POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Applying Nonlinear Mixed-Effects Modeling to Model Patient Flow in the Emergency Department

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Umberto Rosamilia

Advisors: PhD Adam Darwich, PhD Jayanth Raghothama, MSc Luca Marzano

Co-advisor: Associate Professor Enrico Gianluca Caiani

Academic year: 2021-2022

1. Introduction

Emergency departments (EDs) strive to provide high-quality 24/7 emergency care to severely ill or injured patients. ED performance and overcrowding affect the functioning of other parts of the hospital and, indirectly, the "*healthcare systems and communities at large*" [1]. Poor performance of the ED and overcrowding lead to delays, prolonged hospitalization, and improper resource allocation, which reduce the quality of the provided care and increase costs. Moreover, these negative consequences can lead to worse patient health outcomes and high admission and re-admission rates or produce adverse effects for the providers, the healthcare system, and the community [1]. Exposing the providers to heavy workload, for instance, hinders timely service provision and clinical decision-making, thus increasing the length of stay (LOS) [1]. This consequence is particularly relevant since longer LOS increases the risk of contracting hospital-acquired infections and is "*associated with higher patient mortality and worse outcomes*" [2].

1.1. Purpose and goals

Due to the issues introduced above, to guarantee the proper functioning of the hospitals in their entirety and, thus, improve patient outcomes, it is crucial to monitor and enhance ED performance continuously. To achieve this, this thesis aims to find a suitable way to evaluate the clinical impact of complex patient characteristics on ED logistics and to support hospital management in better understanding and intervening regarding the problems leading to excessive LOS within the ED. More specifically, the main goal consists in achieving such an aim by designing and implementing a simplified and empirical process model describing the ED system. In this perspective, patient flow modeling based on real-world data can help find which factors impact the system performance in given situations, support decisions concerning resource allocation and utilization, and help improve the process pathways and perform patient stratification. The primary benefit of achieving this thesis' goals is that the employed approach for evaluating the impact of clinical covariates on logistical outcomes could become the starting point for future operational research studies aiming to test LOS optimization procedures in the ED. Consequently, it could become easier to avoid the discussed negative consequences of longer LOS on patients, staff, and management.

2. Materials and methodology

Data sampling was performed by the hospital Akademiska sjukhuset, which provided two datasets regarding the patients who sought care from their ED in 2019. The datasets were then analyzed and bridged through Python programming, and relevant information was extracted. The ICD-10 terminology associated with the main diagnosis of each patient was simplified to the corresponding macro diagnostic area by being shortened to the first letter (variable "*simple_ diag*"). Other variables of interest were computed. The patients with no age information, those discharged in 2020, and those who visited the ED during June, July, and August, were censored from the dataset. The latter was reshaped to introduce a time coordinate and a state variable to interpret the process as a Continuous-Time Markov Chain (CTMC) for performing the parameter estimation from the data. After having grouped four possible modalities of discharge ("taken in charge by consultants", "death of the patient", "redirected", and "other, unspecified"), a numerical identifier was associated with each of the remaining states as in figure 1. Contextually, these were assigned to have the patients in one of the two initial states at time 0, in state 3 after one minute, and in one of the final states at time "*LOS + 1 minute*". Five **independent** sub-sets were extracted with different random seeds and a stratified sampling by proportionate allocation of the values taken by *simple_ diag*. Four samples contained 933 or 934 patients each and were used to extract suitable covariate sets; a sample of 5031 patients was used to assess the model's validity. For all such samples, the distribution of all the covariates resulted consistently among the sub-sets.

2.1. Modeling approach & technique

The employed modeling approach is "nonlinear mixed-effects modeling" (NLMEM). It includes fixed (F.E) and random effects (R.E), where F.E represent typical population values, and R.E represent inter-individual, intra-individual, and residual variability. It describes a response variable as a function of the predictor variables while recognizing correlations within sample subgroups, providing a good compromise between ignoring data groups entirely and fitting each group separately. A nonlinear mixed-effects model for M individuals and n_i observations on the i^{th} individual can be described as:

$$y_{ij} = f(\boldsymbol{\phi}_{ij}, \boldsymbol{v}_{ij}) + \epsilon_{ij},$$
$$i = 1, ..., M, \;\; j = 1, ..., n_i. \qquad (1)$$

ϵ_{ij} is a normally distributed within-individual *error term* and f a general, real-valued, differentiable function of an individual-specific parameter vector $\boldsymbol{\phi}_{ij}$ and a covariate vector $\boldsymbol{v}_{ij}$. When equation 1 describes a mixed-effects model that is nonlinear, f must be nonlinear for at least one component of $\boldsymbol{\phi}_{ij}$. The latter is modeled as:

$$\boldsymbol{\phi}_{ij} = \boldsymbol{A}_{ij}\boldsymbol{\beta} + \boldsymbol{B}_{ij}\boldsymbol{b}_i, \qquad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad (2)$$

where $\boldsymbol{\beta}$ is the F.E vector, $\boldsymbol{b}_i$ is the R.E vector for the i^{th} individual, and $\boldsymbol{\Psi}$ is the variance-covariance matrix. The matrices $\boldsymbol{A}_{ij}$ and $\boldsymbol{B}_{ij}$ are individual-dependent and possibly dependent some covariates at the j^{th} observation [3]. Given the choice of NLMEM and the intention to exploit its data longitudinalization, but also given the performed data analysis and comparison among modeling techniques, it was chosen to describe the process as a Markov Chain with "memory 1". In this framework, the observed data take values in a fixed finite set of categories, and the observations for any i^{th} individual are of a sequence of random variables. The dependence between observations from the same individual is defined so that, for each observation, to determine the distribution of y_{ij}, no older value than the one of the immediately preceding observation is needed. Being the observations reported at different times for each patient, a CTMC approach was selected. In a CTMC, the system stays "*in the current state for some random time before transitioning*" to a new one [4].

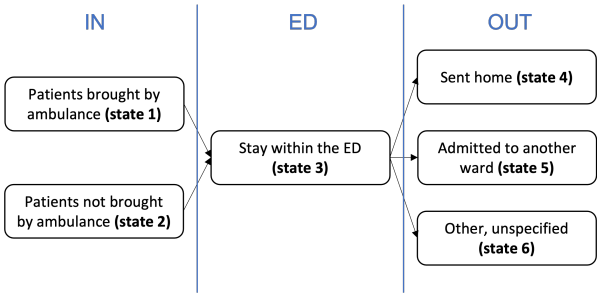

Figure 1: 6-states Markov Chain Model.

The base model in figure 1 was translated into a structural model in Monolix, the chosen software for implementing the NLMEM approach.

2.2. Parameter estimation and additional tasks

Once designed the structural model, it was chosen to estimate the parameters via "Maximum Likelihood Estimation" and to describe *q34* (transition from the ED to the home), *q35* (ED to a hospital ward), and *q36* (ED to another discharge type) with a log-normal probability distribution. The **first task** performed on the dataset estimated the population parameters via the "Stochastic Approximation Expectation-Maximization" (SAEM) algorithm. The **second task** estimated the conditional distribution for each individual and parameter and computed conditional means and standard deviations. The **third task** estimated the mode of the conditional distribution for each individual. The **fourth task** returned the correlation matrix and the uncertainty of the estimated parameters. Moreover, it computed the condition number (C.N) to help detect any model overparameterization. The **fifth task** estimated the log-likelihood and computed the corrected Bayesian Information Criterion (BICc).

2.3. Model and procedure design

An **experimental protocol** was defined to process the available data and select meaningful covariates to be included in the process model for describing complex patient characteristics in relation to the LOS within the ED. The protocol is shown in figure 2, and the extraction of the data sub-sets mentioned in such a scheme is discussed in chapter 2. Concerning the employment of the **automatic covariate model building** tool called "COSSAC", this algorithm exploits the information contained in the base model run to choose which covariate to try. It uses the correlation between individual parameters (I.P) or R.E and covariates as hints at possibly relevant parameter-covariate relationships. Such correlation values are calculated using samples from the a posteriori conditional distribution (see section 2.2). For evaluating the continuous covariates, Pearson's correlation test is used to derive a p-value, whereas ANOVA is employed for the categorical covariates. These p-values are then used to sort which random effect-covariate relationships to include in the model [4]. The COSSAC iteration yielding the best covariate set was mainly chosen based on the value of the BICc.

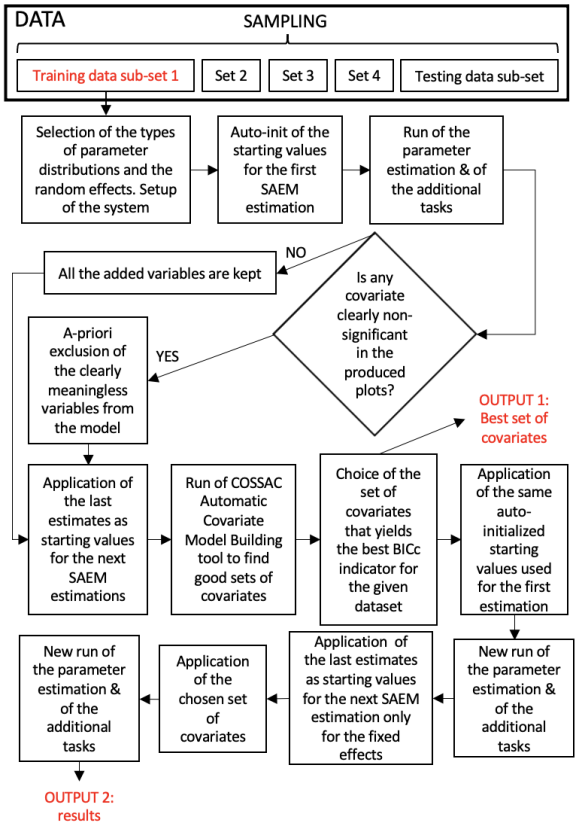

Figure 2: Experimental protocol.

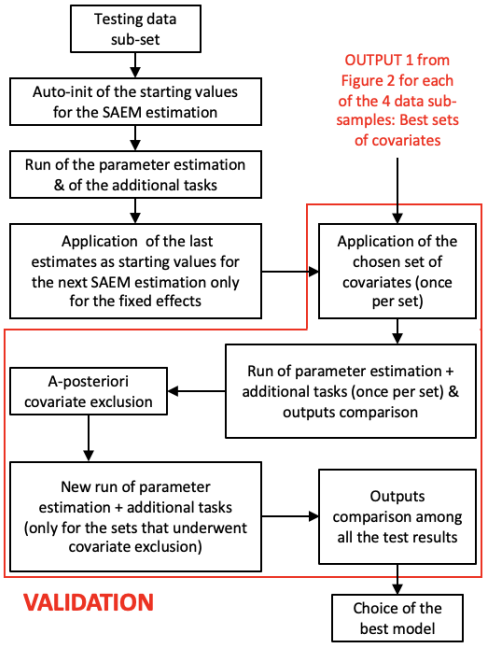

Figure 3: Validation protocol.

Concerning **validity assessment**, this is shown in figure 3. Starting from the base model, the associated initial estimates for the "testing" data sub-set, and the covariate sets selected with the COSSAC algorithm, each set was individually

applied to such a base model. For each set, a parameter estimation run was executed on the "testing" sub-set, the outputs were compared, and one covariate was a posteriori excluded. The covariate sets subjected to covariate exclusion were tested again on the "testing" data sub-set. All model outputs were compared, and the best-performing model was determined accordingly.

3. Results

After comparing the existing modeling techniques for process modeling in healthcare, results related to the steps of the experimental protocol (figure 2) were produced. In the first estimation run of the base model without covariates on the five data sub-sets, the relative standard error (R.S.E.) for all parameters resulted reasonably across the data sub-sets. The Shapiro-Wilk (S.W.) normality test on the R.E and the transformed I.P confirmed their normality. The T-test among R.E showed no correlation. The values of the C.N indicated good confidence in the model not being over-parameterized, and the normalized prediction distribution errors (NPDEs) resulted normally distributed. At this stage, two covariates on which the I.P clearly showed no dependence were excluded. The application of the COSSAC algorithm to the four "training" data sub-sets then produced the results shown in section 3.1.

3.1. CTMC Modeling

For random **seed n° 1** the selected covariates are: "T" for $q34$; "AmbulYN", "K", and "ScanYN" for $q35$; "MA_unit" and "Z" for $q36$. For **seed n° 2**: "R" for $q34$; "K" and "age" for $q35$; "Z" for $q36$. For **seed n° 3**: "M" and "age" for $q34$; "MA_unit" and "age" for $q35$; "A", "K", and "ScanYN" for $q36$. For **seed n° 4**: "E", "K", "M", "ScanYN", and "age" for $q34$; "A", "I", "K", and "Z" for $q35$; "Z" for $q36$. Once for each set of covariates, a set was applied to the base model, and a new estimation run was performed on the corresponding data sub-set from which the covariate set had been extracted. For all sets, the new estimation resulted in good R.S.E. on the population parameters, normally distributed NPDEs, and improved likelihood values. However, for seeds 1 and 3, it became impossible to compute the standard error and R.S.E. for a F.E of the covariate *MA_unit*. Moreover, the C.N. resulted in 65.19 for seed 2 but was not computable for seeds 1,3, and 4. Lastly, the correlation test between I.P and covariates suggested removing two possible classes of *MA_unit* for seed 1.

3.2. Validity analysis

The extracted covariate sets were applied to the base model and tested on the "testing" data sub-set. The S.W. normality tests confirmed the normality of all R.E and transformed I.P in the test with no covariates. The Kolmogorov Smirnov adequacy test confirmed that the I.P were samples from a mixture of transformed normal distributions. No correlation between R.E was found, and the NPDEs were normally distributed. The C.N values indicated good confidence in the model not being over-parameterized for the model with no covariates and the one with covariate set number 2 since such C.Ns were much smaller than 100. For the model with the fourth covariate set, the C.N was equal to 98.01. For the covariate sets 1 and 3, the C.N indicated a high risk of overfitting. Moreover, the distribution of $q35$ showed two peaks only for the model with covariate set number 4. The covariate sets 1 and 3 included dependences on the covariate MA_unit. For such two cases, rerunning the same test after excluding MA_unit led to a mild worsening of some likelihood indicators but a great improvement of the C.Ns.

4. Discussion

Concerning the base model's performance on the five sub-sets, no clinically meaningful parameter showed any large R.S.E., the transformed R.E were normal and uncorrelated, and the transformed I.P and NPDEs were normal. Additionally, the base model showed no signs of overfitting. By applying "COSSAC" to four data sub-sets to find covariates able to capture the variability affecting *q34*, *q35*, and *q36*, each covariate could be selected up to 12 times. The covariate sets producing the greatest improvement in likelihood were chosen. Among the 28 explored covariates, figure 4 shows those selected at least once but only those selected at least thrice are discussed here. ***K*** (digestive system diseases) seems to be the variable that best describes LOS variability in the model. This may be because more than 50% of the patients with "K" as their

simple_ diag had "abdominal pain" as their chief complaint, which is affected by great LOS variability due to how various the causes of such pain could be and the effort to evaluate them. ***Z*** represents conditions of no specific disorder but in which treatment was warranted (e.g., due to self-poisoning), and it was selected by three COSSAC runs for *q36*. Such a result implies *Z*'s strong ability to explain the LOS variability for patients who are not sent home or admitted to a ward. This is reasonable since around 50% of the patients with "Z" as their *simple_ diag* were discharged in a way included in the category "other". In most cases, the ED staff provided them with basic cures and then sent them to specialized clinics. ***age*** was chosen by two COSSAC runs for *q34* and *q35*. A reasonable explanation for this output is that several young patients visited the ED for intoxication or poisoning, thus requiring prompt care and soon becoming transferable to the hospital wards (short LOS within the ED) due to the ease in understanding the causes of their conditions, whereas several elderly patients visited the ED for simple needs (e.g., disorientation) but could not leave until a special mean of transportation would be available for taking them home (long LOS within the ED). The medical imaging covariate ***ScanYN*** was selected three times, which is reasonable since medical imaging can increase patient LOS regardless of how they are discharged. This result also proves the importance of such a variable in LOS variability, thus constituting a hint for where to focus future improvements.

Variable	q34	q35	q36	TOT	% over 12 possibilities
K	1	3	1	5	41,67
age	2	2	0	4	33,33
Z	0	1	3	4	33,33
ScanYN	1	1	1	3	25,00
M	2	0	0	2	16,67
MA_unit	0	1	1	2	16,67
A	0	1	1	2	16,67
T	1	0	0	1	8,33
R	1	0	0	1	8,33
E	1	0	0	1	8,33
AmbYN	0	1	0	1	8,33
I	0	1	0	1	8,33

Figure 4: Count of the selected covariates.

4.1. Final model assessment

After employing the COSSAC algorithm with all covariate sets, it was possible to assess their ability to help the model describe the clinical variability embedded into complex patient characteristics. On the "training" sets, the models failed in computing the effects of *MA_ unit* for seeds 1 and 3 and the C.N for seeds 1, 3, and 4. However, the latter was a sample size issue. It was then possible to test the covariate sets on the "testing" data sub-set. Once again, the achieved R.E and NPDEs resulted normal and uncorrelated. On such a larger data sample, the C.N could be computed in all cases, but its value indicated overfitting with sets 1 and 3. Grouping the information given by the C.Ns and the estimations on the "training" data subsets led to excluding *MA_ unit* from models 1 and 3. A new estimation run for these models, this time without *MA_ unit*, produced good C.Ns. At this stage, it was possible to evaluate the best-performing covariate set and propose a final model. The fourth covariate set yielded the best likelihood values but a relatively high C.N (98.01). Moreover, the plot of the I.P estimated with this model showed two peaks for *q35*, which had never happened in the other estimations. Due to the second peak, the higher C.N, and the highest number of included covariates, the model was excluded for its low generalization capability. Then the third model was excluded since it showed the worst likelihood values and the second-worst C.N. The two remaining models were compared regarding likelihood values, better in the second one, in terms of C.N, better in the first one, and concerning number and quality of the included covariates. The first model included five covariates (no *MA_ unit*), two of which were not selected in any other model. The second model included only four covariates, two of which had been selected in two other models and one in another model. Eventually, considering all the mentioned factors, the second model was defined as the best.

4.2. Value of the approach

NLMEM is not commonly employed to evaluate the impact of clinical covariates on logistical outcomes in process modeling. However, the data analysis revealed high complexity in patient characteristics and some data sparseness. Moreover, the assessment of the conventional approaches showed several relevant limitations that these face when trying to describe the variability embedded into health logistics data. More specifically, these tend not to allow for

much differentiation among the covariates that are significant for each modeled state transition nor for the consideration of R.E. Therefore, it was reasoned that a NLMEM approach would have helped design an effective process model by allowing to differentiate among the significant covariates for each modeled state transition. Furthermore, it was reasoned that such an approach would have made the designed model less affected by data sparseness than with more conventional approaches. Lastly, NLMEM incorporates both F.E and R.E and allows extracting insight from the data using a population approach, thus fitting a model to data coming from all the subjects without losing the notion of individuals and allowing for discrimination between inter-individual and intra-individual variability. The ability to estimate variability and covariate effects is very relevant for this application area.

4.3. Limitations

The model does not represent the system in its full complexity, and the boundaries are limited to the ED and the employment of the imaging department for ED patients. The process model does not automatically update the parameters over time if the system's conditions change. Explicit outlier removal was not performed on LOS values, but precautions were taken. The model was informed by real-world data, which may include wrong information. The high computational demand of the approach led to the need to sub-sample the dataset and carefully design the number of transitions allowed by the model.

4.4. Future work

A **supplementary validation** could be performed. The employed approach could be used to **inform operational research**. A **computational parallelization feature** could be designed. The results achieved with the COSSAC algorithm could be compared to other covariates selection techniques. **Time-varying covariates** could be introduced to account for the yearly variability within the system, or several model scenarios could be analyzed separately.

5. Conclusions

This thesis applied mixed-effects modeling to hospital medical records. Within the chosen approach, a Markov Chains model of patient flow that could capture and describe the impact of patient complex characteristics on the logistics of the ED was designed, tested, and validated. This was done to bridge logistical systems and the clinical insights of the hospital, which is particularly challenging due to the difficulty in dealing with high patient volumes and clinical variability embedded into real clinical data. Accordingly, this work aimed at improving the understanding of how such data could be better exploited for healthcare modeling to potentially achieve a better organization of the hospitals in the future, and it managed to develop an approach for estimating covariate effects on parameters linked to the process description in the ED. Furthermore, due to how much the performance of the ED affects the functioning of the other hospital wards and, indirectly, healthcare systems and communities, the technique applied in this thesis, as well as the deriving model, were designed so that they could become the starting point for future operational research studies to test LOS optimization procedures on the ED.

6. Acknowledgments

Thanks to my supervisors and all the others who supported me in this challenging academic path.

References

[1] H. R. Rasouli, A. A. Esfahani, M. Nobakht, M. Eskandari, H. Goodarzi, and M. A. Farajzadeh, "Outcomes of Crowding in Emergency Departments; a Systematic Review," p. 10, 2019.

[2] S. Paling, J. Lambert, J. Clouting, J. González-Esquerré, and T. Auterson, "Waiting times in emergency departments: exploring the factors associated with longer patient waits for emergency care in England using routinely collected daily data," *Emergency Medicine Journal*, Sep. 2020. doi: 10.1136/emermed-2019-208849

[3] C. J. Pinheiro and M. B. Bates, *Mixed-Effects Models in S and S-PLUS*, ser. Statistics and Computing. New York: Springer-Verlag, 2000. ISBN 978-0-387-98957-0

[4] Lixoft, "Monolix documentation - LIXOFT." [Online]. Available: https://monolix.lixoft.com/