for slope or intercept("lin10S" and "lin10I" dataset respectively).

In order to resume the goodness of fit of NLNPEM method and the `npmlreg` one, we finally compare the normalized Wasserstein distances between the true discrete random effects distribution and the estimated one through the two methods, for each simulated set of linear curves. Results are reported in Table 8.3, together with the goodness of fit index $-2\log L$.
To be noticed is that, in the case of Wasserstein distance, results are similar for all datasets where both algorithm perform well. On the other hand, significant differences exist in cases with large number of groups, where NLNPEM performances are much better both in terms of Wasserstein distance and $-2\log L$.

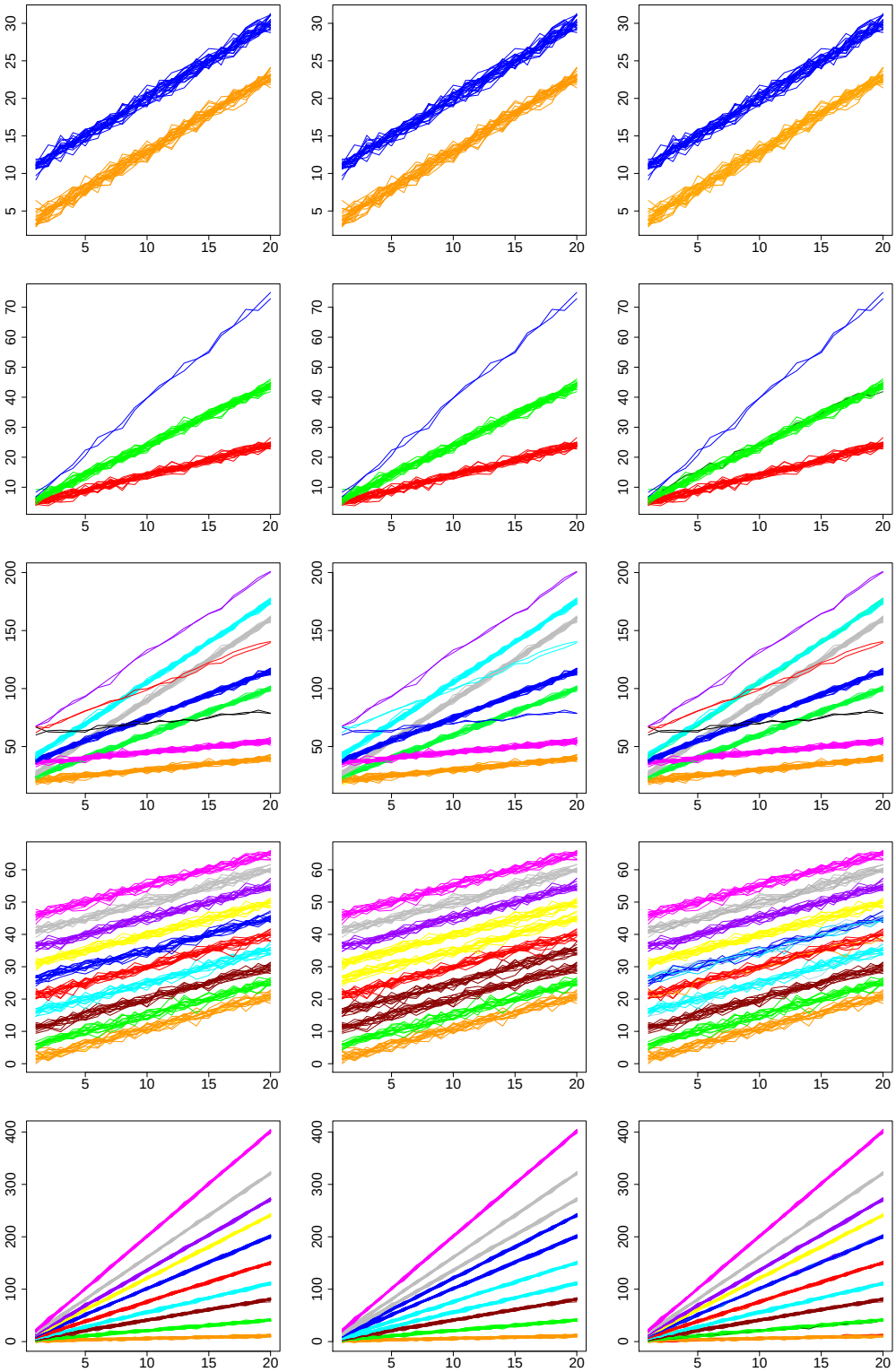


Figure 8.5: Simulated data (left panels), `npmlreg` (central panels) and NLNPEM classification (right panels) in lin2I, lin3S, lin9SI, lin10I and lin10S datasets respectively. Different colors are used to represent real groups (left panels), groups identified by `npmlreg` and NLNPEM methods (central and right panels respectively).

Model	Wasserstein distance		$-2\log L$	
	`npmlreg`	NLNPEM	`npmlreg`	NLNPEM
lin2S	0.013572	0.013724	2861.2	942.0
lin2I	0.004538	0.005187	2097.7	190.5
lin4SI	0.008121	0.006298	5974.4	2017.7
lin3S	0.003041	0.004651	2839.8	912.7
lin3I	0.003454	0.003454	2938.3	1017.2
lin9SI	0.017756	0.001565	16127.0	5376.7
lin10S	0.033632	0.000410	76716.1	18025.9
lin10I	0.023045	0.001649	12795.8	2947.3

Table 8.3: Normalized Wasserstein distances and $-2\log L$ index for `npmlreg` and NLNPEM algorithm respectively in the simulated linear cases.

In the following we describe two nonlinear case studies: the exponential and the logistic growth model. These two models are among the most used in nonlinear mixed effects framework because they find application in several areas like pharmacokinetics and epidemiological studies.
Since other nonlinear nonparametric methods are not available for free software, we are not able to compare the NLNPEM results with those obtained with other methods; for this reason we will only test NLNPEM performances, providing the normalized Wasserstein distance between the true distribution and the estimated one.

8.2.2 Exponential growth model

We first describe the exponential case, in which we consider the following nonlinear function in model (4.44):

$$g(t) = \alpha\left(1 - e^{-\lambda t}\right)$$

which is nonlinear in λ. The two parameters α and λ represent respectively the asymptote and the growth rate. In this case study we consider only random effects for the asymptote, that means that the mixed effects model becomes

$$\mathbf{y}_i = a_i\left(1 - e^{-\lambda \mathbf{t}}\right) + \boldsymbol{\varepsilon}_i$$

where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ are i.i.d. errors, a_i are the random effects for the asymptote ($b_i = a_i$) and λ is the fixed effect for the growth rate ($\beta = \lambda$).

We simulated 3 datasets of exponential growth curves where only asymptote varies and is considered as random. All datasets are then addressed with the name "A". They are:

- exp2A: 2 balanced groups, each one composed by 25 curves, with the same growth rate ($\lambda = 0.5$), 2 different asymptotes ($\mathbf{c} = (c_1, c_2) = (1, 1.5)$) and $\sigma = 0.04$;
- exp3A: 3 unbalanced groups of 24, 24 and 2 curves respectively, with the same growth rate ($\lambda = 0.5$), 3 different asymptotes ($\mathbf{c} = (c_1, c_2, c_3) = (1, 1.5, 2.3)$) and $\sigma = 0.04$;
- exp10A: 10 balanced groups, each one composed by 5 curves, with the same growth rate ($\lambda = 0.5$), 10 different asymptotes ($\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}) = (1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25)$ and $\sigma = 0.04$.

The starting random effects distribution has M support points, randomly chosen in a proper range, and the starting fixed effects are estimated through nonlinear least squares. The tuning tolerance parameter D is set equal to 0.01 and $\tilde{\omega} = 0.05$. Figure 8.6 shows original datasets, where each curve is colored according to the group estimated by NLNPEM method.

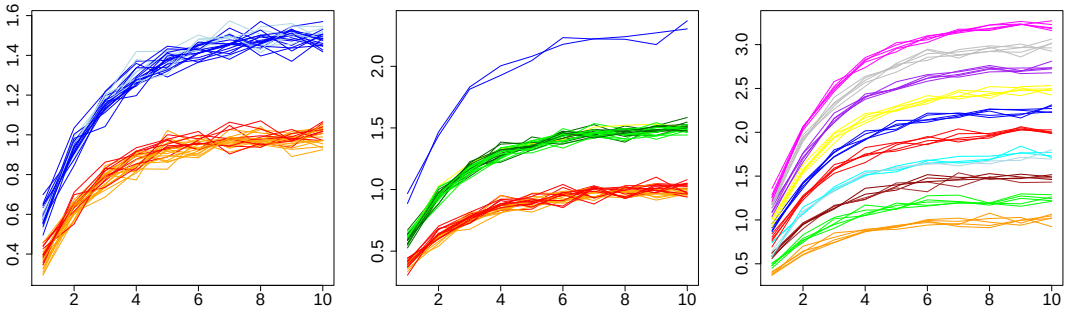


Figure 8.6: NLNPEM classification in exp2A, exp3A and exp10A datasets respectively with exponential model.

The estimated number of groups is larger than the real one in all the three cases; however the estimated random effects create the right number of clusters located close to the correct points. In the exp3A case the NLNPEM method is also able to identify the *outlier* group estimating well the location and the weight of the random effects. The performance of NLNPEM method is evaluated in this case only in terms of normalized Wasserstein distance, shown in Table 8.4.

Model	Wasserstein distance
exp2A	0.030048
exp3A	0.015025
exp10A	0.011524

Table 8.4: Normalized Wasserstein distances for NLNPEM algorithm in the simulated exponential cases.

8.2.3 Logistic growth model

The second nonlinear model tested is the logistic growth model. In this case the nonlinear function is:

$$g(t) = \frac{\alpha}{1 + e^{-\frac{t-\delta}{\gamma}}}$$

where α represent the asymptote, δ is the inflection point, which correspond to the time at which the growth curve reaches the half of the asymptote, and γ is the time elapsed between δ and the time at which the growth curve reaches 3/4 of the asymptote level. The parameter γ will always be treated as a fixed effect while the asymptote and the inflection point will be treated either as fixed or as random effect according to different cases. The general model, which is nonlinear in λ and γ, include three different cases:

$$\mathbf{y}_i = \begin{cases} \dfrac{a_i}{1+e^{-\frac{\mathbf{t}-\delta}{\gamma}}} + \boldsymbol{\varepsilon}_i & \text{(random-asymptote case)} \\ \dfrac{\alpha}{1+e^{-\frac{\mathbf{t}-d_i}{\gamma}}} + \boldsymbol{\varepsilon}_i & \text{(random-inflection case)} \\ \dfrac{a_i}{1+e^{-\frac{\mathbf{t}-d_i}{\gamma}}} + \boldsymbol{\varepsilon}_i & \text{(random-asymptote and inflection case)} \end{cases} \tag{8.4}$$

where $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ are i.i.d. errors, a_i and d_i represent the random effects for the asymptote and the inflection point, while α, δ and γ represent the fixed effects. In particular in the varying asymptote case $b_i = a_i$ and $\beta = (\delta, \gamma)$, in the varying inflection case $b_i = d_i$ and $\beta = (\alpha, \gamma)$ and in the varying asymptote and inflection case $\mathbf{b}_i = (a_i, d_i)$ and $\beta = \gamma$.

We simulated 8 datasets of logistic growth curves that include all the cases resumed in (8.4). Each dataset is composed by a different number of balanced or unbalanced groups (from 2 to 10 clusters) similar to those presented in the linear framework. Datasets addressed with the name "A" represent random asymptote cases, "I" datasets contain groups where only inflection point is random and "AI" ones contain curves where both asymptote and inflection point are random. We then have:

- logis2A: 2 balanced groups, each one composed by 25 curves, with $\delta = 6$, $\gamma = 1$, 2 different asymptotes ($\mathbf{c} = (c_1, c_2) = (1, 2)$) and $\sigma = 0.04$;
- logis2I: 2 balanced groups, each one composed by 25 curves, with $\alpha = 1$, $\gamma = 1$, 2 different inflection points ($\mathbf{c} = (c_1, c_2) = (6, 8)$) and $\sigma = 0.04$;
- logis4AI: 4 balanced groups, each one composed by 25 curves, where location points $\mathbf{c} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4)$ are obtained from all possible combinations of 2 different asymptotes (equal to 1 and 2) and 2 different inflection points (equal to 6 and 10), i.e., $\mathbf{c}_1 = (1, 6)$, $\mathbf{c}_1 = (1, 10)$, $\mathbf{c}_1 = (2, 6)$ and $\mathbf{c}_4 = (2, 10)$ with $\gamma = 1$ and $\sigma = 0.04$;
- logis3A: 3 unbalanced groups of 24, 24 and 2 curves respectively, with $\delta = 6$, $\gamma = 1$, 3 different asymptotes ($\mathbf{c} = (c_1, c_2, c_3) = (1, 2, 3.5)$) and $\sigma = 0.04$;
- logis3I: 3 unbalanced groups of 24, 24 and 2 curves respectively, with $\alpha = 1$, $\gamma = 1$, 3 different inflection points ($\mathbf{c} = (c_1, c_2, c_3) = (6, 8, 11.5)$) and $\sigma = 0.04$;
- logis9AI: 9 unbalanced groups of curves (6 of whom containing 24 curves and 3 containing 2 curves), where location points $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9)$ are obtained from all possible combinations of 3 different asymptotes (equal to 1, 2 and 4) and 3 different inflection points (equal to 6, 8 and 11.5) with $\gamma = 1$ and $\sigma = 0.04$;
- logis10A: 10 balanced groups, each one composed by 5 curves, with $\delta = 6$, $\gamma = 1$, 10 different asymptotes ($\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}) = (1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25)$ and $\sigma = 0.04$;
- logis10I: 10 balanced groups, each one composed by 5 curves, with $\alpha = 1$, $\gamma = 1$, 10 different inflection points ($\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}) = (4.5, 5.5, 7, 8, 9.5, 10.5, 12, 13, 14.5, 16)$ and $\sigma = 0.04$.

Since the NLNPEM method is able to fit all three models resumed in (8.4), we fit the right model for each dataset. The starting random effects distribution has M support points, randomly chosen in

a proper range, and the starting fixed effects are estimated through nonlinear least squares. We set the tolerance D equal to 0.05 and $\tilde{\omega} = 0.05$.

Figure 8.7 shows original datasets, where each curve is colored according to the group estimated by NLNPEM method. We notice in Figure 8.7 that, even if we don't specify a priori the correct number of groups, we are able to cluster correctly the subjects both when there are few groups and when there are many. The method is also able to capture correctly *outliers* groups; in all the unbalanced cases the proposed method recognize the *outliers* groups and estimate well both the location and the weight of random effects.
In order to test the NLNPEM method we can compare these results with those obtained considering always both asymptote and inflection point as random effects. For the two varying asymptote and inflection cases we have obviously fitted only the model with two random effects.

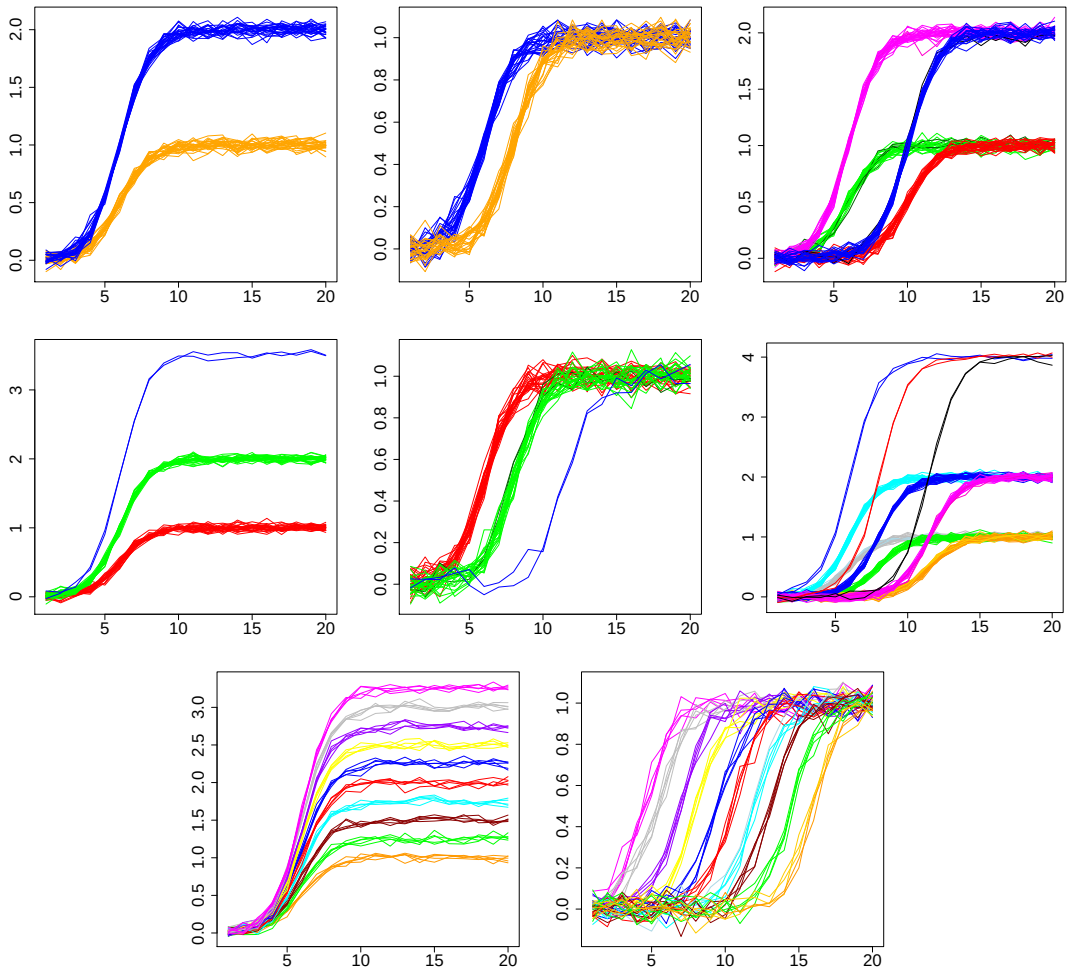


Figure 8.7: NLNPEM classification in logis2A, logis2I, logis4AI, logis3A, logis3I, logis9AI, logis10A and logis10I datasets respectively with logistic growth model.

The normalized Wasserstein distances are shown in Table 8.5; the left column represents the normalized Wasserstein distance for a model with one random effect while the right one represent the same distance for models with two random effects.

Model	Wasserstein distance	
	$q=1$	$q=2$
logis2A	0.000150	0.000450
logis2I	0.003202	0.012171
logis4AI	–	0.004869
logis3A	0.000396	0.000629
logis3I	0.007243	0.010250
logis9AI	–	0.006477
logis10A	0.001286	0.0.0015
logis10I	0.004664	0.005207

Table 8.5: Normalized Wasserstein distances for NLNPEM algorithm in the simulated logistic cases.

We first notice that the normalized Wasserstein distances are always very low, that means that the NLNPEM method is able to estimate well both random and fixed effects even in presence of a high number of groups. We also notice, comparing results for the same case study with one or two random effects, that the normalized Wasserstein distances are always very close together. Both observing the Wasserstein distances and the fitted curves obtained with the model with two random effects, we notice that in the NLNPEM method we are allowed to consider more parameters as random effects than needed, without damaging the parameter estimation. In particular this approach could be useful when we don't know which are the parameters to be considered random. For this purpose we could perform a first analysis considering all parameters as random effects and then fit a second model fixing the parameters that show a very low variability. This approach could be performed with the NLNPEM method because it can handle both random and fixed effects whereas other previous methods cannot.

8.2.4 Application to NON STEMI data

In this example, we study a dataset concerning Hospital Discharges of patients affected by Acute Myocardial Infarction (AMI) without ST-segment elevation (NON-STEMI), a dataset arising from the Administrative database *RICOVERI* described in Paragraph 3.1.1. It contains the same data that have been already studied in [77] through Nonlinear Parametric Mixed Effects Models and that have been detailed in Section 8.1. Figure 8.8 represents the normalized number of NON-STEMI diagnoses along the time period 2000-2007 grouped by hospital and relative to the 30 largest clinical institutions of Regione Lombardia. For each hospital the yearly number of diagnoses has been standardized by the hospital total number of diagnoses in the time period 2000-2007.

As pointed out in Section 8.1, the random-inflection case in model (8.4) seems to capture the common "S-shaped" growing pattern. The NLNPEM algorithm clusters the hospitals in $N=2$ different groups, according to the estimated discrete distribution of the random effect for the inflection point (see Figure 8.8). The estimated fixed effects are $\hat{\alpha}=0.16$ and $\hat{\gamma}=1.31$, the estimated discrete measure $\hat{\mathscr{P}}^*$ is concentrated on $(\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2)=(-3.76,-2.43)$ with weights $(\hat{\omega}_1,\hat{\omega}_2)=(0.2,0.8)$ and the estimated variance is $\hat{\sigma}^2=7.7\cdot10^{-4}$. This analysis, performed with $D=0.05$ and $\tilde{\omega}=0.05$, backs up the presence of two groups of hospitals according to different inflection points and automatically detects an unsupervised cluster structure. Even if clinical best practice maintains that there is no evidence for a greater incidence of NON-STEMI in this period it is known that since the early 2000s a new diagnostic procedure - the *troponin* exam - has been introduced and this could have produced an increased number of positive diagnoses, by easing NON-STEMI detection.

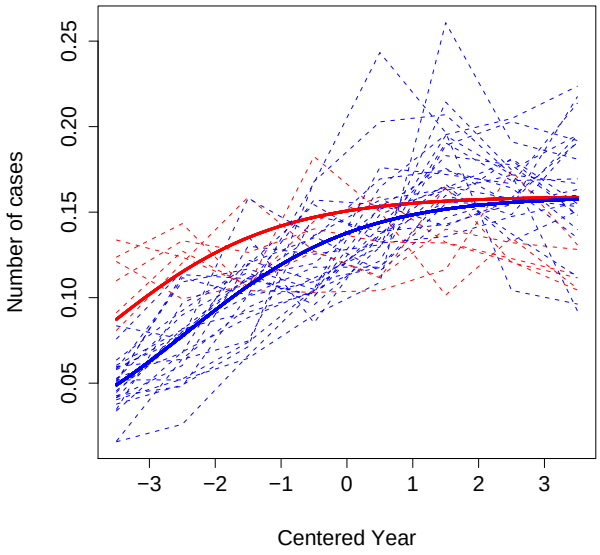


Figure 8.8: Standardized number of AMI without ST-segment elevation diagnoses in the period 2000 - 2007 in the 30 largest clinical institutions of Lombardia Region. The year has been centered and normalization has been carried out standardizing the yearly number of diagnoses for each hospital by total number of diagnoses in the time window 2000 – 2007. Real data are colored according to the NLNPEM clusters and NLNPEM fitted models are superimposed.

Hence, the presence of 2 clusters could be a consequence of the different hospital timings in the introduction and adoption of this practice. This hypothesis cannot be validated directly since the timings of adoption of the troponin exam by the 30 different hospitals included in the analysis are not available. The agreement with previous results of Section 8.1 together with the great advantage of a non-parametric approach advocates the real profit in using this new estimation algorithm.

8.2.5 Comparison of results

Comparison of estimates carried out by `npmlreg` and NLNPEM method are reported here, for some cases of interest mentioned before.

- Linear case - Random-intercept case (lin2I)

effects		True	`npmlreg`	NLNPEM
fixed	slope	1	1.0021	1.0022
random	intercept 1	3	2.9382	2.9368
	(weight 1)	(0.5)	(0.5)	(0.5)
	intercept 2	10	10.0150	10.0136
	(weight 2)	(0.5)	(0.5)	(0.5)

Table 8.6: Estimates carried out by `npmlreg` and NLNPEM method on lin2I dataset, where intercept is considered as random, with 2 balanced groups.

- Linear case - Random-slope case (lin3S)

effects		True	npmlreg	NLNPEM
random	slope 1	1	1.0107	1.0107
	(weight 1)	(0.48)	(0.48)	(0.48)
	slope 2	2	1.9982	2.0030 1.9637
	(weight 2)	(0.48)	(0.48)	(0.4214) (0.0585)
	slope 3	3.5	3.5250	3.5250
	(weight 3)	(0.04)	(0.04)	(0.04)
random	intercept	4	3.9326	3.9326
	intercept	4	4.0751	3.9954 4.6502
	intercept	4	3.3717	3.7174

Table 8.7: Estimates carried out by npmlreg and NLNPEM method on lin3S dataset, where slope is considered as random, with 3 unbalanced groups.

- Linear case - Random-intercept case (lin10I)

effects		True	npmlreg	NLNPEM
fixed	slope	1	1.001857	1.001232
random	intercept 1	1	0.9114 0.9114	0.9185
	(weight 1)	(0.1)	(0.00050) (0.09949)	(0.1)
	intercept 2	5	5.0257	5.0328
	(weight 2)	(0.1)	(0.1)	(0.1)
	intercept 3	10	-	10.048
	(weight 3)	(0.1)	-	(0.1)
	intercept 4	15	12.5442	14.8397 15.1058
	(weight 4)	(0.1)	(0.2)	(0.0192) (0.0807)
	intercept 5	20	19.9818	19.9312 20.0026
	(weight 5)	(0.1)	(0.1)	(0.0368) (0.0631)
	intercept 6	25	27.4750	24.9215 25.1181 25.1975
	(weight 6)	(0.1)	(0.2)	(0.0325) (0.0371) (0.0302)
	intercept 7	30	-	29.886
	(weight 7)	(0.1)	-	(0.1)
	intercept 8	35	35.0050	34.9582 35.2459
	(weight 8)	(0.1)	(0.1)	(0.0813) (0.0186)
	intercept 9	40	39.9516	39.6837 39.9624 40.4505
	(weight 9)	(0.1)	(0.1)	(0.0186) (0.0714) (0.0098)
	intercept 10	45	45.0017 45.0017	45.008
	(weight 10)	(0.1)	(0.09949) (0.000507)	(0.1)

Table 8.8: Estimates carried out by npmlreg and NLNPEM method on lin10I dataset, where intercept is considered as random, with 10 balanced groups.

8.3 Bayesian decision rules for provider profiling in cardiovascular context

In this section, we develop Bayes rules for several families of loss functions for hospital report cards under a Bayesian semiparametric hierarchical model like those proposed in Section 5.4. Moreover, we present some robustness analysis with respect to the choice of the loss function, focusing on the number of hospitals our procedure identifies as "unacceptably performing". The analysis is carried out on a case study dataset arising from MOMI^2 survey (see Paragraph 3.2.2) on patients admitted with ST-Elevation Myocardial Infarction to the hospitals of Milan Cardiological Network. The major aim of this analysis is the ranking of the health care providers performances, together with the assessment of the role of patients' and providers' characteristics on survival outcome.

Performance indicators have recently received increasing attention; they are mainly used with the aim of assessing quality in health care research (see [9], [10], [62], [115], [116], [117], [125]). In this analysis, the aim is to point out similar behaviours among groups of hospitals and then classify them according to some acceptability criteria, suitably modelling the survival outcome of patients affected by a specific disease and admitted to different clinical structures. In general, provider profiling of health care structures is obtained producing report cards comparing their global outcomes or performances of their doctors. These cards have mainly two goals:

- to provide information that can help individual consumers (i.e., patients) making a decision,
- to identify hospitals that require investments in quality improvement initiatives.

Here we are interested not only in the point estimation of the mortality rate, but also to decide whether investing in quality improvement initiatives for each hospital with "unacceptable performances", as explained and implemented in [65].

8.3.1 Statistical support to decision-making in health-care policy

Even in a perfect risk-adjustment framework, random errors will be present. Therefore, when classifying hospital performances as "acceptable" or not, some mistakes could occur, so that some hospitals could be misclassified. Anyway, different players in the health care context would pay different costs on misclassification errors. By *False Positive* we mean the hospital that truly had acceptable performances but was classified as "unacceptably performing", and by *False Negative* the hospital that truly had unacceptable performances but was classified as "acceptably performing". Then a health care consumer would be presumably willing to pay a higher charge for decisions that minimize false negatives, whereas hospitals might pay a higher cost for information that minimizes false positives. On the other hand, the same argument could be used to target hospitals for quality improvement: false positives would yield unneeded investments in quality improvement, but false negatives would lead to loose opportunities in improving the hospital quality. According to its plans, any health care government could be interested in minimizing false positives and/or false negatives.

In order to provide support to decision-making in this context, we carry out the statistical analysis in the following way: firstly we estimate the in-hospital survival rates after fitting a Bayesian semiparametric generalized linear mixed effects model like in Section 5.4, in particular modelling the random effect parameters via a Dirichlet process (Paragraph 5.3.1); then we develop Bayes decision rules in order to minimize the expected loss arising from misclassification errors, comparing four different loss functions for hospital report cards (Section 5.4).

The Bayesian generalized mixed effects model for binary data we fitted for unit (patient) $i = 1,\dots,n_j$, in group (hospital) $j = 1,\dots,J$ is the following: let Y_{ij} be a Bernoulli random variable with mean p_{ij}, i.e.,

$$Y_{ij}|p_{ij} \overset{ind}{\sim} Be(p_{ij}).$$

The p_{ij}s are modelled through a logit regression of the form

$$logit(p_{ij}) = \log\frac{p_{ij}}{1-p_{ij}} = \theta_0 + \sum_{h=1}^{p}\theta_h x_{ijh} + \sum_{l=1}^{J} b_l z_{jl} \tag{8.5}$$

where $z_{jl} = 1$ if $j = l$ and 0 otherwise. In this model, $\boldsymbol{\theta} = (\theta_0,\dots,\theta_p)$ represents the $(p+1)$-dimensional vector of the fixed effects, $\mathbf{x}_{ij}$ is the vector of patient covariates and $\mathbf{b} = (b_1,\dots,b_J)$ is the vector of the additive random effects parameters of the grouping factor. According to [91], we assume a nonparametric prior for $b_1,\dots,b_J$, namely the b_js will be i.i.d. according a Dirichlet process, to include robustness to miss-specification of the prior at this stage, since it is known that the regression parameters can be sensitive to the standard assumption of normality of the random effects; the prior for $\boldsymbol{\theta}$ is parametric. Model (8.5) is a generalized linear mixed model with $p+1$ regression coefficients and one random effect. In [67] the same model was fitted on a different dataset to classify hospitals taking advantage of the in-built clustering property of the Dirichlet process prior. Here we use Bayesian estimates to address a new decision problem concerning hospitals' performances.

Bayesian inferences are based on the posterior distribution, i.e., the conditional distribution of the parameters vector, given the data. Once the posterior distribution has been computed, suitable loss functions can be defined in order to *a posteriori* weigh the decision of wrongly classifying the hospital as having acceptable or unacceptable performances. The random intercepts of model (8.5), i.e., $\theta_0+b_1, \theta_0+b_2,\dots,\theta_0+b_J$ represent the hospital performances quantifying the contribution to the model after patients' covariates adjustment. Let us denote by β_j the sum of θ_0 and b_j. The class of loss functions we are going to assume is then

$$L(\beta_j,d) = c_I \cdot f_1(\beta_j)\cdot d\cdot \mathbb{I}(\beta_j > \beta_t) + c_{II}\cdot f_2(\beta_j)\cdot(1-d)\cdot\mathbb{I}(\beta_j<\beta_t), \tag{8.6}$$

where d is the decision to take ($d=1$ means that the hospital has "unacceptable performances", $d=0$ stand for "acceptable performances"), c_I is the weight assigned to the cost $f_1(\beta_j)$, occurring for a false positive, c_{II} is the weight assigned to cost $f_2(\beta_j)$, occurring for a false negative and β_t is defined as $\log(p_t/(1-p_t))$, being p_t a reference value for survival probabilities.

Without loss of generality, we can assume a proportional penalization, i.e., $f_2(\beta_j) = k\cdot f_1(\beta_j)$, taking k as the ratio c_{II}/c_I. In this sense, the parameter k quantifies our beliefs on cost, being greater than 1 if we credit that accepting a *false negative* should cost more than rejecting a true negative and less than 1 otherwise. An acceptable performance is then defined comparing the posterior expected losses associated with the decision that the hospital had "acceptable performances"

$$R(\mathbf{y},d=0) = E_\pi\left(L(\beta_j,d=0)|\mathbf{y}\right) = \int f_2(\beta_j)\mathbb{I}(\beta_j<\beta_t)\Pi(\beta_j|\mathbf{y})d\beta_j$$

and the decision that the hospital had "unacceptable performances"

$$R(\mathbf{y},d=1) = E_\pi\left(L(\beta_j,d=1)|\mathbf{y}\right) = \int f_1(\beta_j)\mathbb{I}(\beta_j>\beta_t)\Pi(\beta_j|\mathbf{y})d\beta_j.$$

Here $\Pi(\beta_j|\mathbf{y})d\beta_j$ denotes the posterior distribution of β_j. In short, we classify an hospital as being "acceptable" (or with "acceptable performances") if the risk associated with the decision $d=0$ is less than the risk associated with the decision $d=1$, i.e., if $R(\mathbf{y},d=0) < R(\mathbf{y},d=1)$.

Within this setting, four different loss functions of the form (8.6) will be considered in the next paragraph, to address the decision problem, namely

$$\begin{aligned}
\textbf{0/1 Loss} &: \\
L(\beta_j, d) &= d \cdot \mathbb{I}(\beta_j > \beta_t) + k \cdot (1-d) \cdot \mathbb{I}(\beta_j < \beta_t), \\
\textbf{Absolute Loss} &: \\
L(\beta_j, d) &= |\beta_j - \beta_t| \cdot d \cdot \mathbb{I}(\beta_j > \beta_t) + k \cdot |\beta_j - \beta_t| \cdot (1-d) \cdot \mathbb{I}(\beta_j < \beta_t), \\
\textbf{Squared Loss} &: \\
L(\beta_j, d) &= (\beta_j - \beta_t)^2 \cdot d \cdot \mathbb{I}(\beta_j > \beta_t) + k \cdot (\beta_j - \beta_t)^2 \cdot (1-d) \cdot \mathbb{I}(\beta_j < \beta_t), \\
\textbf{LINEX Loss} &: \\
L(\beta_j, d) &= l(\beta_j - \beta_t) \cdot d \cdot \mathbb{I}(\beta_j > \beta_t) + k \cdot l(\beta_j - \beta_t) \cdot (1-d) \cdot \mathbb{I}(\beta_j < \beta_t).
\end{aligned}$$

For instance, this means that, to recover the 0/1 loss function above, the functions $f_i(\beta_j),\ i = 1,2$ in (8.6) are both constant, $f_i(\beta_j) = |\beta_j - \beta_t|, i = 1,2$ for the Absolute Loss case, $f_i(\beta_j) = (\beta_j - \beta_t)^2,\ i = 1,2$ for the Squared Loss case and $f_i(\beta_j) = l(\beta_j - \beta_t) = \exp\{a \cdot (\beta_j - \beta_t)\} - a \cdot (\beta_j - \beta_t) - 1,\ i = 1,2$ to obtain the LINEX Loss function. Note that all the loss functions but the last one are symmetric, and the parameter k is used to introduce an asymmetry in weighting the misclassification error costs.

8.3.2 Application to MOMI^2 data

In this section we apply the model and the method proposed in the previous paragraph to 536 patients of MOMI^2 data underwent to primary PCI treatment. For this sample, 17 hospitals of admission are involved, and a in-hospital survival rate of 95% is observed. Among all possible covariates (mode of admission, clinical appearance, demographic features, time process indicators, hospital organization etc.) available in the survey, only age and Killip class have been selected as being statistically significant. The killip class is binary here, i.e., the killip covariate is equal to 1 for the two more severe classes and equal to 0 otherwise. Moreover we considered the total ischemic time in the logarithmic scale too, because of clinical best practice and know-how. The choice of the covariates and the link function was suggested in [78], according to frequentist selection procedures and clinical best-practice, and confirmed in [66] using Bayesian tools.
Summing up, the model (8.5) we considered for our dataset is

$$logit(\mathbb{E}[Y_{ij}|b_j]) = logit(p_{ij}) = \theta_0 + \theta_1 \cdot age_i + \theta_2 \cdot \log(OB)_i + \theta_3 \cdot killip_i + b_j \tag{8.7}$$

for patient i ($i = 1, \ldots, 536$) in hospital j ($j = 1, \ldots, 17$). As far as the prior is concerned, we assume

$$\begin{gathered}
\theta \perp \mathbf{b} \qquad \theta \sim \mathcal{N}(\mathbf{0}, 100 \cdot \mathbb{I}_4) \\
b_1, \ldots b_J | P \overset{i.i.d.}{\sim} P \qquad P|\alpha, P_0 \sim Dir(\alpha P_0) \\
P_0|\sigma \sim \mathcal{N}(0, \sigma^2) \quad \sigma \sim Unif(0,10) \quad \alpha \sim Unif(0,30).
\end{gathered} \tag{8.8}$$

See details in [67]. The estimated posterior expected number of distinct values among the b_js, computed on 5000 iterations of Markov chain, is close to 7. In Table 8.9 the performances of different loss functions for different values of k and different threshold β_t are reported. The different values of p_t we considered (that determine the β_t values), were fixed in a range of values close to the empirical survival probability, in order to stress the resolution power of different loss in detecting unacceptable performances. Of course, when increasing the threshold p_t (and therefore β_t), more

hospitals will be labelled as unacceptable. The tuning depend on the sensitivity required by the analysis. The parameter a of the LINEX loss is set to be equal to -1.

Loss	$k = 0.5$ $p_t = 0.96$ $\beta_t = 3.178$	$k = 1$ $p_t = 0.96$ $\beta_t = 3.178$	$k = 2$ $p_t = 0.96$ $\beta_t = 3.178$
0/1	None	None	None
Absolute	None	None	None
Squared	None	None	9
LINEX	None	None	9

Loss	$k = 0.5$ $p_t = 0.97$ $\beta_t = 3.476$	$k = 1$ $p_t = 0.97$ $\beta_t = 3.476$	$k = 2$ $p_t = 0.97$ $\beta_t = 3.476$
0/1	None	9	3,5,9,10
Absolute	None	9	3,5,9,10
Squared	9	9	3,5,9,10
LINEX	9	3,5,9,10	3,5,9,10

Loss	$k = 0.5$ $p_t = 0.98$ $\beta_t = 3.892$	$k = 1$ $p_t = 0.98$ $\beta_t = 3.892$	$k = 2$ $p_t = 0.98$ $\beta_t = 3.892$
0/1	3,5,9,10	All	All
Absolute	2,3,4,5,9,10, 13,15	1,2,3,4,5,6,7,8,9,10, 11,13,14,15,16,17	All
Squared	2,3,4,5,9,10, 13,15	1,2,3,4,5,6,7,8,9, 10,13,14,15,17	All
LINEX	2,3,4,5,6,7,8,9, 10,13,15,17	All	All

Table 8.9: Providers labelled as "unacceptable", for different loss functions and different values of the threshold.

Some comments are due, observing results of the Table 8.9. Firstly, as mentioned before, k describes the different approach to evaluating misclassification errors. For example, people in charge with health care government might be more interest in penalizing useless investments in quality improvements, choosing a value less than 1 for k. On the other hand, patients admitted to hospitals are more interested in minimizing the risk of wrongly declaring as "acceptably performing" providers that truly behave "worse" than the gold standards; therefore, they would probably choose a value greater than 1 for k. Moreover, when fixing the loss functions among the four proposed here, and k equal to $0.5, 1$ or 2, as the threshold β_t increases, we obtain the same "implicit ranking" of providers:

$$9, 3, 5, 10, 2, 4, 13, 15, 6, 7, 8, 17, 1, 14, 16$$

(i.e., hospital 9 was classified as "unacceptable" even with small values of β_t, then, when increasing β_t, hospital 3 was classified as "unacceptable", etc.). This result is in agreement with the provider profiling pointed out also in [62]. On the other hand, Figure 8.9 shows the number of hospitals labelled as "unacceptable" as k increases, for a fixed value of the threshold β_t, under the Squared and the LINEX Loss functions.

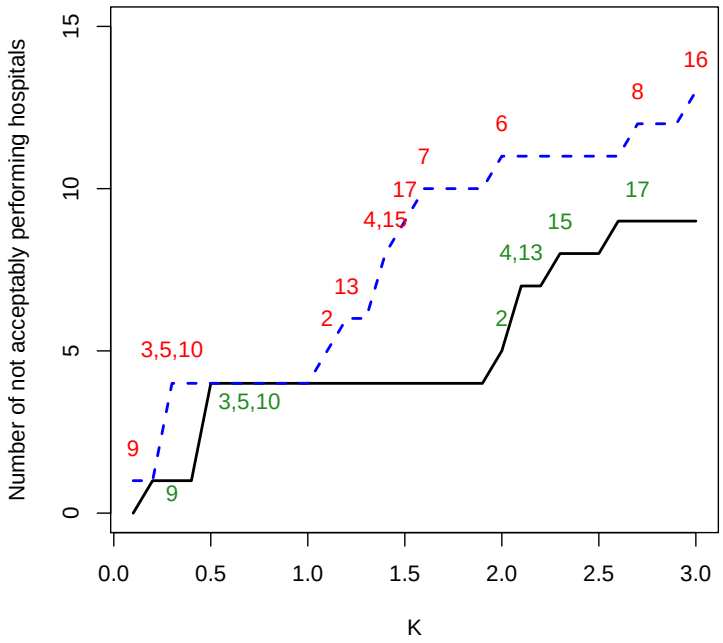


Figure 8.9: Number of hospitals labelled as "unacceptable" as a function of k, under the Squared Loss function (solid black) and the LINEX Loss function (dotted blue). The threshold parameter β_t is 3.6635.

Of course, the choice of the most suitable loss function is problem-driven: in our case, it seems reasonable to consider an asymmetric loss in order to penalize departures from threshold in different ways. For this reason we suggest the LINEX Loss with $k \neq 1$.

Chapter 9

Statistical analysis of ECG signals

In this chapter, the analyses carried out on data arising from PROMETEO datawarehouse (see Paragraph 3.2.4) are presented. In Section 9.1 the aim is to point out an unsupervised classification technique for semi-automatic diagnosis of Bundle Brunch Block, whereas in Section 9.2, multivariate functional depths are used to rank multivariate ECG signals and to carry out outliers detection and nonparametric comparison tests on them.

9.1 Semiautomatic diagnosis for Bundle Brunch Block

In this section we analyse a sample ($n = 198$) of data coming from PROMETEO datawarehouse with the aim of identifying, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. To this aim, it is extremely important to understand the link between cardiac physiology and ECG trace shape. As detailed in following paragraphs, we focus on physiological traces in contrast to Right and Left Bundle Branch Block (RBBB and LBBB respectively) traces. Bundle Branch Block (BBB) is a cardiac conduction abnormality seen on the ECG. In this condition, activation of the left (right) ventricle is delayed, which results in the one ventricle contracting later than the other. Details on Bundle Branch Blocks and their connection with non-physiological shape of ECG signal have been treated in Paragraph 3.2.4, where also clinical details about ECG signals have been given. As mentioned in Section 6.1, the analysis of ECG signals passes through a preprocessing step consisting of wavelet smoothing and landmark registration, which are necessary to denoise signals and removing the variability we are not interested in. From a statistical point of view, we will focus our analysis on shape modifications induced on the ECG curves and their first derivatives by the BBB pathology, and we will investigate these shape modifications only in a statistical perspective, i.e., not using clinical criteria to classify ECGs. The exploitation of these morphological modifications in the clustering procedure will be the focus of the following paragraphs.

9.1.1 Data smoothing and registration

We considered a sample of ECG signals consisting of $n = 198$ subjects, among which 101 are Normal and 97 are affected by BBB (49 RBBB and 48 LBBB). The basic statistical unit is the multivariate function which describes heart dynamics, for each patient, on the eight leads. However, in practice we have only a noisy and discrete observation of the function describing ECG trace for each patient. Moreover, each patient has his own “biological” time, i.e., the same event of the heart

dynamics may happen at different time measurements for different patients: this is only misleading from a morphological point of view. These two problems are common in Functional Data Analysis (FDA) applications and they can be addressed respectively with data smoothing and registration (see [41]).

Wavelets smoothing

The first step of the statistical analysis consists in data smoothing starting from noisy measurements: to this aim, the choice of the functional basis is crucial. Wavelet bases seem suitable for our data because every basis function is localized both in time and in frequency, being therefore able to capture ECG strong localized features (peaks, oscillations...). In particular we use a Daubechies wavelet basis with 10 vanishing moments (see [31] for details), because we are interested also in derivatives of the ECG traces and thus we need a basis smooth enough for this purpose. As in most smoothing methods based on wavelet expansion, it is necessary to deal with a grid of 2^J points, $J \in \mathbb{N}$. Thus, in the further analysis we use only the central $2^{10} = 1024$ observation points. There is no loss of significant information: the portion of the signal on which we focus the analysis contains all the important features of the ECG trace. For this reason, we choose not to turn to non-decimated wavelets, which could be applied also to non dyadic grid but require a larger computational effort.

Since the eight leads (i.e., I, II, V1, V2, V3, V4, V5 and V6) jointly describe the complex heart dynamic, when smoothing these data it is appropriate to use a technique which takes into account all the eight leads simultaneously. This helps in detecting significant features, which reflect on more then one leads. To this aim, in [120] it is developed a wavelet based smoothing technique for multivariate curves. This technique is used to obtain the estimation of 8 dimensional ECG signals. It has also the advantage to provide an estimate of derivatives, which is straightforward when the estimate is provided in functional basis expansion: it can be obtained simply by a linear combination of the basis functions derivatives.

Thus, starting from the vectorial raw signal, we estimate the vectorial function

$$\mathbf{f}_i(t) = (I_i(t), II_i(t), V1_i(t), V2_i(t), V3_i(t), V4_i(t), V5_i(t), V6_i(t)),$$

and its derivatives, for each patient $i = 1, \ldots, 198$. See [120] for a detailed description of this smoothing procedure. Figure 9.1 shows raw data and functional estimates obtained with this wavelet smoothing procedure for a normal subject. Observations are now in a functional form and thus we can use FDA techniques. The smoothing procedure is essential also for an accurate derivative reconstruction, as shown in Figure 9.2, where the estimate of the first derivative is superimposed to the first central finite difference (i.e., a rough indication of first derivative behavior).

Landmarks registration

Functional observations usually show both phase and amplitude variation, i.e., each curve has its own biological time so that the same feature can appear at different times among the patient. It is well known that a correct separation between these two kind of variability is necessary for a successful analysis [41]. We address this problem through a registration procedure based on landmarks, which are points of the curve that can be associated with a specific biological time. Five of these landmarks are provided by Mortara-Rangoni procedure and can be found in the *Details* file. They identify the P wave (P_{onset}, P_{offset}), QRS complex (QRS_{onset}, QRS_{offset}) and T wave (T_{offset}). We add one more landmark corresponding with the R peak on the I lead ($Ipeak$). We choose this landmark

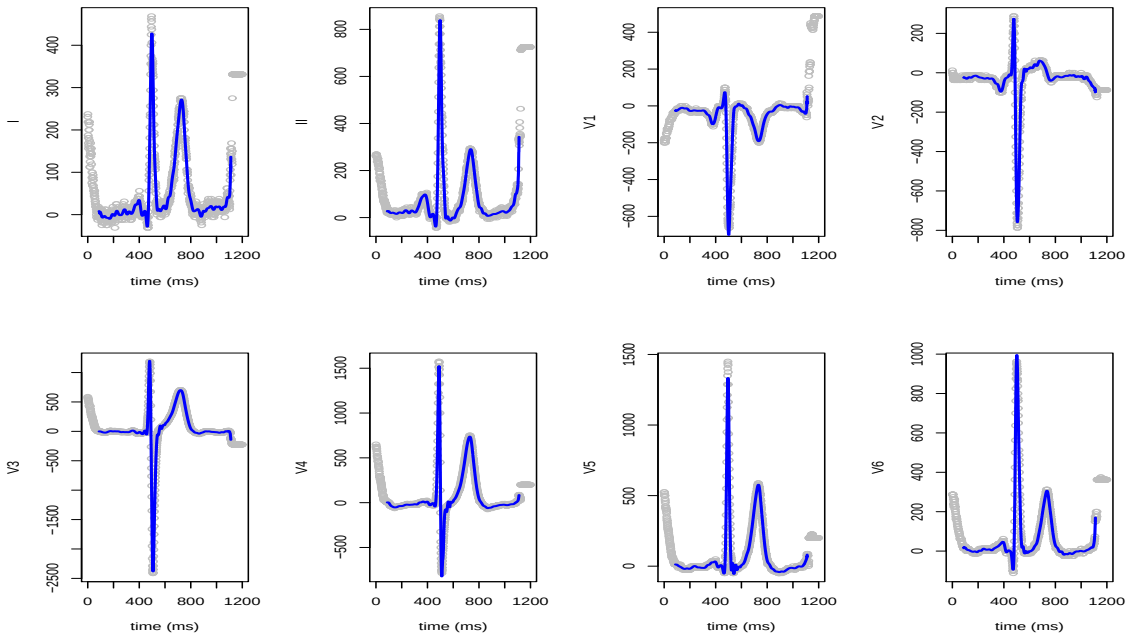


Figure 9.1: Raw data of the eight leads (black points) and wavelet functional estimates (blue) for a normal subject.

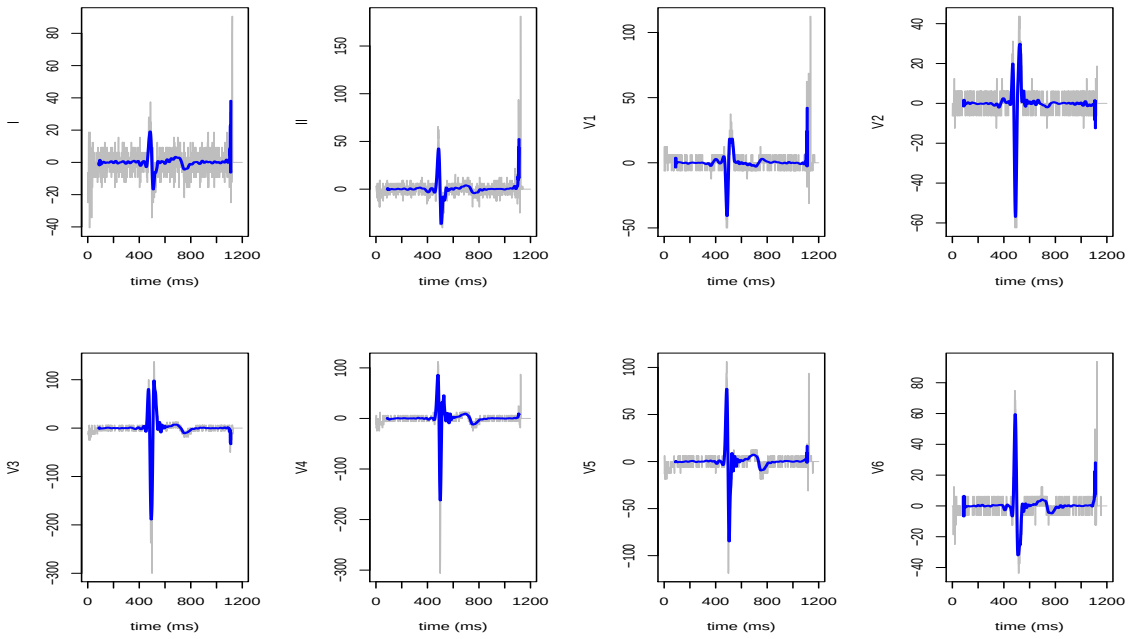


Figure 9.2: First central finite difference of the eight leads (gray) and wavelet estimates of the first derivatives (blue) for a normal subject.

because only on the I lead both physilogical and pathological ECG traces present a clearly identifi-

able R peak. Since all the leads capture the same heart dynamics, biological time must be the same. Thus, these landmarks can be used to register all the leads. For each patient i we look for a warping function h_i such that

$$h_i(P_{onset}) = P^0_{onset} \qquad h_i(P_{offset}) = P^0_{offset}$$

$$h_i(QRS_{onset}) = QRS^0_{onset} \quad h_i(Ipeak) = Ipeak^0$$

$$h_i(QRS_{offset}) = QRS^0_{offset} \quad h_i(T_{offset}) = T^0_{offset}$$

where P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset} and T^0_{offset} are the mean values of the correspondent landmarks. These values are reported in Table 9.1, together with the associated standard deviations. We solve this problem using spline interpolation of degree 3. Thus, the registered vectorial function will be

$$\mathbf{F}_i(t) = \mathbf{f}_i(h_i(t)),$$

for every patient $i = 1, \ldots, 198$. Figure 9.3 shows both unregistered and registered I leads for all the 198 patients. This is a non linear registration procedure, since in this framework there is no simple affine transformation which can take in account the subject specific variability.

	P^0_{onset}	P^0_{offset}	QRS^0_{onset}	$Ipeak^0$	QRS^0_{offset}	T^0_{offset}
mean	184.3	298.2	354.8	407.2	476.9	755.8
standard deviation	39.7	37.4	18.9	15.4	21.4	44.2

Table 9.1: Landmarks obtained at the end of the registration procedure, as the mean of landmarks of all the curves, and used to select the portion of smoothed and registered ECG curves relevant to our analysis (first line of the table); in the second line, landmarks standard deviations. Landmarks values are referred to a registered time in ms.

The registration procedure separates morphological information (i.e., amplitude variability) from duration of the different segments of ECG (i.e., phase variability). The former is captured by the registered ECG traces, while the latter is described by warping functions, determined by landmarks. In clinical practice the duration of different segments of ECG and particularly the QRS complex length is one of the most important parameters to identify pathological situations. However, this kind of information is not able to distinguish among different pathologies, such as Right and Left BBB. This can be seen also in our dataset. If we perform a multivariate 3-means algorithm on interval lengths ($P_{offset} - P_{onset}$, $QRS_{onset} - P_{offset}$, $QRS_{offset} - QRS_{onset}$ and $T_{offset} - QRS_{offset}$), with the aim of identifying the existing 3 groups, we obtain the result shown in Table 9.2: this method correctly separates physiological traces from pathological ones but it gives no information on the pathology.

	Normal	RBBB	LBBB
Cluster 1	96	6	0
Cluster 2	2	17	25
Cluster 3	3	26	23

Table 9.2: Confusion matrix related to patients disease classification. Results are obtained performing 3-means clustering algorithm on interval lengths.

For this reason, we focus our analysis on the registered curves, in the attempt to extract other diagnostic information from ECG morphology. In clinical practice, the result of our analysis should be considered together with traditional diagnostic tools based on segment lengths.

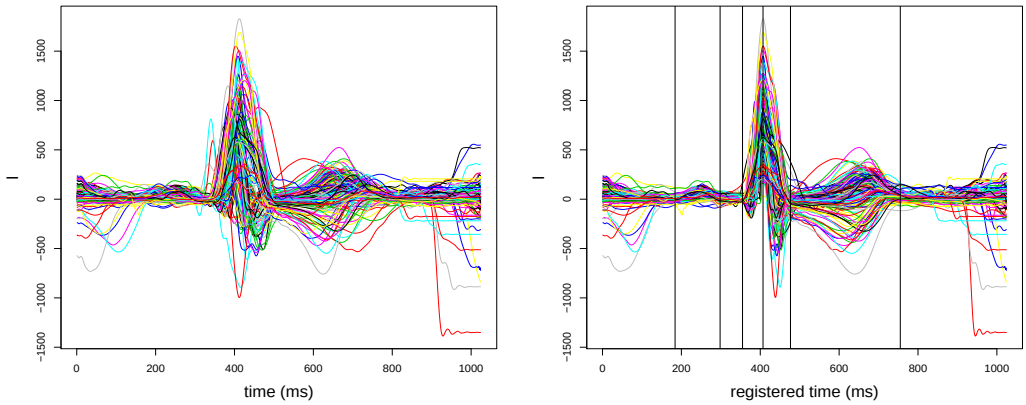


Figure 9.3: Original I leads for the 198 patients (left) and registered ones (right). Vertical lines indicate position of mean landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset} .

9.1.2 Data analysis

In this paragraph we propose the use of FDA techniques to perform clustering of smoothed and registered ECG traces. Aim of the analysis is the development of a proper classification procedure, able to distinguish the grouping structure induced in the sample of ECGs by the presence of different pathologies, on the basis of the sole shape of the considered curves.
As previously discussed, ECG traces are very complex functional data, in which different portions of the domain can be analyzed in order to detect different pathologies. The main focus of our analysis stands in the investigation of BBB pathology, which mainly expresses in the ECG trace through a lengthening of the QRS complex and a modification of the T wave. In fact, the diagnosis of BBB is not concerned with modifications in P wave, since this portion of the ECG curve deals with cardiac rhythm dysfunctions our patients are not affected by. We thus focus our classification analysis on the QT-segment. Since we have already registered the ECG signals, all the curves show relevant features at the same time points, corresponding to the reference landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset} (see paragraph 9.1.1): this fact allows us to select, for all the registered curves of the dataset, only the portion of ECG trace belonging to the interval $[P^0_{offset}, T^0_{offset}]$, which is relevant to our diagnostic purposes. In particular, we select only the portion of

$$\mathbf{F}(t) = \{F^r(t)\}_{r=1}^8 = (I(t), II(t), V1(t), V2(t), V3(t), V4(t), V5(t), V6(t))$$

such that $t \in T := [P^0_{offset}, T^0_{offset}]$, where P^0_{offset} and T^0_{offset} are the values reported in the first line, second and sixth columns of Table 9.1.

Functional classification

We analyze the n patients according to a functional k-means clustering procedure, in which all the eight leads $\mathbf{F}_i(t) : T \to \mathbb{R}^8$, for patients $i = 1, \ldots, n$, are simultaneously clustered. To develop

this clustering procedure we suppose that $\mathbf{F}_i(t) \in H^1(T;\mathbb{R}^8)$. Since we consider all the eight leads simultaneously in the analysis, we name the employed clustering procedure *multivariate functional k-means*, to distinguish it from *standard functional k-means*, which would treat each lead separately. A proper definition of functional *k*-means procedure and an introduction to its consistency properties can be found in [138]. We develop a similar *k*-means procedure, choosing the following distance between ECG traces

$$d_1(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8}\int_T (F_i^r(t)-F_j^r(t))^2 dt + \int_T (DF_i^r(t)-DF_j^r(t))^2 dt}, \tag{9.1}$$

for $i,j = 1,\ldots,n$, and with $DF_i^r(t)$ being the wavelet estimate of the first derivative of the *r*-th lead in the ECG trace of the *i*-th patient. Note that the distance defined in (9.1) is the natural distance in the Hilbert space $H^1(T;\mathbb{R}^8)$.

In order to perform comparisons, and to test the robustness of our clustering procedure, we considered two more distances between two ECG traces

$$\tilde{d_1}(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8}\int_T (DF_i^r(t)-DF_j^r(t))^2 dt}, \tag{9.2}$$

$$d_2(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8}\int_T (F_i^r(t)-F_j^r(t))^2 dt}. \tag{9.3}$$

The distance defined by (9.2) is the natural semi–norm in the Hilbert space $H^1(T;\mathbb{R}^8)$, while the one defined in (9.3) is the norm in the Hilbert space $L^2(T;\mathbb{R}^8)$: they are both considered in the clustering procedure not only to compare performances of multivariate functional *k*-means under different specifications of the distance, but also to have an insight on the role of curves first derivatives: we claim that both the ECG trace and its first derivative are essential to distinguish more similar morphologies from less similar ones.

Functional *k*-means clustering algorithm is an iterative procedure, which alternates a step of *cluster assignment*, in which all curves are assigned to a cluster, and a step of *centroid calculation*, in which a relevant functional representative (the centroid) for each cluster is identified. More precisely, in the cluster assignment step each curve is assigned to the cluster whose centroid (computed at the previous iteration) is nearer according to the distances defined in (9.1), (9.2) or (9.3) respectively. Instead, the identification of centroids $\varphi_l(t)$ for $l = 1,\ldots,k$, is performed solving the following optimization problem

$$\varphi_l(t) = \underset{\varphi\in\Omega_d}{\operatorname{argmin}} \sum_{i:C_i=l} d(\mathbf{F}_i(t),\varphi(t))^2,$$

where C_i is the cluster assignment of the i^{th} patient at the current iteration, d is one of the three distances defined in (9.1-9.3), and Ω_d is the Hilbert space with respect to which the chosen distance d is natural. The solution to this infinite dimensional optimization problem obviously depends on the choice of the distance: it is possible to prove that, both when the distance is measured with (9.1), and when it is measured with (9.3), the minimizer $\varphi_l(t)$ corresponds to the functional mean of curves belonging to the same cluster. An immediate consequence of this result is that, when the semi–norm in H^1 (eq. (9.2)) is used, the centroid is the functional mean of the first derivatives of curves belonging to the same cluster.

There are many different implementations of functional *k*-means algorithm in the literature on functional data analysis, among which some procedures integrate registration in the classification

steps (see for example [18], [103] and [104]). Here, instead, we chose to separate registration and clustering in two subsequent steps of the analysis, since the latter doesn't use any information beside morphology of the ECG traces, while the former is based on a strong clinical indication provided by landmarks supplied by the Mortara-Rangoni VERITAS^{TM} algorithm.

The k-means clustering procedure clearly depends not only on the choice of the distance, but also on the number of clusters k. Being the number of clusters a–priori unknown, we also consider a way to select the optimal number of clusters k^* via silhouette values and plot of the final classification [136]. Note that a patient which alone constitutes a cluster, has silhouette value equal to 1, but he is not considered in the silhouette plot for choosing k^*.

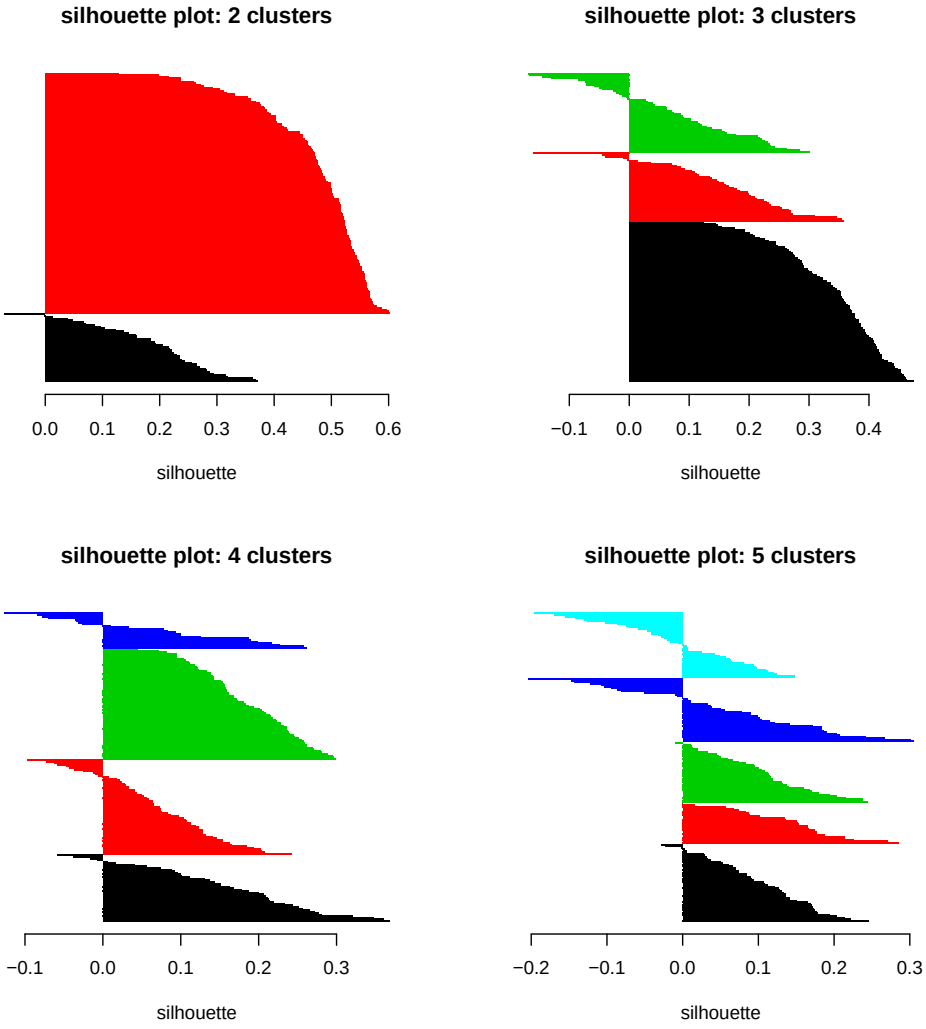


Figure 9.4: Silhouette plots of the clustering result obtained via multivariate functional k-means procedure, setting $k = 2,3,4,5$ and with distance given by (9.1); data are ordered according to an increasing value of silhouette within each cluster, and are coloured according to the cluster assignment.

9.1.3 Results and discussion

Aim of the analysis is to detect the underlying grouping structure in our sample of 198 ECG traces. We thus perform clustering of the whole dataset via the multivariate functional k-means algorithm

previously described, using the different definitions of the distance between curves given in (9.1-9.3). The final silhouette plots obtained by clustering the sample of 198 ECG traces according to a multivariate functional k-means procedure with distance d_1 (9.1), and setting $k = 2,3,4,5$, are shown in Figure 9.4. As we can appreciate from the picture, the grouping structure obtained setting $k = 3$ seems the best one, both in terms of silhouette profile, and in terms of wrong assignments. A similar result is obtained measuring the distance between curves via (9.2) or (9.3); however, the procedure seems to detect the best grouping structure when both the curves and their derivatives are considered in the distance. We thus set $k^* = 3$. The final classification obtained with this choice of the distance, and setting $k = 3$, is shown in Figure 9.5, where the whole functional dataset is coloured according to cluster assignments; each panel corresponds to a different lead. From inspection of this picture a different shape of ECGs assigned to different clusters can be immediately appreciated, especially looking at the final centroids (functional mean) of each group, drawn in black in each panel of the picture. We shall now verify whether this difference in the ECGs morphology across clusters is due to the different pathology.

Since we have an indication of the different pathologies of the patients included in the sample, we can analyze the confusion matrix associated to the final cluster assignments, with respect to the Mortara-Rangoni algorithm classification (Normal, RBBB and LBBB). The confusion matrices obtained via multivariate functional k-means with different choices of the distance between curves (given by d_1, $\tilde{d}_1$ or d_2) are shown in Table 9.3. We remark that the final cluster assignments are based on the sole shape of the smoothed and registered ECG curves and their first derivatives, analyzed via a unsupervised classification procedure. Both choosing the H^1 norm and the L^2 norm, the results seem appreciable, and slightly better in the former case: the final grouping structure traces out quite coherently the patients disease classification, with only few cases wrongly assigned. Moreover, we remark the improvement in the results obtained via multivariate functional 3-means with respect to the results of 3-means clustering algorithm on interval lengths (see Table 9.2): we are now able not only to detect pathological subjects, but also to distinguish between the two different pathologies present in the dataset. The result obtained via multivariate functional 3-means clustering with H^1 semi-norm, instead, is not so positive, since cluster 1 and 2 apparently merge physilogical traces with ECGs of patients affected by RBBB.

	H^1 norm			H^1 semi-norm			L^2 norm		
	Normal	RBBB	LBBB	Normal	RBBB	LBBB	Normal	RBBB	LBBB
1	95	7	1	71	12	0	94	6	2
2	6	42	3	30	36	5	7	43	3
3	0	0	44	0	1	43	0	0	43

Table 9.3: Confusion matrices related to patients disease classification. Results are obtained by application of multivariate functional 3-means clustering algorithm to smoothed and registered QT-segment of ECG curves, with different choices of the distance between ECGs: H^1 norm (eq. (9.1), first table), H^1 semi-norm (eq.(9.2), second table) and L^2 norm (eq. (9.3), third table). In the first table, cluster 1,2,3 respectively correspond to orange, green and red in Figure 9.5.

The effectiveness of the clustering procedure in detecting the grouping structure among data suggests the definition of a semi–automatic diagnostic procedure based on the multivariate functional k-means algorithm: in fact, the final result of our clustering procedure is a set of k centroids, representative of each cluster, which can be used as reference signals to compare a new ECG trace. Suppose a new ECG signal is available: we could have an immediate hint on the new patient's diagnosis by smoothing its ECG trace, registering it and finally assigning it to the group characterized by the nearest centroid.

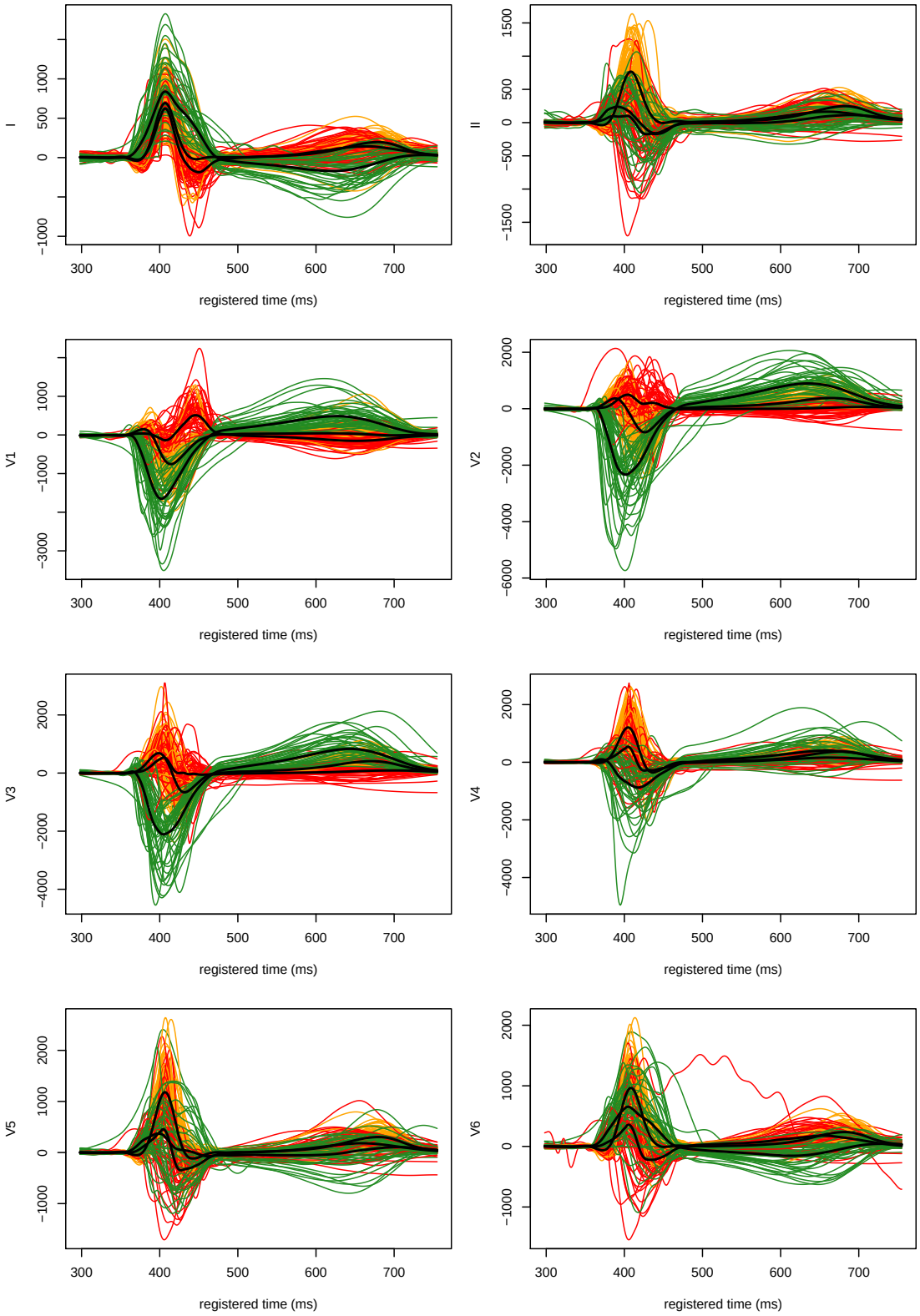


Figure 9.5: Smoothed and registered ECG traces (QT-segment): the whole dataset is coloured according to the final cluster assignments of multivariate functional 3-mean clustering, with distance given by (9.1); the superimposed black lines are the three final cluster centroids (functional means). Each panel correspond to a different lead of the ECG traces.

It is important to evaluate the *misclassification cost* for this procedure, with the choice of the different functional distances. To this aim, we perform a cross-validation analysis. We randomly choose among ECGs a training set of 80 Normal subjects, 40 to RBBBs and 40 to LBBBs, for a total of $n_{training} = 160$ curves. A multivariate functional 3-means clustering is performed on the selected training set; we then consider the remaining $n_{test} = 38$ curves, and we assign each of them to the cluster whose centroid is nearer, according to distances (9.1)-(9.3). Given the patients disease classification, we compute misclassification cost using the following index

$$cost_{CV} = \frac{\lambda_1 \cdot misc_N + \lambda_2 \cdot (misc_{RN} + misc_{LN}) + \lambda_3 \cdot (misc_{RL} + misc_{LR})}{n_{test}}, \tag{9.4}$$

where $misc_N$ is the number of healthy patients assigned to a pathological cluster[1], $misc_{RN}$ and $misc_{LN}$ are the number of patients respectively affected by RBBB and LBBB which are assigned to the cluster of healthy patients, while $misc_{RL}$ and $misc_{LR}$ are the number of patients whose ECGs are detected as pathological, but whose pathology is wrong. The parameters λ_1, λ_2 and λ_3 are misclassification weights: they are chosen according to the suggestion of the clinicians, who believe that assigning a BBB patient to the cluster of healthy patients is approximately 4 times more serious than treating as pathological a normal subject, which indeed is two times more serious than assigning a RBBB patient to the LBBB cluster (or viceversa); in order to determine the values of the weights we introduce a further request: $cost_{CV} = 1$ in the worst case, i.e., when all Normal subjects are classified as BBB and all BBB subjects are classified as Normal. This led to the choices $\lambda_1 = 0.4270$, $\lambda_2 = 1.7079$ and $\lambda_3 = 0.2135$. We repeat this procedure 20 times, computing each time the misclassification cost according to equation (9.4): the mean and standard deviation computed along the 20 cross-validation repetitions are shown in Table 9.4. Even if all the distances (9.1)-(9.3) provide good results, we notice that the norm in the Hilbert space $H^1(T;\mathbb{R}^8)$ seems to give best results, thus confirming our initial claim: both registered curves and first derivatives are needed to accurately compare ECGs morphology.

distance	d_1	$\tilde{d}_1$	d_2
mean $cost_{CV}$	0.1227563	0.2286588	0.1275316
std dev $cost_{CV}$	0.1112663	0.1050911	0.1220574

Table 9.4: Mean misclassification cost (first row) and standard deviation (second row) computed over 20 repetitions of the cross-validation procedure via equation (9.4).

[1] given the final cluster assignments, the cluster of healthy patients is detected as the one that includes the most physiological traces. The pathological ones are subsequently chosen, first the one that contains the more RBBB traces, while the cluster that remains is the LBBB one.

9.2 Depth measures for multivariate functional data

In Section 9.1, a statistical framework for analysis and classification of ECG curves starting from their sole morphology is proposed. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. In order to do this, a real time procedure consisting of preliminary steps like reconstructing signals, wavelets denoising and removing biological variability in the signals through data registration is tuned and tested. Then, a multivariate functional k-means clustering of reconstructed and registered data is performed. Since when testing new procedures for classification the performances of classification method are to be validated through cross validation, it is mandatory a suitable training of the algorithm on data. This would lead to robustify classification algorithm and would improve reliability in prediction. The procedure proposed in the previous Section is an effective way to reach this goal. In fact, it leads to select for the training set the proportion of multivariate curves whose depth is greater. Considering the ECG of the j-th patient as a 8-variate function $\mathbf{f}_j = (f_{j;1}, ..., f_{j;8})$, the $f_{j;k}$, $(k = 1, ..., 8)$ correspond to the eight leads I, II, V1, V2, V3, V4, V5 and V6. Then the procedure discussed in Section 6.2 is applied in order to carry out functional boxplots and to perform outliers detection for two different groups: physiological and pathological patients, i.e., people affected by a particular kind of heart disease, called Bundle Branch Block (BBB). This is a pathology which is easy to detect through the observation of shape modifications it induces on ECG pattern and divides in Right Bundle Branch Block (RBBB) and Left Bundle Branch Block (LBBB) according to the heart side it affects. In the following, we will consider a sample of 100 physiological signals and 50 pathological ones, where the latter come from patients affected by LBBB.

In Figures 9.6 and 9.7 the raw data are shown, whereas Figures 9.8 and 9.9 show the corresponding functional boxplots, one for each lead of the ECG, (see [81] and [82] for details on statistical analysis and procedures). Functional Boxplots are produced according to the ranking induced by the multivariate functional index where the weights p_k, $(k = 1, \ldots, 8)$ are all equal to $1/8$, weigthing in the same way all the leads. Since there is a common ranking of all components of $\mathbf{f}_j$s, induced by the multivariate index of depth, the central band is defined with the same curves in each component of the functional boxplot, since the multivariate funtional index of depth defined in (6.4) takes jointly into account the order of each component (lead) of the multivariate function (ECG). This is the main and most important difference between functional boxplots reported in Figures 9.8 and 9.9 and those we would have obtained simply asking for functional boxplots of each lead.

As described in Section 6.2, given the order in the sample of curves induced by the multivariate functional depth measure, it is possible to widen to this framework a non parametric rank test in order to compare two samples of multivariate functions. Actually, we will adopt the rank test to check for differences in the underlying process generating the LBBB curves with respect to the physiological ones. Then the combined dataset consists of 150 8-variate functional ECG signals. The $p-$value of the test carried out on these curves using the multivariate functional index computed on them all is equal to $3.38 * 10^{-16}$. The statistical evidence is still very strong ($p-$value $= 2.96 * 10^{-16}$) if we compute the depth measure (6.4) setting $(p_1, ..., p_8)$ equal to $(1/10, 1/10, 2/10, 1/10, 1/10, 1/10, 1/10, 2/10)$, stressing the weight of leads V1 and V6, since they are the most important for carrying out the LBBB diagnosis, as confirmed by cardiologists. That is, a strong evidence for the LBBB to be considered as arising from a different latent process exists. This is also detectable looking at the functional boxplots arising from the combined database of physiological and pathological signals, shown in Figure 9.10: almost all the outliers are those related to LBBB signals.

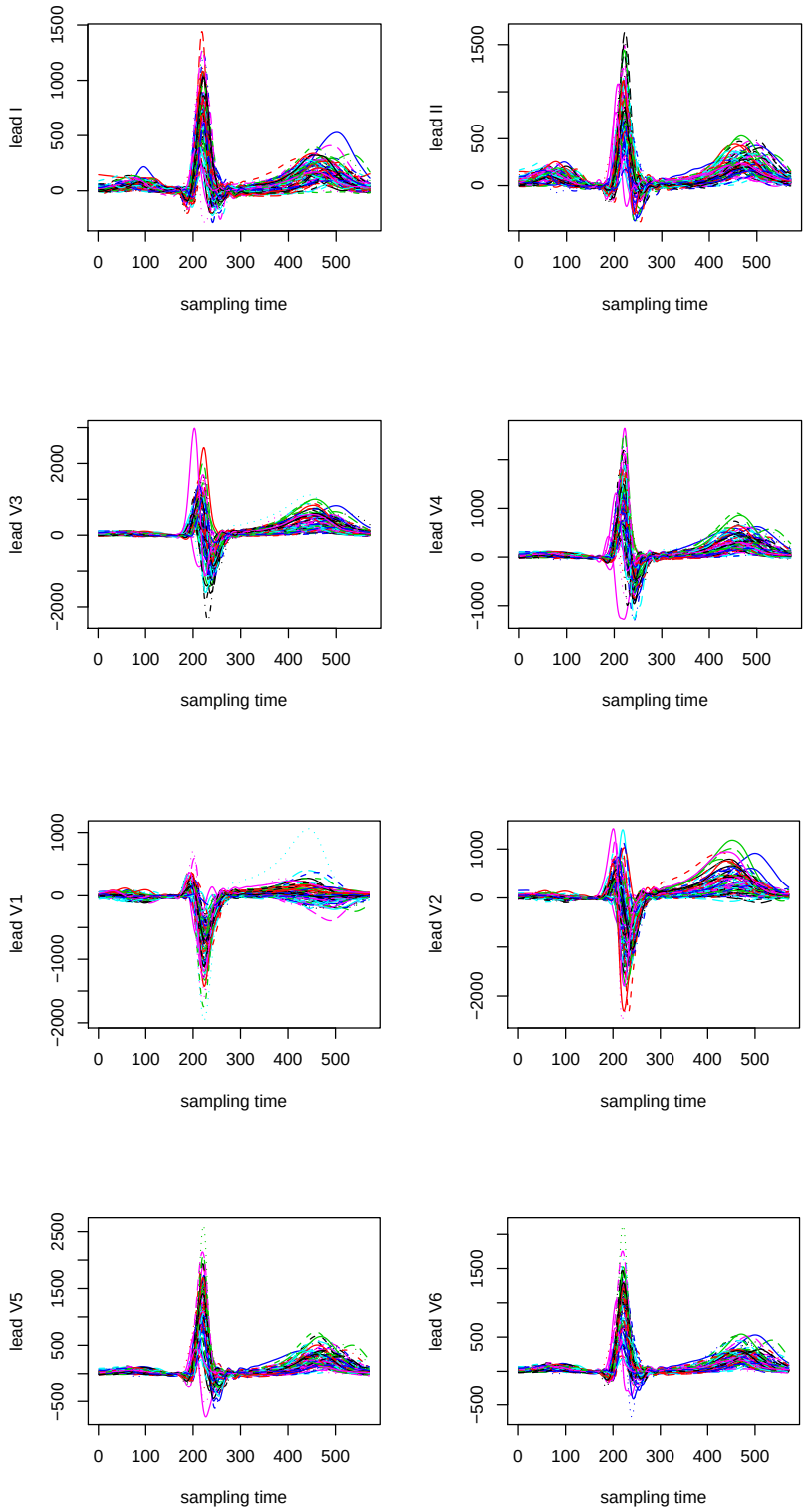


Figure 9.6: Raw signals of the 100 physiological patients.

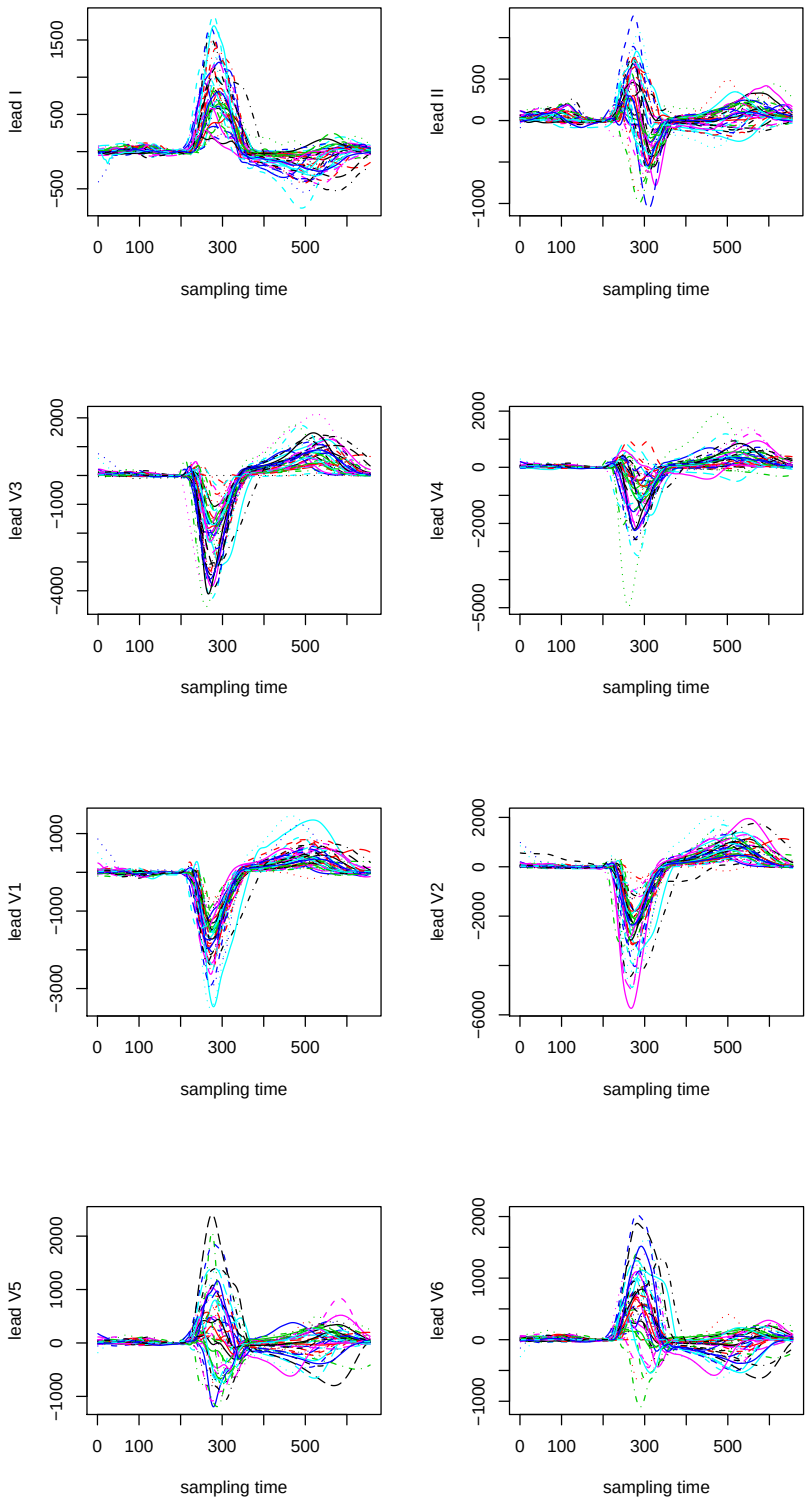


Figure 9.7: Raw signals of the 50 pathological patients.

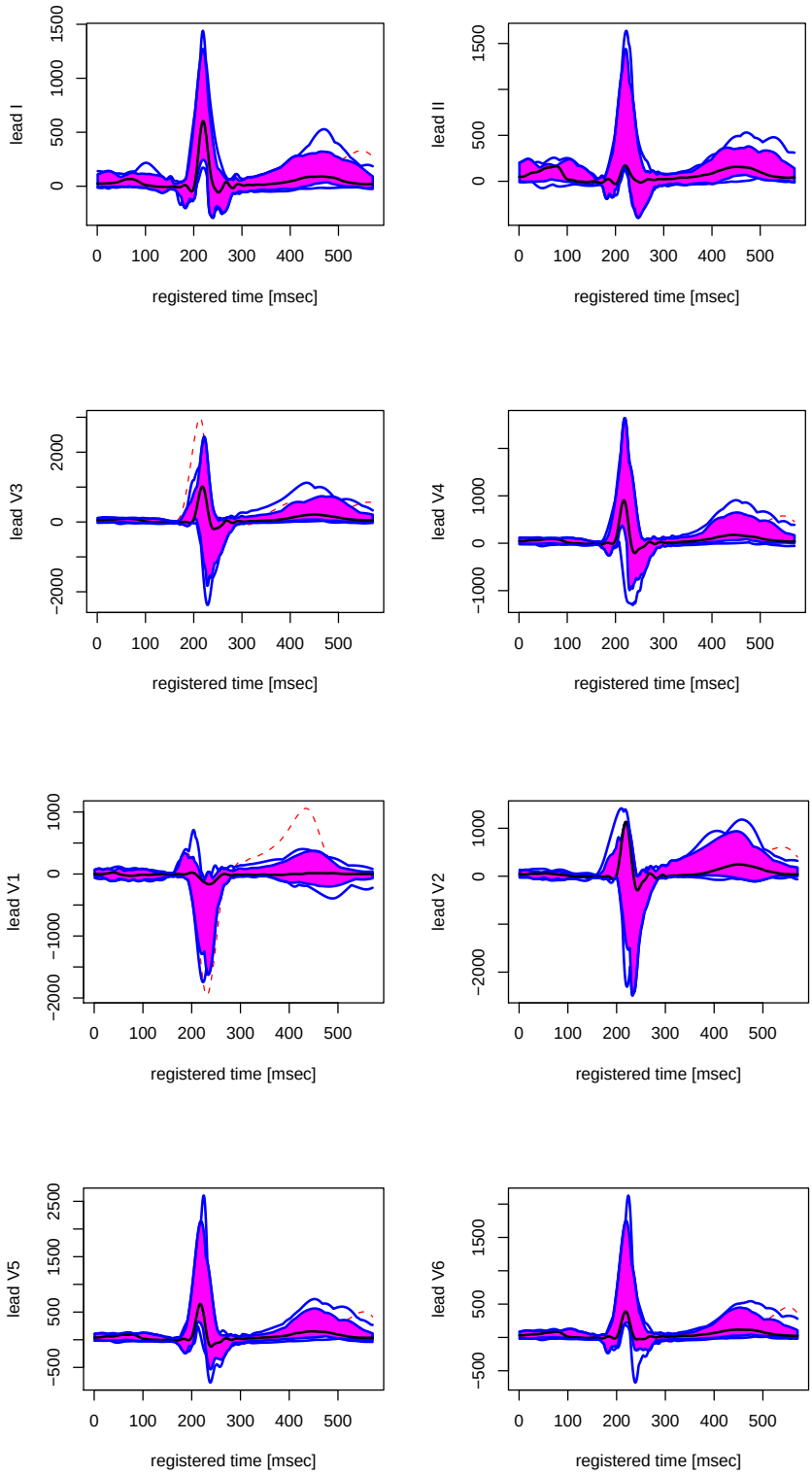


Figure 9.8: Functional boxplots of each component (*lead*) of the 100 physiological ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4).

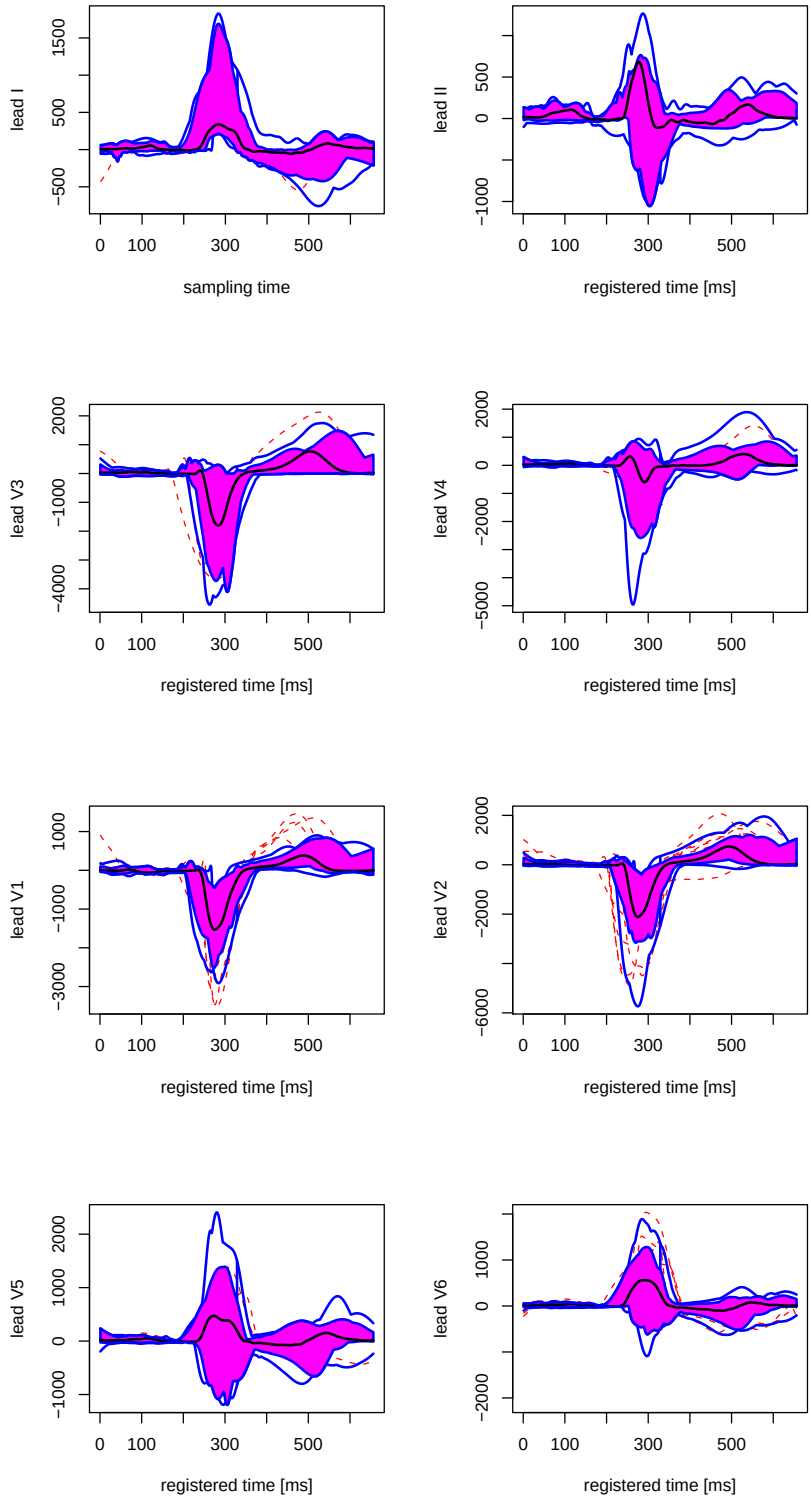


Figure 9.9: Functional boxplots of each component (*lead*) of the 50 pathological (Left Bundle Brunch Block) ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4).

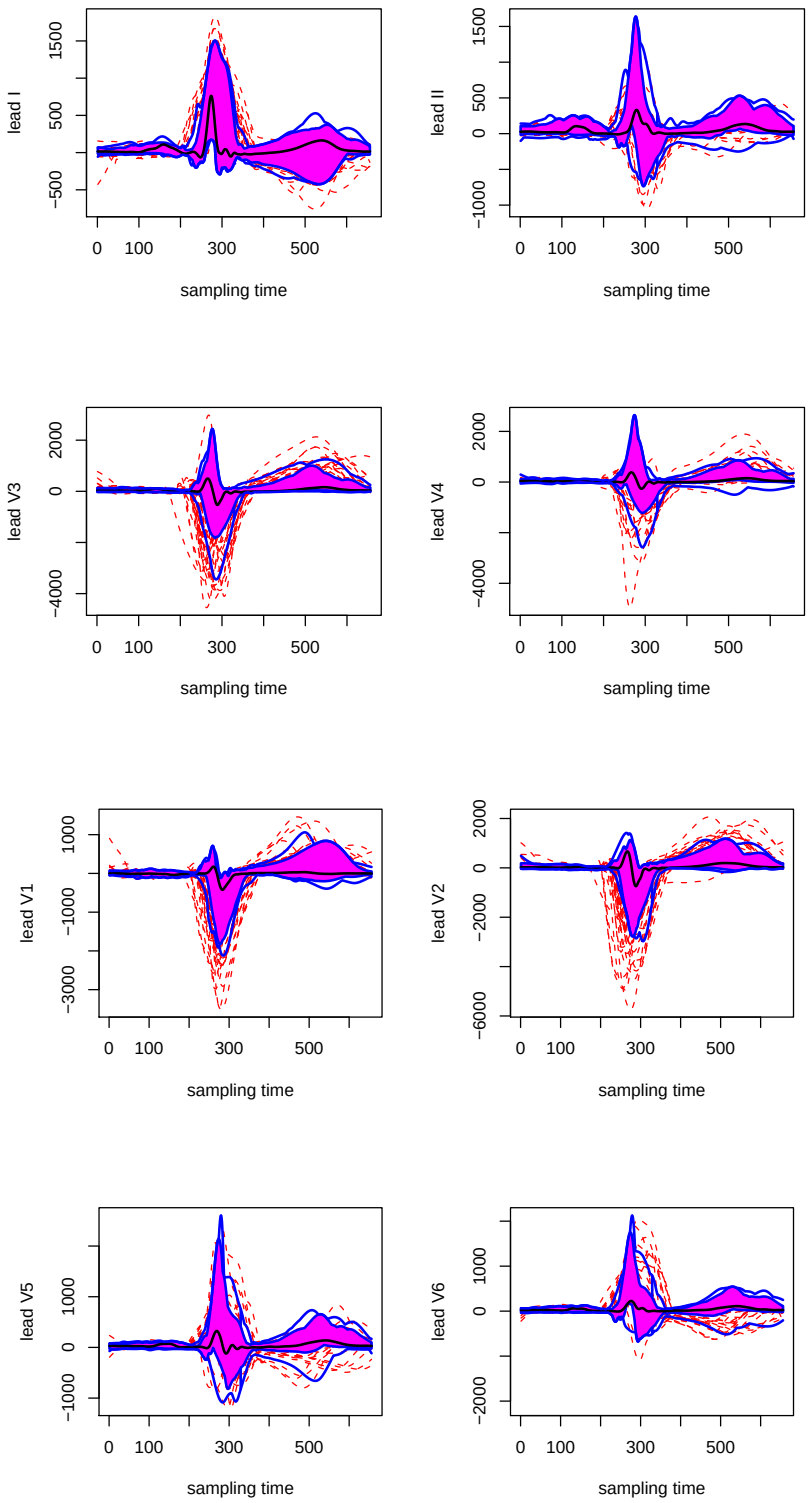


Figure 9.10: Functional boxplots of each component (*lead*) of the 150 physiological (100) and pathological (50 Left Bundle Brunch Block) ECGs. The central bands (purple coloured area) and outliers (red dotted lines) of each lead are defined as described in Section 6.2, according to the ranking induced by $MBD_n^J(\mathbf{f})$ defined in (6.4).

Conclusions and Future Works

In this thesis, we develop statistical methods focused at supporting decisions in healthcare context. In particular, we focus on optimization and improvement of patterns of care for cardiological and cardiovascular patients treated in any hospital of the Cardiologica Network of Regione Lombardia within the Strategic Program [36], and on development of statistical tools for the semi automatic diagnosis of specific types of infarction.

Concerning these frameworks, we designed a new clinical registry (namely STEMI Archive) to collect data about patients affected by Acute Coronary Syndromes and we used this tool to monitor and evaluate the process consisting of care delivery and treatment in the pre- and in-hospital phase, as well as after patient's discharge. This is made possible by integration of the clinical registry with the administrative datawarehouse of Regione Lombardia. Such instruments provide a real time picture of the healthcare offer of our Regional District with respect the cardiological and cardiovascular syndromes, enhancing at the same time the already exsisting administrative resources of data of Regione Lombardia, and depict how the Cardiological Network works in managing these patients, highlighting critical situations to be acted upon. The model we propose for monitoring and evaluating the healthcare process of interest is sustainable and effective, as proved also by previous pilot experiences performed on Milanese urban area. It is also general and flexible, and can then be applied to any other pathology or care process. Moreover, it enables people in charge with healthcare government to plan activities and make investments according to real epidemiological evidence and needs.

One of the main results achieved thanks to our collaboration to Strategic Program in these three years concerns the Design and activation of a clinical registry for collecting data on Acute Coronary Syndromes (STEMI Archive [36]), which involves all public and private cardiological divisions of Regione Lombardia, since it is a mandatory data collection, part of the goals of *Direttori Generali*. Moreover, the work carried out within the Strategic Program turned out in the strengthening and extension of the Network among providers. In fact, thanks to the shared protocols and information technology systems developed in these years, it is now active a Regional Network among hospitals, connected by 118 rescue service, that could be extended also to those pathologies that, similarly to Acute Coronary Syndromes, have a wide spread, a high incidence and takes benefits by well timed organization.

From a statistical perspective, we focused our research on regression models and classification techniques for longitudinal, grouped and multivariate functional data. In particular, we studied and adopted regression models, namely mixed effects models, able to handle grouped and longitudinal data arising from clinical context. The aim was profiling providers acording to the clustering of grouping factor's effect on outcome of interest, as shown in [62], [78] and [79]. We mainly considered generalized linear mixed effects models, since the most part of the outcomes of interest consisted in binary variables, and we explored linear and nonlinear dependencies among outcomes and model parameters, as well as parametric and nonparametric approaches to the modelling of the

random effects distributions. This led to face some complexities in terms of parameters' estimation. Firstly, since for GLME models likelihood integrals cannot be worked out analytically, numerical approximations are needed, based on the approximation of the integrand, on an approximation of the data, or on the approximation of the integral itself. Within GLME models, also problems due to unbalanced shares are treated. Then the choice of linear or nonlinear models has been considered, developing also new algorithms for the nonlinear nonparametric case [12], in order to fit data at best. Finally, both frequentist and Bayesian approaches have been considered, studied and implemented. Comparisons among different methods within the same framework [62] as well as among different approaches have been carried out. We have been confirmed in our initial guess that nonparametric techniques suit better the unsupervised clustering aims of our research, especially in nonlinear context, since more and more often parametric assumptions are too restrictive for modelling complexity of data arising from clinical surveys. Anyway, the joint sequential use of nonparametric and parametric techniques (as shown, for example, in [77]), together with the information coming from the clinical best practice, could lead to a more refined choice of the model, increasing its goodness of fit and its predictive power. Moreover, it often happens that results arising from parametric models are easier to communicate to the clinical community, and this remains an important issue to be considered in the monitoring and evaluating healthcare processes. Concerning the model predictive power, we saw how Bayesian posterior densities can be adopted to point out new decison mechanisms and thresholds for patients' classification [64]. Bayesian nonparametrics provide also several advantages in terms of in-built classification of the random effects [67]. Moreover the decisional issue of hospitals' profiling can be effectively addressed setting the problem within the Bayesian decision theory, as proposed in [65].
We considered also statistical methods for dealing with multivariate functional data, with the aim of addressing problems connected with the semi automatic diagnosis of cardiac diseases that are detecteable through electrocardiogram. We developped suitable clustering tools for the unsupervised classification of functional data [80], and generalized some results on functional depth measures to the multivariate functional case and performed nonparametric tests for carrying out inference on the difference between families of multivariate functional curves, as detailed in [81] and [82].

Further developments of this work are actually ongoing both from clinical and statistical perspective. As previously mentioned, Regione Lombardia is considering the extension of this paradigm of monitoring and evaluation of healthcare process also to other family of diseases. Moreover, the STEMI Archive linked to the administrative database may also be used for specific clinical enquiries concerning health technology assessment, cost-effectiveness studies on drugs utilization and so on.

There are several potential further directions to be investigated in statistics. We are actually working on generalization of results on nonparametric modelling of random effects to mixtures models, as well as on consistency properties and convergence of multivariate functional indexes od depth. It would also be interesting to move towards the theory of Generalized Additive Mixed Models (GAMMs), in order to use them for better accounting for the complex relationship among outcomes at patient's level and the process he/she undergoes. Another current topic where our research is moving toward is the use of ECG signals comning from PROMETEO database for studying and validating numerical simulated ECGs.

In general, this work showed how statistical methods can be adopted for handling complex problems arising from clinical and healthcare context. The strength of this experience consists in the profitable collaboration among Regional District government of healthcare, statisticians and physicians or in general players involved in the delivery of care. We believe that this way of monitoring and evaluating processes starting from on-going, shared and well structured data collections can really

highlight critical points and enable people involved in the process to act upon them. The continuous statistical control aimed at improving healthcare service can really be helpful in optimizing resources and in making decisions, this unavoidably reflecting in improvements of services offered to patients to safeguard their health.

Bibliography

Statistical references:

[1] Aalen, O.O. and Borgan (2006), Survival and event history analysis: a process point of view, Springer-Verlag, New York

[2] Agresti, A. (2002), Categorical Data Analysis, Wiley-Interscience, New Jersey, 2nd edition

[3] Aitkin, M. (1996), A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251-262.

[4] Aitkin M. (1999), A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models, *Biometrics*, **55**: 117–128.

[5] Albert, J.H. and Chib, S. (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 422, 669–679.

[6] Andersen, K., Borgan, O., Gill, R.D., Keiding, N. (1993), Statistical models based on counting processes, Springer-Verlag, New York

[7] Antic, J., Laffont, C.M., Chafaï, D., Concordet, D. (2009), Comparison of Nonparametric Methods in Nonlinear Mixed Effect Models. *Computational Statistics and Data Analysis* **53**, 3, 642–656.

[8] Antoniak, C.E. (1974), Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, *Annals of Statistics*, **2**, 6, 1152–1174

[9] Austin, P.C. (2008), Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals, *BMC Medical Research Methodology*, 8–30

[10] Austin, P.C., Lawrence, J.B. (2008), Optimal Bayesian probability levels for hospital report cards, *Health Service and Outcomes Research Method*, **8**, 80–97

[11] Azzimonti, L., Ieva, F., Paganoni, A.M. (2011), A new unsupervised classification algorithm for nonlinear non parametric mixed effects models, *Proceedings* of SCo2011, Seventh Conference

[12] Azzimonti, L., Ieva, F., Paganoni, A.M. (2011), Nonlinear nonparametric mixed-effects models for unsupervised classification, *Submitted* [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/22-2011.pdf

[13] Barbieri, P., Grieco, N., Ieva, F., Paganoni, A.M., Secchi, P. (2010), Exploitation, integration and statistical analysis of Public Health Database and STEMI archive in Lombardia Region, *Complex data modeling and computationally intensive statistical methods - Series "Contribution to Statistics"*, Springer, 41–56

[14] Baraldo, S., Ieva, F., Paganoni, A.M., Vitelli, V. (2010), Statistical models for hazard functions: a case study of hospitalizations in health failure telemonitoring, *Acts* of XLV Scientific Meeting of the Italian Statistical Society 2010, Padua (Italy), June 16-18, 2010.

[15] Baraldo, S., Ieva, F., Paganoni, A.M., Vitelli, V. (2010), Generalized functional linear models for recurrent events: an application to re-admission processes in heart failure patients, *Submitted* [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/42-2010.pdf

[16] Bates, D., Maechler, M. (2010), lme4: linear mixed-effects models using S4 classes. Available at http://CRAN.R-project.org/package=lme4

[17] Blackwell, D., MacQueen, J.B. (1973), Ferguson distributions via Polya urn schemes, *Annals of Statistics*, **1**, 353–355

[18] Boudaoud, S., Rix, H., and Meste, O. (2010), Core Shape modelling of a set of curves, *Computational Statistics and Data Analysis*, **54**, 308–325

[19] Breslow N.E., Clayton D.G. (1993), Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, **88**, 421, 9–25

[20] Breslow, N.E., Lin, X. (1996), Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion *Journal of the American Statistical Association*, **91**, 435, 1007–1016

[21] Browne W.J., Goldstein H., Jones K., Subramanian, S.V.(2005), Variance partitioning in multilevel logistic models that exhibit overdispersion, *Journal of the Royal Statistical Society A*, **68**, 3, 599–613

[22] Browne W.J., Draper D. (2006), A comparison of Bayesian and likelihood-based methods for fitting multilevel models, *Bayesian Analysis*, **1**, 3, 473–514

[23] Casella, G., George, E.I. (1992), Explaining the Gibbs Sampler, *The American Statistician*, **46**, 3, 167–174

[24] Chib, S., Greenberg, E. (1995), Understanding the Metropolis-Hastings Algorithm, *American Statistician*, **49**, 4, 327–335

[25] Clayton, D. and Kaldor, J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, **43**, 671–681

[26] Congdom, P.D. (2010), Applied Bayesian Hierarchical Methods, Chapman & Hall

[27] Cox, D.R. (1972), Regression models and life tables, *Journal of the Royal Statistical Society B*, **34**, 2, 187–220

[28] Cramer J.S. (1999), Predictive Performance of the Binary Logit Model in Unbalanced Samples, *Journal of the Royal Statistical Society D*, **48**, 1, 85–94

[29] Cuevas, A., Febrero, M., Fraiman, R. (2007), Robust Estimation and Classification for Functional Data via Projection-based Depth Notions, *Computational Statistics*, **22**, 481–496.

[30] Daniel M.J., Gatsonis C. (1999), Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization, *Journal of the American Statistical Association*, **94**, 445, 29–42

[31] Daubechies, I. (1988), Orthonormal basis of compactly supported wavelets, *Communictions on Pure and Applied Mathematics*, **41**, 909–996.

[32] Davidian, M., Gallant, A.R. (1993), The Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Biometrika* **80**, 3, 475–488.

[33] De Iorio, M., Muller, P., Rosner, G.L., Mac Eachern, S.N. (2004), An ANOVA Model for Dependent Random Measures, *Journal of American Statistical Association*, **99**, 465, 205–215

[34] De Lalla, C., Rinaldi, A., Montagna, D., Azzimonti, L., Bernardo, M.E., Sangalli, L.M., Paganoni, A.M., Maccario, R., Di Cesare Merlone, A., Zecca, M., Locatelli, F., Dellabona, P., Casorati, G. (2011), Invariant Natural Killer T-cell reconstitution in pediatric leukemia patients given HLA-haploidentical stem cell transplantation defines distinct CD4+ and CD4- subset dynamics and correlates with remission state. *The Journal of Immunology* **186**, 7, 4490–4499.

[35] Decreto N^o 20592, 11/02/2005, Direzione Generale Sanità - Regione Lombardia (2005), Patologie cardiocerebrovascolari: interventi di prevenzione, diagnosi e cura

[36] Decreto N^o 10446, 15/10/2009, Direzione Generale Sanità - Regione Lombardia (2009), Determinazioni in merito alla Rete per il trattamento dei pazienti con Infarto Miocardico con tratto ST elevato(STEMI)

[37] Einbeck, J., Darnell, R., Hinde, J. (2009), npmlreg: nonparametric maximum likelihood estimation for random effect models. Available at http://CRAN.R-project.org/package=npmlreg

[38] Febrero, M., Galeano, P., Gonzalez-Manteiga, W. (2008), Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, *Environmetrics*, **19**, 331–345

[39] Fellegi, I., Sunter, A. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, **64**, 328, 1183–1210

[40] Ferguson, T.S. (1973), A Bayesian analysis of some non-parametric problems, *Annals of Statistics*, **1**, 209–230

[41] Ferraty, F., Vieu, P. (2006), Nonparametric functional data analysis, Springer-Verlag, New York

[42] Fox, J. (2002), Linear Mixed Models, Appendix to An R and S-PLUS Companion to Applied Regression

[43] Fraiman, R. and Meloche, J. (1999), Multivariate L-estimation, *Test*, **8**, 255–317

[44] Fraiman, R. and Muniz G. (2001), Trimmed means for functional data, *Test*, **10**, 419–440

[45] Freeman, E.A., Moisen, G.G. (2008), A comparison of the performances of threshold criteria for binary classification in terms of predicted prevalence and kappa, *Ecological Modeling*, **217**, 58–58

[46] Gallant, A.R. (1987), Nonlinear Statistical Models, Wiley, New York

[47] Gamerman, D. (1997): Markov Chain Monte Carlo, Chapman & Hall.

[48] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2006), Bayesian Data Analysis, Chapman & Hall/CRC Texts in Statistical Science, 2nd edition

[49] Gelman, A., Hill J. (2007), Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, Cambridge, 2nd edition

[50] Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, **1**, 3, 515–533

[51] Goldstein H. (1989), Restricted unbiased iterative generalized least-squares estimation, *Biometrika*, **76**, 3, 622–623

[52] Goldstein H. (1991), Nonlinear Multilevel Models, with an Application to Discrete Response Data, *Biometrika*, **78**, 1, 45–51

[53] Goldstein H., Rabash J. (1996), Improved Approximations for Multilevel Models with Binary Response, *Journal of the Royal Statistical Society A*, **159**, 3, 505–513

[54] Goldstein H., Spiegelhalter D.J. (1996), League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance, *Journal of the Royal Statistical Society A*, **159**, 3, 385–443

[55] Goldstein, H. (2011), Multilevel Statistical Models, Wiley Series in Probability and Statistics, 4th edition

[56] Grieco, N., Sesana, G., Corrada, E., Ieva, F., Paganoni, A.M., Marzegalli, M. (2007), The Milano Network for Acute Coronary Syndromes and Emergency Services, *MESPE journal*, First Special Issue

[57] Grieco, N., Corrada, E., Sesana, G., Fontana, G., Lombardi, F., Ieva, F., Paganoni, A.M., Marzegalli, M. (2008), Predictors of reduction of treatment time for ST-segment elevation myocardial infarction in a complex urban reality. The MoMi2 survey, [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/10-2008.pdf

[58] Grieco, N., Sesana, G., Ieva, F., Marzegalli, M., Paganoni A.M. (2008), Door to Balloon Time in Patients with ST-Segment Elevation Myocardial Infarction, *Acts* of XLIV Scientific Meeting of the Italian Statistical Society 2008, Arcavacata di Rende, Cosenza (Italy), June 25-27, 2008

[59] Grieco, N., Corrada, E., Sesana, G., Fontana, G., Lombardi, F., Ieva, F., Paganoni, A.M., Marzegalli, M. (2008), Le reti dell'emergenza in cardiologia : l'esperienza lombarda, *Giornale Italiano di Cardiologia*, Supplemento "Crema Cardiologia 2008. Nuove Prospettive in Cardiologia"

[60] Grieco, N., Corrada, E., Sesana, G., Lombardi, F., Paganoni, A.M., Ieva, F., Marzegalli, M. (2008), On site ECG transmission reduces door-to-balloon time in patients referred for primary PCI, *European Heart Journal*, **29** (Abstract Supplement), 655–656

[61] Grieco, N., Sesana, G., Paganoni, A.M., Ieva, F., Marzegalli, M. (2008), Door to Balloon time in patients with ST-segment elevation myocardial infarction. A study in a complex urban reality, *Acts* of IX Congresso SIMAI, Roma (Italy), September 15-19, 2008

[62] Grieco, N., Ieva, F., Paganoni, A.M. (2011), Performance assessment using mixed effects models: a case study on coronary patient care, *IMA Journal of Management Mathematics*. In press. [Online] http://imaman.oxfordjournals.org/content/early/2011/05/27/imaman.dpr007

[63] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2010), A hierarchical random-effects model for survival in patients with Acute Myocardial Infarction, *Acts* of XLV Scientific Meeting of the Italian Statistical Society 2010, Padua (Italy), June 16-18, 2010

[64] Guglielmi, A., Ieva, F., Paganoni, A.M., Soriano, J, Ruggeri, F. (2011) Semiparametric Bayesian modeling for the classification of patients with high observed survival probabilites, *In progress*

[65] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2011), Hospital clustering in the treatment of acute myocardial infarction patients via a Bayesian nonparametric approach, *submitted*

[66] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2012), A Bayesian random-effects model for survival probabilities after Acute Myocardial Infarction, *Chilean Journal of Statistics*. In press. [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/17-2010.pdf

[67] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2012), Process indicators and outcome measures in the treatment of Acute Myocardial Infarction patients, *Statistical Methods in Healthcare*, Wiley (2011). Editors: Faltin, F., Kenett, R., Ruggeri, F.

[68] Harding M., Hausman J., (2007), Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Non-linear Least Squares, *International Economic Review*, **48**, 4, 1311–1328

[69] Hartigan, J.A., Wong, M.A. (1979), A k-means clustering algorithm, *Applied Statistics*, **28**, 100–108

[70] He, X., Pin, T. (1999), COBS: Qualitatively constrained smoothing via linear programming, *Computational Statistics*, **14**, 315–337

[71] Hobert, J.P., Casella, G. (1996), The effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models, *Journal of the American Statistical Association*, **91**, 436, 1461–1473

[72] Hoff, P.D. (2009), A first course in Bayesian Statistical Methods, Springer Texts in Statistics

[73] Hox, J. (2002), Multilevel Analysis, Techniques and Applications, Lawrence Erlbaum Associates, New Jersey

[74] Ieva, F. (2008), Modelli statistici per lo studio dei tempi di intervento nell'Infarto Miocardico Acuto, *Master Thesis*. Available at http://mox.polimi.it/it/informazioni/personale/viewpers.php?id=91&en=en&tesi=on

[75] Ieva, F., Paganoni, A.M. (2009), Statistical Analysis of an Integrated Database Concerning Patients With Acute Coronary Syndromes, *Proceedings* of SCo2009, Sixth Conference, Maggioli editore, Milano

[76] Ieva, F., Paganoni, A.M. (2009), Integrazione tra registri clinici database amministrativi: il progetto IMASTE della Regione Lombardia, *SIS - Magazine* [Online] http://www.sis-statistica.it/magazine/spip.php?article161

[77] Ieva, F., Paganoni, A.M., Secchi, P. (2010), Mining Administrative Health Databases for epidemiological purposes: a case study on Acute Myocardial Infarctions diagnoses, *Submitted*, [Online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/45-2010.pdf

[78] Ieva, F., Paganoni, A.M. (2010), Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI2 survey, *Communications in Applied and Industrial Mathematics*, **1**, 1, 128-147 [Online] http://cab.unime.it/journals/index.php/caim/article/view/2010CAIM477

[79] Ieva, F., Paganoni, A.M., Secchi, P. (2010), Data mining the Lombardia Public Health Database: a pilot case study on hospital discharge data for Acute Myocardial Infarctions, *Acts* of XLV Scientific Meeting of the Italian Statistical Society 2010, Padua (Italy), June 16-18, 2010.

[80] Ieva, F., Paganoni, A.M., Pigoli, D., Vitelli, V. (2011), Multivariate functional clustering for the analysis of ECG curves morphology. *Submitted* [Online] http://www1.mate.polimi.it/biblioteca/qddview.php?id=1434&L=i

[81] Ieva, F. (2011), Outlier detection for training sets in an unsupervised functional classification framework: an application to ECG signals, *Proceedings* of the 17th European Young Statisticians Meeting, Lisboa (Portugal), September 5-9, 2011.

[82] Ieva, F., Pagnoni, A.M. (2011), Depth Measures for Multivariate Functional Data, *Submitted*

[83] Ieva, F., Paganoni, A.M. (2011), Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes, [Online] http://www1.mate.polimi.it/biblioteca/qddview.php?id=1443&L=i

[84] Ieva, F., Paganoni, A.M. (2011), Process Indicators for Assessing Quality of Hospital Care: a case study on STEMI patients, *JP Journal of Biostatistics*, **6**, 1, 53–75

[85] Ieva, F, Paganoni, A.M. (2011), Reportistica Archivio STEMI, Rapporto Tecnico di fine Programma Strategico

[86] Inmon, W.H. (1996), Building the Data Warehouse, John Wiley & Sons, second edition.

[87] Ishwaran, H. and Zarepour, M.(2000). Exact and approximate sum representations for the Dirichlet process, *Canadian Journal of Statistics*, **30**, 269–283

[88] Jackman, S. (2009), Bayesian Analysis for the Social Sciences, Wiley

[89] Kalbfleisch, John D. and Prentice, Ross L. (1980), The statistical analysis of failure data, John Wiley & Sons, New York

[90] Kaufman, L., Rousseeuw, P. (1990), Finding Groups in Data. Wiley Series in Probability and Mathematical Statistics

[91] Kelinman, K.P., Ibrahim, J.G. (1998), A Semi-parametric Bayesian Approach to the Random-Effects Models, *Biometrics*, **54**, 265–278

[92] Kelinman, K.P., Ibrahim, J.G. (1998), A Semi-parametric Bayesian Approach to Generalized Linear Mixed Models, *Statistics in Medicine*, **17**, 2579–2596

[93] Kiefer, J. and Wolfowitz, J. (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Annals of Mathematical Statistics*, **27**, 887–906

[94] Kuhn, E. and Lavielle, M. (2005), Maximum Likelihood estimation in nonlinear mixed effect models, *Computational Statistics and Data Analysis* **49**, 4, 1020–1038

[95] Lai, T.L. and Shih, M.C. (2003), Nonparametric estimation in nonlinear mixed-effects models, *Biometrika*, **90**, 1, 1–13

[96] Laird, N.M. (1978), Nonparametric maximum likelihood estimation of a mixing distribution, *Journal of the American Statistical Association*, **73**, 805–811.

[97] Laird N.M., Ware J.H. (1982), Random-Effects Models for Longitudinal Data, *Biometrics*, **38**, 4, 963–974

[98] Lee Y., Nelder J.A. (1996), Hierarchical Generalized Linear Models, *Journal of the Royal Statistical Society B*, **58**, 4, 619–678

[99] Li, J. and Liu, R. (2004), New Nonparametric Tests of Multivariate Locations and Scales using Data Depth, *Statistical Science*, **19**, 686–696

[100] Lindsay, B.G. (1983), The geometry of mixture likelihoods: a general theory, *The Annals of Statistics*. **11**, 1, 86–94

[101] Lindstrom, M.J. and Bates, D.M. (1990), Nonlinear Mixed Effects Models for Repeated Measures Data, *Biometrics*, **46**, 673–687

[102] Liu, R. (1990), On a Notion of Data Depth based on Random Simplices, *The Annals of Statistics*, **18**, 405–414

[103] Liu, X., and Müller, H.-G. (2003), Modes and clustering for time–warped gene expression profile data *Bioinformatics*, **19**, 15, 1937–1944.

[104] Liu, X., and Yang, M. (2009), Simultaneous curve registration and clustering for functional data *Computational Statistics and Data Analysis*, **53**, 1361–1376

[105] Lopez-Pintado, S., Romo, J. (2003), Depth-based classification for functional data, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*.

[106] Lopez-Pintado, S., Romo, J. (2009), On the Concept of Depth for Functional Data, *Journal of the American Statistical Association*, **104**, 486, 718–734

[107] MacEachern, S.N. (1994), Estimating normal means with a conjugate style Dirichlet process prior, *Communications in Statistics*, **23**, 727–741

[108] MacEachern, S.N. (2000). Dependent Dirichlet processes. *Technical Report*, Department of Statistics, The Ohio State University

[109] Mahalanobis, P.C. (1936), On the generalized distance in statistics, *Proceedings of National Academy of Science of India*, **12**, 486, 49–55

[110] Mallat, S. (1999), A Wavelet Tour of Signal Processing, Accademic Press

[111] Mallet, A. (1986), A Maximum Likelihood method for random coefficient regression models, *Biometrika*. **73**, 3, 645–656

[112] Marzegalli, M., Tridico, C., Fontana, G., Borghi, G., Grieco, N., Ieva, F., Paganoni, A.M. (2009), Integrazione tra registri clinici e database amministrativi: il progetto IMASTE Lombardia, *Acts* of "Cardiologia 2009" of XVIII International conference of Cardiovascular Department A. De Gasperis, Milan (Italy), June 16-18, 27–31

[113] Müller, H.G., Stadtmüller, U. (2005), Generalized Functional Linear Models, *Annals of Statistics*, **33**, 2, 774–805

[114] Muller, P., Quintana, F.A. (2004), Nonparametric Bayesian Data Analysis, *Statistical Science*, **19**, 1, 95–110

[115] Normand, S.T., Glickman, M.E., Gatsonis, C.A. (1997), Statistical methods for profiling providers of medical care: issues and applications, *Journal of the American Statistical Association*, **92**, 803–814

[116] Normand, S.T., Shahian, D.M. (2007), Statistical and Clinical Aspects of Hospital Outcomes Profiling, *Statistical Science*, **22**, 2, 206–226

[117] Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J. (2007), A hierarchical modelling framework for identifying unusual performance providers, *Journal of the Royal Statistical Society A*, **170**, 4, 865–890

[118] Peña, E.A., Hollander, M. (2004), Models for recurrent events in reliability and survival analysis, *Mathematical Reliability: An Expository Perspective*, **6**, 493–514. Editors: Soyer, R. and Mazzuchi, T. and Singpurwalla, N., Kluwer Academic Publishers,

[119] Peña, E.A., Slate, E.H., González, J.R. (2007), Semiparametric inference for a general class of models for recurrent events, *Journal of Statistical Plannning and Inference*, **137**, 6, 1727–1747

[120] Pigoli, D., Sangalli, L.M. (2011), Wavelets in Functional Data Analysis: estimation of multidimensional curves and their derivatives, *Computational Statistics and Data Analysis*. In press

[121] Pinheiro, J.C. (1994), Topics in Mixed Effects Models, *PhD Thesis*

[122] Pinheiro, J.C., Bates, D.M. (2000), Mixed-Effects Models in S and S-plus, Springer-Verlag, New York

[123] Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 20–22

[124] R Development Core Team (2009), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] http://www.R-project.org

[125] Racz, J., Sedransk, J. (2010), Bayesian and Frequentist Methods for Provider Profiling Using Risk-Adjusted Assessments of Medical Outcomes, *Journal of the American Statistical Association*, **105**, 489, 48–58

[126] Ramsay, J.O., Silverman, B.W. (2005), Functional data analysis, Springer Science and Business Media, New York

[127] Ripley, B.D. (2008), Stochastic Simulation, John Wiley & Sons, Inc., Hoboken, New Jersey

[128] Robert,C.P., Casella, G. (2004), Monte Carlo Statistical Methods (second edition), Springer-Verlag, New York

[129] Schumitzky, A. (1991), Nonparametric EM Algorithms for estimating prior distributions, *Applied Mathematics and Computation*, **45**, 2, 143–157.

[130] Serfling, R. (2004), Depth Functions in Nonparametric Multivariate Inference, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*

[131] Sethuraman, J. (1994), A constructive definition of Dirichlet Process, *Statistica Sinica*, **4**, 639–650

[132] Sheiner, L.B., Beal, S.L. (1980), Evaluation of methods for estimating population pharmacokinetic parameters. III. Monoexponential model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Pharmacodynamics*. **11**, 3, 303–319.

[133] Spiegelhalter, D.J., Sharples, L.D., Ohlssen, D.I. (2007), A hierarchical modelling framework for identifying unusual performance in health care providers, *Journal of the Royal Statistical Society A*, **170**, 4, 865–890

[134] Steele F., Zhang W. (2004), A Semiparametric Multilevel Survival Model, *Applied Statistics*, **53**, 2, 387–404

[135] Steele, F. (2008), Multilevel models for longitudinal data, *Journal of the Royal Statistical Society A*, **171**, 1, 5–19

[136] Struyf, A., Hubert, M., and Rousseeuw, P. (1997), Clustering in an Object-Oriented Environment, *Journal of Statistical Software*, **1**, 4, 1–30

[137] Sun, Y., Genton, M.G. (2011), Functional Boxplots, *Journal of Computation and Graphical Statistics*, to appear

[138] Tarpey, T., and Kinateder, K. K. J. (2003), Clustering Functional Data, *Journal of Classification*, **20**, 93–114

[139] Tierney, L., Kass, R.E., Kadane, J.B. (1989), Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions, *Journal of the American Statistical Association*, **84**, 407

[140] Tierney, L. (1994), Markov Chains for exploring posterior distributions, *The Annals of Statistics*, **22**, 4, 1701–1762

[141] Tukey, J. (1975), Mathematics and Picturing Data, *Proceedings* of the 1975 International Congress of Mathematic, **2**, 523–531

[142] Verbeke, G., Molenberghs, G. (2000), Linear Mixed Models for Longitudinal Data, Springer-Verlag, New York

[143] Verbeke, G., Molenberghs, G. (2005), Models for Discrete Longitudinal Data, Springer, New York

[144] Vermunt, J.K. (2004), An EM Algorithm for the Estimation of Parametric and Nonparametric Hierarchical Models. *Statistica Neerlandica* **58**, 2, 220–233

[145] Walker, S. (1996), An EM Algorithm for Nonlinear Random Effects Models. *Biometrics* **52**, 3, 934–944

[146] Wolfinger, R. (1993), Laplace's approximation for nonlinear mixed models, *Biometrika* **80**, 4, 791–795

[147] Wood, S.N. (2006), Generalized Additive Models: An Introduction with R, Chapman & Hall/CRC

[148] Zeger, S.L., Karim, M.R. (1991), Generalized Linear Models with Random Effects: a Gibbs sampling approach, *Journal of American Statistical Association*, **86**, 79–86

[149] Zuo, Y.Z., Serfling, R. (2000), General Notions of Statistical Depth Function, *The Annals of Statistics*, **28**, 2, 461–482

[150] Zuo,Y. (2003), Projection Based Depth Functions and Associated Medians, *The Annals of Statistics*, **31**, 2, 1460–1490

Clinical references:

[151] Antman, E.M., Hand, M., Armstrong, P.W., Bates, E.R., Green, L.A., Halasyamani, L.K., Hochman, J.S., Krumholz, H.M., Lamas, G.A., Mullany, C.J., Pearle, D.L., Sloan, M.A. and Smith, S.C. Jr. (2008), Update of the ACC/AHA 2004 guidelines for the management of patients with ST-elevation myocardial infarction, *Circulation*, **117**, 296–329

[152] Avezum, A., Makdisse, M., Spencer, F., Gore, J.M., Montalescot G, *et al.*, for the GRACE Investigators (2005), Impact of age on management and outcome of acute coronary syndrome: observations from the Global Registry of Acute Coronary Events (GRACE), *American Heart Journal*, **149**, 67–73

[153] Balzi, D., Barchielli, A., Battistella, G. et al. (2008), Stima della prevalenza della cardiopatia ischemica basata su dati sanitari correnti mediante un algoritmo comune in differenti aree italiane, *Epidemiologia e Prevenzione*, **32**, 3, 22–29

[154] Barendregt, J.J., Van Oortmarssen, J.G., Vos, T. et al. (2003), A generic model for the assessment of disease epidemiology: the computational basis of DisMod II, *Population Health Metrics*, **1**

[155] Bird, S.M., Cox, D., Farewell, V.T., Goldstein, H., Holt, T., Smith, P.C. (2005), Performance indicators: good, bad and ugly. The report of a Working Party on Performance Monitoring in the Public Services, *Journal of the Royal Statistical Society A*, **168**, 1–27

[156] Bradley, E.H., Herrin, J., Wang, Y., Barton, B.A., Webster, T.R., Mattera, J.A., Roumanis, S.A., Curtis, J.P., Nallamothu, B.K., Magid, D.J., McNamara, R.L., Parkosewich, J., Loeb, J.M. and Krumholz, H.M. (2006), Strategies for reducing the door-to-balloon time in acute myocardial infarction, *New England Journal of Medicine*, **355**, 2308–2320

[157] Brodie, B.R., Stuckey, T.D., Muncy, D.B., Hansen, C.J., Wall, T.C., Pulsipher, M. and Gupta, N. (2003), Importance of time-to-reperfusion in patients with acute myocardial infarction with and without cardiogenic shock treated with primary percutaneous coronary intervention, *American Heart Journal*, **145**, 4, 708–715

[158] Cannon, C.P., Gibson, C.M., Lambrew, C.T., Shoultz, D.A., Levy, D., French, W.J., Gore, J.M., Weaver, W.D., Rogers, W.J. and Tiefenbrunn, A.J. (2000), Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction, *Journal of American Medical Association*, **283**, 22, 2941–2947

[159] Chen J., Normand S.T., Wang Y., Drye E.E., Schreiner G.C., Krumholz H.M. (2010), Recent Declines in Hospitalizations for Acute Myocardial Infarction for Medicare Fee-for-Service Beneficiaries Progress and Continuing Challenges, *Circulation*, **121**, 1322-1328

[160] Dalby, M., Bouzamondo, A., Lechat, P. and Montalescot, G. (2003), Transfer for primary angioplasty versus immediate thrombolysis in acute myocardial infarction: a metaanalysis, *Circulation*, **108**, 1809–1814

[161] Data Mart Rete IMA: ETL, Data model e Reportistica, Lombardia Informatica S.p.A., Direzione Sviluppo - Area Sistemi Direzionali SISS

[162] de Lemos, J.A., Braunwald, E. (2001), ST Segment Resolution as a Tool for Assessing the Efficacy of Reperfusion Therapy, *Journal of the American College of Cardiology*, **38**, 5, 1283–1294

[163] Devlin, G.P., Anderson, F.A., Heald, S., Lòpez-Sendòn, J., Avezum, A., Elliott, J., Dabbous, O.H., Brieger D. (2005), Management and outcomes of lower-risk patients presenting with acute coronary syndromes in a multinational observational registry, *Heart* doi:10.1136/hrt.2004.05400

[164] Diercks, D.B., Kontos, M.C., Chen, A.Y., Pollack, C.V. *et al.* (2009), Utilization and impact of pre-hospital electrocardiograms for patients with acute ST-segment elevation myocardial infarction: data from NCDR (National Cardiovascular Data Registry) ACTION (acute coronary treatment and intervention outcome network), *The American College of Cardiology*, **53**, 161–166

[165] Eagle, K.A., Goodman, S.G., Avezum, A., Budaj, A., Sullivan, C.M., Lòpez-Sendòn, J., for the GRACE Investigators (2002), Practice variation and missed opportunities for reperfusion in ST-segment-elevation myocardial infarction: findings from the Global Registry of Acute Coronary Events (GRACE), *Lancet*, **359**, 373–377

[166] Einthoven, W. (1908), Weiteres über das Elektrokardiogram, *Pflüger Archiv: European Journal of Physiology*, **122**, 517–548.

[167] Einthoven, W., Fahr, G. and de Waart, A. (1950), On the direction and manifest size of the variations of potential in the human heart and on the influence of the position of the heart on the form of the electrocardiogram. *American Heart Journal*, **40**, 2, 163–211.

[168] Every, N.R., Frederick, P.D., Robinson, M. et al. (1999), A Comparison of the National Registry of Myocardial Infarction With the Cooperative Cardiovascular Project, *Journal of the American College of Cardiology*, **33**, 7, 1887–1894

[169] Federazione Italiana di Cardiologia, Società Italiana di Cardiologia Invasiva (2005), Documento di Consenso La rete interospedaliera per l'emergenza coronarica, *Italian Heart Journal*, **6**, (Suppl 6), 5–26

[170] Fine L.G., Keogh B.E., Cretin S., Orlando M., Gould M.M. (2003), How to evaluate and improve the quality and credibility of an outcomes database: validation and feedback study on the UK Cardiac Surgery Experience, *British Medical Journal*, **326**, 4, 25-28

[171] Fox, K.A., Goodman, S.G., Klein, W. *et al.* (2002), Management of acute coronary syndromes. Variations in practice and outcome: findings from the Global Registry of Acute Coronary Events (GRACE), *European Heart Journal*, **23**, 1177–1189

[172] Gersh, B.J., Stone, G.W., White, H.D., Holmes, D.R. (2005), Pharmacological Facilitation of Primary Percutaneous Coronary Intervention for Acute Myocardial Infarction: Is the Slope of the Curve the Shape of the Future? *Journal of American Medical Association*, **293**, 8, 979–986

[173] Glance, L.G., Osler, T.M., Mukamel, D.B. et al. (2008), Impact of the present-on-admission indicator on hospital quality measurement experience with the Agency for Healthcare Research and Qualità (AHRQ) Inpatient Quality Indicators, *Medical Care*, **46**, 2, 112–119

[174] Goldberger, E. (1942), The aVL, aVR, and aVF leads: a simplification of standard lead electrocardiography. *American Heart Journal*, **24**, 378–396.

[175] Goldberger, E. (1942), A simple indifferent electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar extremity leads. *American Heart Journal*, **23**, 483–492.

[176] Hanratty, R., Estacio, R.O., Dickinson L.M., et al. (2008), Testing Electronic Algorithms to create Disease Registries in a Safety Net System, *Journal of Health Care Poor Underserved*, **19**, 2, 452–465

[177] Hasday, D., Behar, S., Wallentin, L. (2002), A prospective survey of the characteristics, treatments and outcomes of patients with acute coronary syndromes in Europe and the Mediterranean basin. The euro heat survey of acute coronary syndromes (euro heart survey ACS), *European Heart Journal*, **23**, 1190–1210

[178] Hatry, H.P. (1999), Performance Measurement Getting Results, Washington D.C.: The Urban Institute Press, 1999. ISBN: 0-87766-692-X.

[179] Hougaard, P. (1984), Life table methods for heterogeneous populations: Distributions describing the heterogeneity, *Biometrika*, **71**, 75–83

[180] Huber, K., De Caterina, R., Kristensen, S.D., Verheugt, F.W.A., Montalescot, G., Badimon Maestro, L. *et al.* (2005), Pre-hospital reperfusion therapy: a strategy to improve therapeutic outcome in patients with ST-elevation myocardial infarction, *European Hearth Journal*, **26**, 2063–2074

[181] Iezzoni, L.I., Hotchkin, E.K., Ash, A.S., Shwartz, M., Mackiernan, Y., MedisGroups Data Bases (1993), The impact of data collection guidelines on predicting in-hospital mortality, *Medical Care*, **31**, 277–283

[182] Jneid, H., Fonarow, G.C., Cannon, C.P., Palacios, I.F., Kilic, T., Moukarbel, G.V., Maree, A.O., LaBresh, K.A., Liang, L., Newby, L.K., Fletcher, G., Wexler, L. and Peterson, E. (2008), Impact of time of presentation on the care and outcomes of acute myocardial infarction, *Circulation*, **117**, 2502–2509.

[183] Kusek, J.Z., Rist, R.C. (2004), Ten Steps to a Results-Based Monitoring and Evaluation System, The World Bank, Washington, D.C.

[184] Lichiello, P. (1999), Guidebook for Performance Measurement, Turning Point

[185] Lindsay, A.E. (2006), ECG learning centre, [online] http://library.med.utah.edu/kw/ecg/-index.html

[186] Manuel, D.G., Lim, J.J.Y., Tanuseputro, P. et al. (2007), How many people have a myocardial infarction? Prevalence estimated using historical hospital data, *BMC Public Health*, **7**, 174–89

[187] Mason, R., Likar, L. (1966), A new system of multiple leads exercise electrocardiography, *American Heart Journal*, **71**, 2, 196–205.

[188] Masoudi F.A., Bonow R.O., Brindis R.G., Cannon C.P., DeBuhr J., Fitzgerald S., Heidenreich P.A., Ho K.K.L., Krumholz H.M., Leber C., Magid D.J., Nilasena D.S., Rumsfeld J.S., Smith S.C. Jr, Wharton T.P. Jr. (2008), ACC/AHA 2008 statement on performance measurement and reperfusion therapy: a report of the ACC/AHA Task Force on Performance Measures (Work Group to Address the Challenges of Performance Measurement and Reperfusion Therapy), *Circulation*, **118**

[189] McNamara, R.L., Wang, Y., Herrin, J., Curtis, J.P., Bradley, E.H., Magid, D.J., Peterson, E.D., Blaney, M., Frederick, P.D. and Krumholz, H.M. (2006), Effect of door-toballoon time on mortality in patients with ST-segment elevation myocardial infarction, *Journal of American College of Cardiology*, **47**, 11, 2180–2186

[190] Mehta, R.H., Sadiq, I., Goldberg, R.J., Gore, J.M., Avezum, A., Spencer, F., Kline-Rogers, E., Allegrone, J., Pieper, K., Fox, K.A.A., Eagle, K.A., for the GRACE Investigators (2004). Effectiveness of primary percutaneous coronary intervention compared with that of thrombolytic therapy in elderly patients with acute myocardial infarction, *American Heart Journal*, **147**, 253–259

[191] Milakovich, M.E., Gordon, G.J. (2001), Public Administration In America, Boston: Bedford/St Martin's, 2001. ISBN: 0-312-24972-1.

[192] Montalescot, G., Dabbous, O.H., Lim, M.J., Flather, M.D., Mehta, R.H., for the GRACE Investigators (2005), Relation of timing of cardiac catheterization to outcomes in patients with non-ST-segment elevation myocardial infarction or unstable angina pectoris enrolled in the multinational Global Registry of Acute Coronary Events (GRACE), *American Journal of Cardiology*, **95**, 1397–1403

[193] Moscucci, M., Fox, K.A.A., Cannon, C.P., Klein, W., Lòpez-Sendòn, J., Montalescot, G., White, K., Goldberg, R.J., for the GRACE Investigators (2003), Predictors of major bleeding in acute coronary syndromes: the Global Registry of Acute Coronary Events (GRACE), *European Heart Journal*, **24**, 1815–23

[194] Oltrona, L., Mafrici, A., Marzegalli, M., Fiorentini, C., Pirola, R., Vincenti, A. (2005), The early management of ST-elevation acute myocardial infarction in the Lombardy Region (GestIMA), Studio GestIMA e della Sezione Regionale Lombarda dell'ANMCO e della SIC, *Italian Heart Journal Supplement*, **6**, 8, 489–97

[195] Piano Socio Sanitario Regionale 20010-2014. [online] www.SSOSA.com/lombardia.pdf

[196] Politi, A., Martinoni, A., Klugmann, S., Zanini, R., Onofri, M., Guagliumi, G., Fiorentini, C., Lettieri, C., Belli, G., Piccaluga, E., De Cesare, N., D'Urbano, M., Ettori, F., Repetto, A., Musumeci, G., Castiglioni, B., Colombo, P., Passamonti, E., Bramucci, E., Cattaneo, L., Ferrari, G., Repetto, S., Bartorelli, A., Pirelli, S., De Servi, S., LombardIMA Study Group (2011) LombardIMA: a regional registry for coronary angioplasty in ST-elevation myocardial infarction, *Journal of Cardiovascular Medicine*, **12**, 1, 43–50

[197] Rathore, S.S., Curtis, J.P., Chen, J., Wang, Y., Nallamothu, B.K., Epstein, A.J. and Krumholz, H.M. (2009), Association of door-to-balloon time and mortality in patients admitted to hospital with ST elevation myocardial infarction: national cohort study, *British Medical Journal*, **338**

[198] Rokos, I.C., French, W.J., Mattu, A., Nichol, G., Farkout, M.F., Reiffel, J., *et al.* (2010), Appropriate cardiac cath lab activation:optimizing electrocardiogram interpretation and clinical decision-making for acute ST-elevtion myocrdial infarction, *American Heart Journal*, **160**, 995–1003

[199] Rutman, L., Mowbray, G. (1983), Understanding Program Evaluation, Beverly Hiss/London/New Delhi: Sage Publications, 1983. ISBN: 0-8039-2093-8.

[200] Saia, F., Marrozzini, C., Guastaroba, P., Ortolani, P. Palmerini, T., Pavesi, P.C. *et al.* (2010), Lower long-term mortality within a regional system of care for ST-elevation myocardial infarction, *Acute Cardiac Care*, **12**, 42–50

[201] Scher, A.M. and Young, A.C. (1957), Ventricular depolarization and the genesis of the QRS, *Annals of New York Academy of Science*, **65**, 768–78.

[202] Schiele F., Hochadel M., Tubaro M., Meneveau N., Wojakowski W., Gierlotka M., Polonski L., Bassand J.P., Fox K.A.A., Gitt A.K. (2010), Reperfusion strategy in Europe: temporal trends in performance measures for reperfusion therapy in ST-elevation myocardial infarction, *European Heart Journal*, **31**, 2614–2624

[203] Schroder R. (2004) Prognostic impact of early ST-segment resolution in acute STelevation myocardial infarction, *Circulation*, **110**, 506–510

[204] Sejerstein, M., Sillsen, M., Hansen, P.R., Nielsen, S.L., Nielsen, H., Trutner, S. *et al.* (2008), Effect on treatment delay of prehospital tele-transmission of 12-lead electrocardiogram to a cardiologist for immediate triage and direct referral of patient with ST-segment elevation acute myocardial infarction to primary percutaneous coronary intervention, *American Journal of Cardiology*, **101**, 941–946

[205] Sibley, L.M., Moineddin, R., Agham,M.M. et al. (2010), Risk Adjustment Using Administrative Data-Based and Survey-Derived Methods for Explaining Physician Utilization, *Medical Care*, **48**, 175–182

[206] Silber, J.H., Rosenbaum, P.R., Williams, S.V., Ross, R.N., Schwartz, J.S. (1997), The Relationship Between Choice of Outcome Measure and Hospital Rank in General Surgical Procedures: Implications for Quality Assessment, *International Journal for Quality in Health Care*, **9**, 3, 193–200

[207] Spertus, J.A., Radford, M.J., Every, N.R. *et al.* (2003), Challenges and opportunities in quantifying the quality of care for acute myocardial infarction: summary from the Acute Myocardial Infarction Working Group of the American Heart Association/American College of Cardiology First Scientific Forum on Quality of Care and Outcomes Research in Cardiovascular Disease and Stroke, *Circulation*, **107**, 1681–1691

[208] Spertus, J.A., Eagle, K.A., Krumholz, H.M. *et al.* (2005), American College of Cardiology and American Heart Association methodology for the selection and creation of performance measures for quantifying the quality of cardiovascular care, *Circulation*, **111**, 1703–1712

[209] Studnek J.R., Garvey L., Blackwell T., Vandeventer S., Ward S.R. (2010), Association Between Prehospital Time Intervals and ST-Elevation Myocardial Infarction System Performance, *Circulation*, **122**, 1464–1469

[210] Thygesen, K., Alpert, J.S., White, H.D. (2007), Universal definition of myocardial infarction, *European Heart Journal*, **28**, 20, 2525–2538

[211] Ting, H.H., Krumholz, H.M., Bradley, E.H., Cone, D.C., Curtis, J.P., Drew, B.J., Field, J.M., French, W.J., Gibler, W.B., Goff, D.C., Jacobs, A.K., Nallamothu, B.K., O'Connor, R.E. and Schuur, J.D. (2008), Implementation and integration of prehospital ECGs into systems of care for acute coronary syndrome, *Circulation*, **118**, 1066–1079

[212] Tu, J.V., Khalid, L., Donovan, L.R., Ko, D.T. for the Canadian Cardiovascular Outcomes Research Team/Canadian Cardiovascular Society Acute Myocardial Infarction Quality Indicator Panel (2008), Indicators of quality of care for patients with acute myocardial infarction, *Canadian Medical Association Journal*, **179**, 9, 909–915

[213] Tubaro, M., Danchin, N., Goldstein, P., Van De Werf, F. *et al.* (2011), Pre-hospital treatment of STEMI patients. A scientific statement of the Working Group Acute Cardiac Care of the European Society of Cardiology, *Acute Cardiac Care*, **13**, 56–67

[214] Widimsky, P., Wijns W., Fajadet, J., de Belder, M., Knot, J., Aaberge, L., Andrikopoulos, G., Baz, J.A., Betriu, A., Claeys, M., Danchin, N., Djambazov, S., Erne, P *et al.* (2010), Reperfusion therapy for ST elevation acute myocardial infarction in Europe: description of the current situation in 30 countries, *European Heart Journal*, **31**, 943–957

[215] Williams, B., Dowell, J., Humphris, G., Themessl-Huber, M., Rushmer, R., Ricketts, I., Boyle, P., Sullivan, F. (2010), Developing a longitudinal database of routinely recorded primary care consultations linked to service use and outcome data, *Social Science and Medicine*, **70**, 473–478

[216] Wilson, F.N., Johnston, F.D., Rosenbaum, F.F., Erlanger, H., Kossmann, C.E., Hecht, H., Cotrim, N., Menezes de Olivieira, R., Scarsi, R., Barker, P.S. (1944), The precordial electrocardiogram, *American Heart Journal*, **27**, 19–85.

[217] White, H.D., Chew, D.P. (2008), Acute myocardial infarction, *Lancet*, **372**, 570–584

[218] Wirehn, A.B., Karlsson, H.M., Cartensen J.M., et al. (2007), Estimating Disease Prevalence using a population-based administrative healthcare database, *Scandinavian Journal of Public Health*, **35**, 424–431

[219] Wiviott, S.D., Antman, E.M., Gibson, C.M., Montalescot, G., *et al.* (2006), Evaluation of prasugrel compared with clopidogrel in patients with acute coronary syndromes: design and rationale for the Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition With Prasugrel Thrombolysis In Myocardial Infarction (TRITON-TIMI 38), *American Heart Journal*, **152**, 627–635

[220] Yeh, R.W., Sidney, S., Chandra, M., Sorel, M., Selby, J.V., Go, A.S. (2010), Population Trends in the Incidence and Outcomes of Acute Myocardial Infarction, *The New England Journal of Medicine*, **362**, 23

Ringraziamenti

Sono ben consapevole che queste righe non potranno mai esprimere completamente la gratitudine che provo, né contenere tutti coloro a cui vorrei rivolgerla. Mai avrei pensato che il tempo di questi tre anni potesse contenere tanto. Tanta vita, tante esperienze, tante persone. Tra tutte quelle a cui rivolgo il mio sincero GRAZIE, un pensiero particolare a

Nico e al ***Dott. Marzegalli***, per avermi sempre ricordato con l'esempio la differenza che intercorre tra lavorare e spendersi con passione per una causa. Per tutto quello che mi hanno insegnato, per il bene che mi vogliono, e perchè nessuno è più titolato di loro a parlare di "cuore".

Anna, per cui non ci saranno mai parole adeguate ad esprimere la mia stima. So solo che senza di lei nulla sarebbe stato, e che per tutto ciò che ha fatto per me non basterebbe un'altra tesi (la quarta!) di ringraziamenti. Perchè è un gigante di umanità prima ancora che una grande insegnante. Per la pazienza, la perenne disponibilità, la professionalità, l'amicizia; e soprattutto, per l'esempio che rappresenta nei confronti di chiunque abbia la fortuna di lavorare con lei.

Piercesare, che con il suo mettere in discussione tutto, mi ha sempre obbligata a confrontarmi con i miei limiti, spesso aprendomi nuovi orizzonti e regalandomi occhi diversi con cui guardare alle cose. Per la capacità che ha di entusiasmare e di rendere tutto un interessante oggetto di riflessione.

Tutti i ***MOXstat***, una famiglia dove ogni giorno è stato possibile crescere professionalmente e umanamente. Grazie ***Laura***, per la presenza costante, sensibile e discreta; ***Simo***, per l'allegria con cui sa ammorbidire ogni cosa; ***Vale***, la "socia" di ogni giorno, viaggio e impresa, amica insostituibile di università, dottorato... e spero molto altro ancora; ***Laura & Ste***, per l'energia e la pacatezza (rispettivamente) che li contraddistinguono; ***Davide***, perchè nessun altro sarebbe riuscito a reggere l'onere di essere quotidianamente il mio compagno di scrivania; ***Paolo*** e ***Andre***, perchè come nessuno sanno farmi ridere... e ovviamente anche ***Bree***!, ***Alessia***, ***Alessandra*** e ***Davide***.

Gli abitanti del ***Laboratorio MOX***, che hanno dato senso e corpo alla parola "condivisione", oltre ad aver migliorato molte delle mie giornate, per il semplice fatto di averle trascorse insieme.

Alessandra, che spesso ho messo a dura prova con scelte di parametri "a cacchio", e ***Fabrizio***, la cui squisita attenzione mi ha sempre accompagnata e guidata nel percorso alla scoperta del mondo Bayesiano.

Paolo, il coordinatore di dottorato che tutti dovrebbero avere. Per lo stile silenzioso, discreto, efficace e concreto con cui in questi 3 anni l'ho visto spendersi per noi dottorandi. E perchè collaborare con lui è stato un divertimento, oltre che un privilegio.

Il mio ***Poli***, che in 4 lettere racchiude un mondo in cui oltre ad infinite opportunità ci sono persone che credono che un futuro migliore si costruisca da un presente migliore, e ci lavorano con entusiasmo.
Un luogo in cui ho avuto l'immensa fortuna di trovare persone capaci di gioire dei miei successi oltre che di costruirli, e di mitigare i momenti difficili: in altre parole insegnanti, colleghi, amici. Sono ancora tanti i nomi che mi vengono in mente scorrendo gli anni, i 7 piani della Nave o gli edifici di piazza Leonardo, ma non è possibile farceli stare tutti. A tutti però la mia gratitudine.

La mia famiglia e le persone che nella mia vita hanno saputo trasmettermi il valore dello studio, della tenacia, della lealtà e delle cose belle. Per avermi insegnato (***papà***) che vale la pena mettere il cuore in tutto ciò che si fa, sempre. Per essere stati (***Chiara*** e ***Chicco***) un motivo per mettercelo.

In particolare un pensiero alla mia ***mamma***, che questo dottorato l'ha visto iniziare ma non finire, ma che non ha smesso un attimo di sostenermi, prima durante e dopo. A lei dedico questa tesi, questi anni, quello che contengono e ciò che hanno significato per me, certa che come nessun altro ne avrebbe capito l'importanza.

... e ***Andre***, che ha colorato ogni giorno degli ultimi 8 anni. Per avermi sempre incoraggiata e protetta, sostenendo con la sua solidità i miei momenti di incertezza. Per non avermi mai permesso di mollare, ma piuttosto costretta a guardare le cose con serena lucidità. Perchè è alla certezza del suo affetto che devo molto dell'entusiasmo con cui affronto la vita. E soprattutto, perchè le nostre strade, sebbene a distanza e in tutta la loro peculiare diversità, non hanno smesso un attimo di correre, parallele, nella medesima direzione.

Grazie a tutti, di cuore
FRA