POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell’Informazione

Corso di Laurea Specialistica in Ingegneria Gestionale

Uncovering Physician Information Needs from Outside Healthcare Facilities via Association Rule Mining

Relatore: Prof. Emanuele Lettieri

Correlatore: Prof. Jose L. Zayas-Castro

Tesi di Laurea di:

Elia Mora Matr. 804541

Martino Gemmani Matr. 795666

Anno accademico 2013 – 2014

TABLE OF CONTENTS

TABLES AND FIGURES

ABSTRACT

Introduction: In the United States, the health care system is fragmented among several different providers. This fact hampers the effective exchange of information about the patients among different health care facilities significantly. This thesis work tries to provide further insights on this topic, by shedding light on what type of patients' clinical information from outside health care facilities is requested by clinicians working in a hospital. Moreover, the study tries to make the process of information retrieval more efficient, by creating a priority scheme of access to this information according to specific types of diagnosis.

Literature Review: Health information exchange (HIE) is today a relevant way of electronically sharing information between providers. HIE could be the possibility of reducing time and costs in patient care, when well implemented. There are many factors influencing the adoption and the usage of HIE system in a hospital, mostly depending on the social and economic situation of the American health care system background. Also, the actual state of healthcare IT presents several problems, such as data quality, timing to access, design and physicians' behavior and information needs issues. The aim of this research is to identify the data most commonly requested by hospital physicians to treat the most frequent diagnosis.

Methods: for the purpose of the research, we selected an affinity analysis, widely used for market basket analysis, which is a tool to discover co-occurrence relationships between events performed by specific agents. More specifically, the Apriori algorithm has been selected. This algorithm is a simple adoption algorithm for identifying frequent item-sets and mining association rule for future cases.

Results: By linking the physician need of information and the most useful data exchanged between the different health structures we obtained some relation rules. Most of them are relations between a diagnosis and an outside information document. Starting from the confidence and support levels, it was possible to classify them in critical, useless, bad and top rules.

Discussion: The best 20 rules generated by the algorithm are the main research contribution and starting from these it is possible to formulate smart guidelines necessary to design and improve a new IT healthcare systems, which is more effective to provide integration among different health information about the same patient generated in diverse health care settings. With respect to the specific methodology, the Apriori algorithm generated several association rules, but the most of them was not interesting for our scope because of their composition, so we made another selection phase in order to obtain only useful rules for our goal.

Conclusion: Our outcome shows clearly how it’s possible discovering, through the use of data mining methods, new useful association rules that could be considered smart guidelines for the HIE system. In particular these relations will be implemented in the new healthcare IT systems, creating a network efficient and which helps physicians in them work.

SOMMARIO

Negli Stati Uniti il sistema sanitario risulta da sempre essere un tema critico, per complessità e rilevanza. La sanità americana va infatti a porsi perfettamente in linea con gli altri sistemi sanitari occidentali, contraddistinti per la maggior parte da una forte frammentazione che, nel contesto statunitense, risulta ulteriormente inasprita da una forte concorrenza dei diversi sistemi privatistici e assicurativi. Tale fenomeno ha conseguenze notevoli sulla totalità del sistema sanitario statunitense in quanto i cittadini risultano fortemente vincolati alla loro realtà sanitaria e nel momento in cui si devono rivolgere a enti esterni questi non posseggono dati a riguardo dei nuovi pazienti. Risulta intuitivo come paziente si venga a trovare in una situazione più rischiosa e meno efficiente: senza i dati necessari disponibili, gli ospedali si vedono ogni volta obbligati a ripetere esami già sostenuti dal paziente, perdendo tempo e sostenendo costi evitabili, ma soprattutto il paziente risulta essere l'unico a conoscenza della propria storia clinica completa, esponendosi cosi volontariamente o no, a possibili pericoli derivanti per esempio dall'assunzione di medicinali incompatibili. Preso atto di questi problemi, dal 2009 il governo americano ha cominciato a mettere in pratica una forte politica di incentivi economici, investendo centinaia di milioni di dollari all'anno, in particolare cercando di creare un sistema più integrato ed efficiente, basato sulla condivisione dei dati e delle informazioni tra i vari enti sanitari. Tale politica a livello di ricerca si è evoluta nell'ambito che oggi si chiama Health Information Exchange (HIE) e il presente lavoro di tesi rientra quindi in questo campo così ampio e rilevante. Consapevoli di affrontare tutte le tematiche relative all'HIE, si è deciso di orientare la ricerca tentando di affrontare le seguenti tematiche:

- **Qualità dei dati:** la mancata integrazione dei sistemi informativi porta questa come una della principali conseguenze. I dati rispetto a pazienti provenienti da fonti rispetto all'ospedale spesse volte risultano incompleti, sbagliati o comunque inconsistenti;

- **Facilità e velocità di accesso ai dati:** per sopperire alla mancata condivisione di informazioni, gli ospedali tendono a creare ingenti depositi di dati che risultano si completi ma troppo strutturati per un facile e veloce accesso da parte dell'utente;
- **Comprensione dei bisogni dei medici**: spesse volte i medici accusano i sistemi informativi ospedalieri di essere inutili o inefficienti in quanto forniscono informazioni inutili mentre invece per trovare quelle a loro necessarie occorre troppo tempo e ricerca;
- **Mancanza di integrazione e mancata condivisione di dati:** risultando concorrenti e non parte invece di uno stesso sistema, gli ospedali non condividono di proposito dati e informazioni relative ai pazienti a meno che non risulti strettamente necessario, ma anche in quel caso le informazioni condivise sono le strette indispensabili.

In particolare questa ricerca è mirata a fornire come risultato finale una serie di linee guida volte alla ristrutturazione e alla ridefinizione degli attuali sistemi informativi ospedalieri. Il metodo utilizzato per ottenere questo risultato è basato su approccio di tipo quantitativo, consistito nell'implementazione di uno strumenti di data mining chiamato algoritmo Apriori su una base di dati fornita dal principale ospedale dell'area di Tampa, in Florida. Questa scelta è stata effettuata considerando diversi fattori. In primo luogo l''innovatività del metodo di ricerca rispetto al tema trattato: esistono pochi precedenti di tecniche di data mining applicate all'ambito HIE, si è quindi optato per questa scelta credendo nell'affidabilità e nella robustezza dei risultati forniti da questa metodologia che normalmente vengono applicati in ambiti diversi, come il marketing. Secondo fattore rilevante è stata disponibilità di dati accessibili, ottenuta grazie alla collaborazione tra Politecnico di Milano e University of South Florida, partner di ricerca del suddetto ospedale. Gli obiettivi concreti preposti prima dell'implementazione dell'algoritmo Apriori erano 2:

1. Scoprire quale fossero le 20 relazioni più frequenti tra sintomi del paziente ricoverato e informazioni richieste ad enti esterni all'ospedale.

2. Fornire una classificazione chiara e intuitiva che aiutasse nell'interpretazione di queste regole associative.

Al fine di conseguire questi risultati, è stato eseguito un accurato processo di selezione, trasformazione e infine implementazione dei dati prima tramite Microsoft Excel e poi tramite R. L'esito del lavoro è risultato in linea con le aspettative, nonostante qualche differenza. La distribuzione particolare dei dati, che ha visto una forte presenza del informazione esterna chiamata "Outside Medical Record" (OMR), ha reso necessario rivedere la struttura delle 20 regole ricercate che ha questo punto sono risultate tutte in funzione di questa tipologia di dato. Ottenuta questa serie di relazioni si è poi eseguito una trasformazione logaritmica dei valori relativi a queste regole in modo da ottenere una struttura di dati facilmente classificabile in quella che poi è stata chiamata HIE Apriori Matrix. Il valore di questa matrice è la capacità di esprimere visivamente la bontà delle regole trovate mostrandone le principali caratteristiche, ovvero la probabilità di occorrenza, che si può vedere tramite il supporto, e la capacità predittiva, attraverso la confidenza, e classificandole poi in 4 categorie: Top, Useless, Critical e Bad Rules. In conclusione si può affermare di essere riusciti a raggiungere gli scopi preposti. Si è infatti mostrato come, tramite l'utilizzo di una tecnica innovativa rispetto al contesto e partendo da alcuni dei temi più rilevanti legati al mondo HIE, sia possibile in primo luogo acquisire una conoscenza maggiore legata all'ambito sanitario, in particolare nel nostro caso rispetto allo scambio di informazioni tra enti sanitari, per poi in seguito valutare e implementare quei cambiamenti necessari al sistema per aumentarne il livello di servizio per i pazienti e l'efficienza totale. Nel caso considerato questo secondo passaggio decisivo è consistito nel fornire alcune linee guida, come i vari livelli di priorità relativi alle diverse categorie di regole associative o l'importanza assoluta dell'accesso alla voce OMR, utili alla ridefinizione dei sistemi informativi ospedalieri. Grazie a questi risultati si può quindi affermare di aver contribuito all'apertura di una nuova via all'utilizzo di tecniche di data mining applicate alla ricerca sanitaria.

1. INTRODUCTION

1.1. Significance

In the United States, the health care service is fragmented. Hospitals and primary care services are often organized in small group of providers with autonomous functions, and very focused on specific areas or treatments. The complex needs of patients would require them to visit several and different health care providers. With a similar system structure, the exchange of the patients' medical information between clinical actors working in different facilities appears clearly as a critical factor to consider and to take care for effective delivery of care [Nicholson, 2003].

In 2010 The National Priorities Partnership (NPP), a partnership of the first 52 major American Organizations with the objectives to achieve better health service and a safety value-driven healthcare system, identified care coordination as one of six national priorities for health care. Indeed, , a lack of care coordination could lead to serious problems and complications, including medication mistakes, avoidable hospital readmissions and potential lost in terms of time, cost and more importantly, human lives [Ross, 2010] .

The need for efficient care coordination clearly emerges if we consider the cycle of care of a patient. Patients usually seek care from many diverse providers, and these providers may prescribe different medications for the same patient. In this way, patients are responsible of keeping track, by themselves, of all their medications. The direct consequence of this is that information about medication may be confusing, especially for those patients with more than one medication. It clearly appears that when care is not well integrated and each provider is not conscious of the entire medications taken by a patient, this is in a high risk situation because the patient could be affected by adverse drug interactions and possible adverse events related with this fact, such as overdosing or underdosing. Further, physicians need to review periodically a patient's therapies to ensure that they are following their treatment plans by doing what is needed at the appropriate time. Another relevant problem is related with the need and usage of laboratory information. Several medications require these kinds of information to monitor patient health but many studies report that often these labs are incomplete, not available or not updated, exposing patient health to a relevant risk [Kaelber, 2007]. With the introduction of a smart designed and implemented unique medical information system this entire problem could be avoided and solved [Ross, 2010].

Another relevant fact to consider is the economic waste behind an inefficient coordination system. The Institute of Medicine estimates that improvements on health care coordination are expected to generate savings for $240 billion per year, as reported on the official website of the National Quality Forum. The National Quality Strategy seeks to improve care coordination through the following objectives:

- Improve communication during transitions of care among patients with chronic conditions or disabilities;
- Improve preventive care services for patients with chronic conditions and disabilities; and
- Reduce healthcare disparities and improve quality of care by integrating communities and healthcare systems.

One of the initiatives to achieve these objectives is the connection of all the patients' electronic records of the stakeholders in the healthcare system. This effort is called health information exchange (HIE), and it implies that systems are electronically connected to share patient-level information. In an effort to inform the design and implementation of HIE networks, we aim to answer the fundamental questions: ***what type of patient-level clinical information is requested by clinicians working in a hospital from outside health care facilities? Is it possible create a priority scheme of access to this information according to specific types of diagnosis?*** The health IT research community believes that a smart HIE system has to provide the right data on time and with the highest accuracy level possible [Kaelber, 2007]. Provided the relevance of the health information exchange, in the next chapter we will examine the state of art of this topic.

2. LITERATURE REVIEW

2.1. Health Information Exchange

HIE represent an effort to facilitate an effective and efficient electronic sharing of health information among unaffiliated stakeholders. These include hospitals, ambulatory practices, skilled nursing facilities laboratories, patients, and others.

Various efforts are underway to lead RHIOs (Regional Health Information Organization) and other data sources into a cohesive national network of healthcare stakeholders, sometimes referred to as the Nationwide Health Information Network (NwHIN, defined at www.healthit.gov/policy-researchers-implementers/nationwide-health-information-network-nwhin) [Shapiro, 2007]. It is expected that HIE will cut down unnecessary duplicate diagnostic tests, avoidable re-admissions, enhance the use of large data sets for public surveillance and research, and, consequently, improving efficiency and lowering healthcare costs. Much evaluation is needed to demonstrate that these promises will come to fruition once the investment on these networks is made [Kaelber, 2007].

In essence HIE should facilitate the movement or migration of patients among/across providers, payers and other stakeholders. Consequently the effective implementation of HIE should reduce the existing fragmentation in the healthcare system [Adler-Milstein, 2014] and allow clinicians to have timely access to the information that they need in delivering care to the patients [Hincapie, 2011].

Although it is believed to be essential for better care coordination [Zwaanswijk, 2011], its adoption in the USA is still low. A barrier in increasing adoption is lack of workflow integration. This adoption has typically been performed through a “portal”, a stand-alone results review application. However, these portals may impose some significant workflow integration issues because they require the clinician to perform a separate task that is not part of his/her normal routine. In addition, trials are being made to import the HIE data directly from various electronic medical records (EMRs), allowing clinicians to view the data in their own native application in a more seamless way.

2.2. Types of HIE Structure

Health information exchange may be enabled by a number of different technical models, including centralized, hybrid peer-to-peer and standards-based document sharing.

In smaller markets with a main provider organization (i.e., one large academic medical center) and/or a dominant commercial payer, a centralized data model may predominate. In this case, the dominant stakeholder originates a centralized server that hosts own data and a master patient index allowing patients to be identified across organizations. In a centralized model, the associated stakeholders (e.g., ambulatory practices, smaller community, hospitals, and commercial labs) allow their data to be stored on the centralized server. This architecture may decrease overall installation and operating costs.

In larger markets, for which there is no preponderant provider or payer, HIE employs a federated peer-to-peer model using “edge” server that are either real or virtualized. These servers lie behind the firewall at the “edge” of each healthcare organization’s IT infrastructure. All the servers in a given exchange use a common data model and are securely connected using encryption or encrypted virtual private networking (VPN) tunnels, allowing them to function collectively as a distributed data repository. All clinical data are stored on the edge servers, and each healthcare organization keeps stewardship of their data.

There are many challenges that remain before a nationwide health information network becomes fully operational, with interoperable electronic health records pushed out to every stakeholder of the healthcare system. Consistent privacy and consent policies need to be developed allowing RHIOs to interoperate among themselves and across state lines. Sustainability and funding issues must be addressed since the financial benefits have yet to be proven in a meaningful way. The recent allocation of federal resources for health IT and HIE development is still only a fraction of the estimated costs.

2.3. How HIE is helping patient care

Between the many potential advantages of HIE, patient safety stands out as one of the most promising. Patient safety can be corroded by both types of errors, of a commission and of omission if the right information is not timely available to the right person. Better patient safety through enriched technology-enabled HIE will directly enhance patient safety because it will provide a complete clinical description of a patient. The more people and more information are related in the HIE, the more relevant the exchange will be for patient care. Up to 18% of patient safety mistakes have been estimated to have occurred because the appropriate information was not available at the time the medical decision was made [Leape, 1995]. Although HIE is implemented in a small number of institutions, initial evidence is showing that HIE can improve healthcare delivery in a number of areas:

1. **Improved medication information processing**. Medication information processing probably represents the most studied area of HIE today. One study has evaluated that over 100,000 deceases occur every year in the United States because of adverse drug events (ADEs), including both non-preventable and preventable ADEs [Lazarou, 1998]. Although most of these ADEs will not be amenable to elimination with improved HIE, a relevant proportion will, and many opportunities and strategies that enhance patient safety through HIE have been studied. Drug–dose information processing has been shown to improve patient safety in a number of

settings. For instance, renal dosing of medications is perhaps the single most important part across the decision support, and when dosing is based on the patient's renal function it is more often appropriate.

2. **Improved laboratory information processing**. Patient safety can also be improved by enhanced laboratory information processing enabled by HIE. Two primary areas are: (1) insuring that the indicated laboratory test is ordered and (2) guaranteeing that laboratory results (especially abnormal results) are followed up.

   A prime instance of the interplay between laboratory information processing and patient safety is medications. Dozens of commonly used medications require labs tests prior to initiation and/or after initial administration to monitor for patient safety [Chen, 2005; Committee, 2005]. Unfortunately, many studies document inappropriate laboratory information processing [Mann, 2006]. Other examples in this area include appropriate ordering and follow-up of Pap smears, prostate-specific antigen (PSA) levels, cholesterol levels, and stool guaiac testing, to name a few.

3. **Improved radiology information processing**. Typically the provider ordering an imaging study differs from the provider interpreting the study. Therefore, health information has to be exchanged between these two professionals for the study to be effectively ordered and interpreted. For example, improved HIE could decrease adverse intravenous contrast reactions and decrease exposure to inappropriate radiology testing [Levy, 2006], as well as radiation exposure [Hadley, 2006]. Probably more important for patient safety is enhancement in HIE to ensure appropriate follow-up of abnormal radiology findings. For example, up to 2% of abnormal mammograms were found to be lost to follow-up without enhanced information exchange [Choksi, 2006].

4. **Improved communication among providers**. In the USA on average a patient has four outpatient appointments per year, with just over half of these visits being to primary care providers, approximately 40% to specialists and 10% to emergency departments [Cullen, 2005]. In addition, there are 114 hospital discharges per 1000 people per year [Bernstein, 2004]. The result of these interactions is that many providers yearly get in contact with each patient. Although many of these interactions are for acute or sub-acute problems, each of these encounters can provide valuable information for the patient's future care. Providers who are not familiar with the patient, either in an inpatient or outpatient setting, make safer decisions with improved HIE [Sutcliffe, 2004].

5. **Improved communication between patients and providers**. Patients and healthcare organizations typically do not sufficiently recognize the key role that patients can play in assuring their own healthcare safety. With the significant concern and impending growth of improved HIE through personal health records (PHRs), many hope this paradigm will change. Over 60% of the people feel that PHRs will help prevent medical mistakes [Foundation, 2003]. Examples include patients checking for errors in their medical past times, adding new valuable information into their medical records, following up on their test outcomes, examining medications and other healthcare instructions, and being able to convey more quickly with healthcare providers when they think their safety may be at risk.

6. **Improved public health information processing**. A rapidly growing area of HIE is public health informatics. Patient safety could be greatly improved through this growth. For instance, in 2006 the Centers for Disease Control recommended expanding the age for childhood influenza vaccination based on enhanced public health informatics HIE [Bourgeois, 2006].

As more and more healthcare information evolves from paper to digital format, the potential for HIE to improve patient care will grow. One challenge will be developing healthcare systems capable of processing and using the dramatic increase in data. If we reduce the time needed to get the right information to the right person, then better healthcare decisions will be made.

2.4. Adoption and usage of HIE

Factors influencing or obstructing the use of HIE help to guide the design and implementation of future HIE. Below we present the state of the art in terms of factors influencing adoption and usage of HIE networks, as well as how these factors may influence HIE design. We divide our findings between medical practice in rural and metropolitan areas.

Medical group practices in rural areas.

Kralewski and colleagues [Kralewski, 2012] studied small medical practices in rural areas from three different HIE networks. First, they found that the adoption and usage of HIE were not associated with EMR's brand or the time since EMR adoption. Second, practices interested in improving administrative work (e.g., a billing process) rather than clinical activities were more likely to adopt and use HIE. Third, practices having a clear mission and vision about their role in the community where they were inserted were more successful in implementing HIE. Practices without a clear mission and vision seemed to deal with these aspects on a more ad-hoc basis and, consequently, decisions about the health IT to adopt often focused on a narrow set of capabilities. Fifth, the provision of technical support and accommodating workload changes were noted as key factors for HIE adoption and use. Sixth, clinician motivations and values greatly impact on the use of EMR functions and support HIE. Seventh, practices described as collegial with a team approach to patient care seem to have more EMR functions in place, but no information exchange with other providers. Eight, organizational leadership is key to successful implementation of HIE. Leadership in this circumstance includes creating sufficient support among stakeholders to get electronic clinical information issues on the agenda, acquiring the needed resources, maintaining support as the details unfold, and problem solving during implementation. Kralewski found that in successful practices this leadership often came from the medical staff. Normally, these leaders were also exceptionally computer literate and provided a great deal of technical help for the other clinicians during implementation. Ninth, costs were noted by all of the practices as a major problem, but none were able to provide estimates of how much it would cost to develop clinical information exchange with specialists in non-owned medical group practices or hospitals. Finally, the will of interconnecting EMRs is apparently trumped by patient preoccupation about information protection, physician concern about releasing information to competitors, and administrators who are very busy trying to keep the practices viable [Kouroubali, 1998].

This study has several discoveries relevant to public policies on HIE:

- RHIOs need to raise their effort to demonstrate the potential quality and cost gains from information exchange.
- RHIOs need to work with the providers to improve strategic planning at the medical group practice level. Strategic planning at the provider level creates a community wide perspective concerning provider relationships and can facilitate HIE adoption.
- RHIOs can increase provider confidence and make an important contribution to health care improvement by providing more electronic information technical help to small rural practices.

Medical group practices in a metropolitan area.

Rudin and colleagues [Rudin, 2011] found a wide range of usage intensity. First, most active users believed accessing the HIE helped them deliver better quality care by supplying them with relevant clinical data in a timely mode. They believed HIE saved time, in part, through prevented phones calls to request clinical data from other physician offices, pharmacies, hospitals and patients' relatives. For instance, a hospitalist believed it avoided more than 75% of such phone calls, saving him significant time. For office-based EMRs, which allowed immediate importing of data from the HIE, the HIE expedited documentation, mainly for inmates new to the practice. Many different clinicians believed that information gathered through the HIE facilitated interviews with patients and reduced the need to ask them as many questions. Second, none of the clinicians adduce cost as a motivating factor for accessing the HIE. Third, HIE was found to be more useful when serving patients who had trouble communicating, who lacked relatives to assist them, and who suffered from various or complex medical conditions. Emergency clinicians thought the HIE held considerable potential value to improve the efficiency by which patient information relevant to an emergency department visit could be found. Fourth, patients who only visited one practice for all their care, or who went outward of the community for care and therefore associated data would not be available in the community's self-contained HIE, clinicians had little motivation to access the HIE. On the other hand, for new patients with data in the system, clinicians discovered the HIE very valuable by saving time in gathering clinical information. Fifth, many clinicians perceived that their particular medical specialty determined how valuable HIE can be. A *pediatrician* who used the HIE seldom did not believe many pediatric visits had problems with missing clinical information because consulting physicians usually forwarded their medical notes back to this clinician via fax. A *psychiatrist* who likewise accessed the HIE infrequently believed the HIE would not be valuable for his specialty because psychiatric issues do not change often and are isolated from other medical conditions [Bailey, 2012].

The interviewed clinicians varied in how effectively they integrated HIE into their complex workflows. Several physicians were unaware of how to access the HIE directly from their EMR, did not know about the capability to import data from the HIE, or clearly did not think to check it to find missing patient data. Many physician noted that information sources they were accustomed to using "competed" to the HIE, such as a hospital portal which contained relatively complete patient data but for hospital visits only.

Existing information exchange processes using paper and fax may also have reduced the frequency with which physicians accessed the HIE. Many offices regularly faxed clinical notes to other providers in the community for referrals or in response to chart requests, reducing the need for the HIE. Physicians believed that specialists outside of the community were far less trustworthy in sending their notes but, since they were not part of the HIE, the HIE could not be used to pick up clinical information from their practices. Asking clinical notes via fax, although more time consuming than using HIE, had the benefit of containing textual notes, which were rejected from this HIE [Callen, 2013].

How clinicians coordinated with each other within their practices also affected HIE accesses. One practice adjusted their workflow so that either the physician or a nurse would routinely check the HIE for all new patients. Another physician, by contrast, thought that it was faster to simply tell his assistant to call another office than for him to check the HIE and had not thought to ask his assistant to check the HIE instead.

Some clinicians confessed that they had a general aversion to changing their practice workflow, mostly after a stressful process of installing an EMR. Time constraints, mainly in primary and emergency care, also tended to decrease motivations for accessing the HIE. Conversely, clinicians working during non-business hours found the HIE particularly valuable because other means of obtaining clinical information were unavailable.

Factors affecting HIE sustainability.

The most recent nationwide survey, using data from 2012, found that the number of operational HIE organizations is growing, and approximately one-quarter of them claim to have a supportable business models [Adler-Milstein, 2010]. However, many HIE organizations still struggle to find a value proposition. It is not clear whether the HIE organizations that are sustainable are centered on the forms of HIE included in this review or on a more basic forms of data exchange, such as the automated delivery of laboratory outcomes. One nationwide survey discovered 3 characteristics to be independent predictors of greater financial viability: having ambulatory physicians as receivers of data, having hospitals as a data receiver, and receiving a 1-time or recurring payment from participants while planning. Some HIE organizations have achieved sustainability, but many are still struggling to make a business case. Moreover the factors for achieving sustainability will likely change over time. Emerging payment models, such as accountable care organizations, grouped payments, and other risk-sharing payment arrangements, may be useful to create a greater value proposition.

Factors affecting HIE data contribution.

Data from each visit would be automatically contributed to the HIE immediately after a clinician “locked” his or her notes, which was achieved when the clinician performed a software action that indicated the documentation for the visit was complete. Note-locking was the only way for a clinician to contribute data to the HIE. It is stated that note-locking was influenced by the following factors: billing concerns, time constraints, and an aversion or lack of awareness of the ability to add addenda to notes. Physician’s note-locking habits varied considerably. One physician compulsively locked her notes within a few hours of the patient visit. One practice adopted the strategy of locking notes exactly one week after the visit to allow time for their billing department to check for mistakes. One primary care physician locked notes on an ad-hoc basis “whenever it pops into my head.” One specialist was nearly 3 months late in his notes. Another clinician, after a billing error resulted in lost income, stopped locking notes completely.

Lastly, however, they have also discovered a long list of potential moderators of these benefits which, if not addressed, may result in physician using this form of HIE minimally or not at all. This underuse could decrease much of the potential value of a HIE. Some types of clinicians accessed the HIE much more than others and had upright reasons for doing so, indicating that incentives targeted at providers may need to be considered. The main two HIE usage moderators are the extent of physician participation and existing electronic and paper processes [Miller, 2007].

Physicians did not have incentives to lock their notes in a timely manner because they were not the ones who directly benefited from having the data available. Implementation of HIE by an organization does not guarantee use by individuals within the organization [Vest, 2011]. Research has repeatedly demonstrated the organizational decision to adopt an innovation is frequently independent of individuals' adoption decisions [Jasperson, 2005]. Their exploratory study supports the objects in the context of voluntary use HIE, but also suggests when and why these systems are actually used and how to improve implementation.

2.5. Issues with HIE design that inhibit use

Drivers and barriers for HIE usage.

Initial studies suggest that time constraint for health-care delivery is a barrier to HIE usage [Vest, 2011]. Physicians press by time are less likely to consult additional sources of information through HIE networks. This simple result suggests that improving the design and functionality of HIE networks might impact utilization of HIE. Healthcare is busy and fast-paced, and some physicians already think HIE may not save time [Sicotte, 2010]. Given that the voluntary usage of additional information sources can be discouraged by time constraints, those desires to implement HIE have two options. One option is to improve the usefulness of the information and the system; in fact, change the equation so the potentially available information is more valuable than the opportunity costs. For example, screen redesign, single sign-on, eliciting user needs, or enhanced record searching could all be means to that aim. The second option is for organizations to dramatically increase the level of functional integration among exchange partners' EMRs and their own. Functional integration can be improved, for example, by directly placing the information made available by HIE into the organization’s EMR. Although this action might be legally, organizationally and even technically difficult, it effectively removes from the user the decision to seek or not supplementary information in the HIE system. Case reports suggest tighter functional integration is associated with higher proportions of usage [Wilcox, 2006], and the aspect of constrained time illustrates why.

On the other hand, previous results provide an argument for addressing low HIE usage by simply mandating user utilization. The decreased odds of utilization for some encounters but increased for others, suggests that users have determined HIE is useful in some, but not all cases. For current encounters that have little to do with previous exercise or information stored in other organizations, HIE may have less immediate value [Wilcox, 2006]. Furthermore, the effect of higher number of comorbidities on novel usage (that is a pattern classification for more complex patients), but not on basic usage, indicates that users employ the system to match their immediate needs. For involved patients, the minimum information provided by the HIE system was probably not sufficient; those encounters required more detailed investigation. Mandating utilization of an alternative information source that changes workflows needlessly, or when little potential value for the problem at hand exists, it is the perfect scenario to incite resistance [Lapointe, 2006]. Therefore, blanket mandatory usage of HIE networks might not solve perceived issues of underutilization.

In addition, HIE usage is less likely for unfamiliar patients contradicting expectations based on a theory and conventional wisdom. An unknown patient is essentially the poster child for justifying HIEs in the Emergency Department (ED) setting [Shapiro, 2006]. Such a view is understandable, as repeated contact should raise provider knowledge around the patient’s history and idiosyncrasies, thereby decreasing the need to seek additional information. However, this unexpected association suggests one very practical explanation why HIE, at least in the emergency setting, is used. In the ED, patient familiarity is not desired because it is indicative of patients with inappropriate origins of care. For the familiar patient, HIE might provide clinicians and organizations the necessary information to get and keep these patients out of the ED. The lack of an association between ED utilization at other locations and HIE utilization reinforces this interpretation. Patient knowledge of a facility is more significant than the frequency of a patient's visits to any ED.

Lastly, HIE usage seems to have non-clinical reasons as well. Evidence from previous studies suggests that HIE is more likely to be used in response to a facility's repeat patient visits [Kleinke, 2005].There are also associations found between payer kind and HIE usage. While Medicaid does not boast the most abundant restitution rates, some payment is better than no payment. Between patients for whom no payment could be expected, HIE system usage was higher, suggesting the possibility of either using HIE to locate past payers or to again help find patients more appropriate sources of care [Kleinke, 2005].

HIE usage variability.

This aspect seemed to us particularly essential, so we followed up this topic trough a study performed in New York [Rudin, 2014].

Reports of HIE usage showed wide variation in terms of rates of access, patterns of usage and types of users. The same study suggested that streamlined consent procedures played an important role in a HIE's high access rate compared with other instances.

The high degree of variability in use within different institutions and providers is consistent with patterns found in other health IT applications such as electronic prescribing and clinical decision support when they were first introduced. It confirms the statement that local context and implementation factors are probably relevant factors (along with software functionality and usability) in promoting high (or low) use of HIE.

The eHealth Initiative study discovered that many HIEs are not sharing data with competing organizations and that interoperability is a challenge to implement and finance [Kern, 2009].

Four studies based on the American Hospital Association survey supplement related to IT and focused only on hospitals [Miller, 2014; Furukawa, 2013; Adler-Milstein, 2011; Vest, 2010]. These studies showed that use of HIE increased over time but that larger hospitals were less likely to have implemented a HIE or exchange data for competitive causes. Many sustainability challenges were identified. Some of them are including competition between regional providers, costs and technical complexity of integrating with a HIE, uncertainty about who benefits, and potential reduction in the need for revenue-generating services [Miller, 2007]. Moreover, these studies suggest that attracting participants to HIE and achieving sustainability is complex and may vary widely across the country. Although the number of HIEs that have achieved sustainability grows, many are still at risk of being unsuccessful.

Although most stakeholders think that HIE will be valuable to health care, particularly in terms of quality and efficiency, there are many obstacles to adoption and utilization, and these vary somewhat by stakeholder. Physicians and other clinicians most frequently mention concerns about disruptions in workflow, trouble with the interface and other technical matters, and cost to some extent. Policymakers and other stakeholders, principally hospitals or other large providers, worry most about legal and ethical aspects, the plethora of available technology and lack of standards, costs, and the lack of a sustainable business case. Patients are most concerned about privacy and security and how permission is given to share information.

The attitudes and barriers to a robust and sustainable use of HIE are comparable to those for many other health IT interventions, which include interface, workflow, cost issues, and patient concerns about privacy.

HIE system design according to users' behavior.

Within a single HIE system, different and discernible users' behaviors have been observed [Vest, 2012]. These usage categories ranged from minimalistic system interaction to very detailed and complex patterns of screen views that targeted defined types of information. The complexity and diversity of usage categories suggest that researchers who employ simple measures such as access or acceptance obscure substantial fluctuation in user behaviors. A recent study shows that physicians and those working in the children's emergency department (ED) were more likely to have minimal interaction with the HIE system whereas nurses, public health workers, and pharmacists usually sought demographic and clinical information within a session. These findings point towards multiple ways of HIE and other health information systems improvement.

First, these outcomes make the case for prioritizing the display of information in the HIE system to make it quickly available to users. The most exhaustive approach to matching the HIE system to user information needs would be displaying information completely tailored to the user instead of a uniform interface for all system users. However, attempting to create custom user information displays or even some simpler form of screen reorganization to the tastes of every individual user of an HIE system might be beyond the technological or financial capabilities of HIE organizations, or even software vendors. The variety of utilization types by broad job categories justifies this conclusion. Otherwise, the vast majority of users saw very few screens, which underscores the significance of the first screens viewed by users. Least possible usage could be either through job requirements, like completing intake forms by managerial employees or out of time restrictions, in the case of physicians. Therefore, the specific data elements should be amenable to customized viewing according to end-user characteristics.

Second, the value of creating a master-patient-index and record locator feature into a HIE system. About 11% of analyzed sessions included repetitive searching by users. These recurrences represented greater investments by the user in terms of time and cognitive effort. Other than the administrative job titles with scheduled encounters, users did not make those attempts. For HIE networks, ensuring accurate record linkage, record de-duplication, and enhance searching algorithms might help resolve information seeking faster and end repeated searches.

Third, researchers recommend the development of new means to analyze system user logs. User logs provide the required audit trail to ensure patient privacy. Understanding and categorizing usage behavior can supplement this application by helping identifying inappropriate uses of the system. Individuals who use the HIE system in a way very different than their peers or on dates after health care visits may represent potential privacy threats. Very low levels of deep system usage, for instance the categories of clinical or demographic usage, might signify a training opportunity. Users may not be aware of applicable data obtainable by using the system in more than a minimal fashion. Similarly, HIE is anticipated to yield many gains in terms of safety and quality. Looking at how HIE networks are really used can help to refine expectations and suggest reasonable evaluation measures.

Fourth, some studies [Shapiro, 2007] give guidance for those working to establish information exchange partnerships with other organizations. Clinical information appears to be a greater value or is at least more often the apparent objective. This is consistent with examinations of the types of information users want and get from HIE networks.

Lastly, this conceptualization of utilization can inform evaluations of other health information system-most notably EMRs, which present information in a similar fashion. While in many systems it is feasible to view discrete data elements or single displays comprised of the same data (e.g. vital signs or recent medications), EMR screens allotted to previous orders, history and physicals comprehend multiple pieces of disparate data. Measures of EMR usage will be analogous in that they will include views of screens with multiple types of data. Considering EMRs as the base of HIE activities also illustrates how this measurement method can be expanded.

2.6. Literature Gaps

2.6.1. Limitations of the current HIE system

The youngness of the research field and the topic complexity could be identified as the principals causes of the several gaps and limitation actually present in the HIE system. All these problems could be synthetize in the following categories:

Data quality and completeness. Almost all physicians complain about completeness in the data provided by HIE's systems [Rudin, 2011]. For example, in a specific study textual notes were not included in the HIE system due to confidentiality reasons and, while many physicians understood the privacy concerns, the absence of notes made the HIE much less valuable. On the other hand, for office-based clinicians, a major issue is that hospitals are not contributing any data into the HIE system, severely limiting its value and demanding clinicians to access separate portals in addition to the HIE for an adequate picture of the patient's previous care. The hospitals had projected to integrate its data into the HIE but that functionality had not been completed at the time of this study.

Other gaps in the data provided by functional HIE system have attributed to local practices that withdrew from or opted out of the HIE, comprising a primary care practice of several physicians, significantly reducing the sum total of potentially valuable data in the HIE. For inmates who did visit participating clinical practices, clinicians could not be confident why their HIE searches sometimes returned an insufficiency of results, but they cited two possible reasons: patients occasionally rejected consent, and contributing physicians sometimes did not "lock their notes" on their EMR, a software action that was required to send the clinical data into the HIE repository. Because the patient consent level was quite high (about 95%), the lack of timely note-locking was probably the major reason for completeness issue with HIE data.

Timing to access – "Excessive clicking". In addition to data completeness challenges, many clinicians mentioned usability complications with the HIE. Hospital clinicians thought accessing the HIE through the Web portal involved "too many clicks" [Rudin, 2011]. This was less of a problem in the practices, which were able to access the data provided by the HIE system more easily. Clinicians were also discouraged from using HIE due to the inability to discover easily what changed since the previous visit, the condition to change passwords frequently, and a login and search process for the network portal that could take more than a minute, yet often did not result in new or helpful data.

HIE access was also influenced by many technical difficulties, such as software glitches and versioning issues with the EMRs and hardware resulting in frequent downtimes that lasted hours or longer, even after two years of operation.

Trustworthiness was not a significant factor in accessing the HIE: all providers trusted the accuracy of the data but many would still check it with the patient or another data origin. Technical support for HIE was not found to be helpful enough to clinicians to access the HIE more frequently.

A lack of understanding of physicians' information needs. Physicians and staff access a variety of data reports using HIE system, but they usually access only the default landing page [Campion, 2013]. This is consistent with a study of query-based HIE portal utilization for medically indigent patients where users most frequently accessed only the default, or "gateway," screen listing a patient's most recent encounters [Vest, 2012]. Also, workflow for updating patient consent forms may be contributing to higher acceptance of HIE system. Together, these findings accentuate the relevance of data displayed on the default landing page of HIE networks, as the data would provide clinical utility to users without needing to display a detailed report, and situation of levels of access, such as default only and detailed, when examining usage of query-based HIE portals [Vest, 2012].

Other health professionals, such as nurses and administrative staff, often prepare HIE data for clinicians to use by generating printouts [Unertl, 2012]. Previous researches suggest that improved

HIE designs should incorporate user- and role-specific display of data to accommodate differing information needs [Unertl, 2012; Vest, 2012]. Furthermore, others have noted that regular HIE utilization by nurses and administrative staff was associated with increased overall levels of HIE access in practice sites [Johnson, 2011]. These result highlight the importance of tailoring data displayed by HIE networks. Exposure to meaningful health information in a timely fashion may improve HIE usability. Still, little is known about information needs of the several types of health professionals interacting with outside health information and HIE networks.

Other common HIE-related issues are [Thorn, 2014; Kripalani, 2007] :

1. HIE access influences decision-making process. Indeed, according as it is useful or not, patients avoid repeated and multiple workups, for instance.
2. The access is difficult too many times. It is imperative to have a unique and integrated system, and not multiple log-ins, instead of 2 or 3 step to get into it.
3. Problematic user privileges occur, indeed obtaining HIE privileges is complicated and a clinician could stop trying to get information rather than become irritated.
4. The systems and software applications in HIE are not able to work together, generating interoperability problems. As result, HIE system has functional issues and could not satisfy the information needs.
5. Physicians want to learn HIE functions to access information faster, but they identify training deficits and a lack of technical support. So it is difficult to navigate in the HIE network; physicians want someone to show them the functions and speed up access. They do not have time to learn on their own.

The traditional methods of completing and delivering discharge summaries are suboptimal for communicating timely, accurate, and medically important patient data to the physicians who will be responsible for follow-up care. Urgent improvements are needed in the processes and formats used for transferring information to primary care physicians at hospital discharge. In this new model of care, the discharge summary becomes a vital tool for communication and information transfer. Research is beginning to show that poor information transfer and discontinuity are associated with lower quality of care on follow-up, as well as adverse clinical outcomes.

Unexpected uses of HIE.

To highlight the gap between the information displayed in HIE networks and the health professionals' information needs, we present below unexpected uses of HIE networks reported in the literature [Ozkaynak, 2013]:

- The information from the HIE could be used to confront with the patients.

  The HIE system was used by physicians to confirm patient statements or confront the patient. During clinician interviews, seven physicians mentioned that they used the HIE to catch patients' incorrect statements and intentionally untold history. A quote from the interviews: 'we use it [HIE system] we did not use it today but for example we used it-a lot last weekend it is useful to catch patient's lie'. Clinicians thought that some of the incorrect patient statements can be due to human factors (not remembering or being unconscious). However, physicians also believed that the majority of the incorrect patient statements were aimed at abusing the health system.

- The HIE was being used mostly for patients only with specific characteristics. Clinicians used the HIE system to verify patients' previous visits to the current ED or to other health systems. Clinicians paid attention to whether the patient visited any other ED for a similar complaint on the same day or lately. The observers noted some workflow issues (e.g., the paper HIE sheet

> was not attached by the administrator staff in timely manner) that might be a barrier for HIE use. Not all the patient care episodes suffered from these workflow issues. Observers noted that the prescribers were spending extra effort to overcome the workflow problems (if they were present) and used the system when adult patients with chronic pain were present. The HIE system was not used for patients with other characteristics, even when there was no workflow issues. Prescribers were also asked for drivers of HIE utilization. Some of the collected answers are that the HIE system was used by physicians to develop an opinion or confirm their initial opinion about whether the patient was visiting the ED to receive narcotics unnecessarily (i.e., drug seeker detection).

HIE networks are providing limited and incomplete information about the patient. This system design issues might lead to snap judgments that affect the quality of the patient clinician interaction, and potentially patient health outcomes.

2.6.2. Understanding HIE Usage Patterns can help Improving HIE design

Data mining is the process of analyzing large amounts of data from different perspectives and summarizing it into useful information. The data can be converted into knowledge about historical patterns and future trends. Data mining has a relevant role in the area of information technology [Pujari, 2001]. Today, the healthcare industry generates large volume of complex data about patients, diseases, hospitals resources, electronic patient records and diagnosis methods. However, huge quantities of healthcare data are not mined to discover hidden information for effective decision-making. The discovered knowledge can be adopted by the healthcare administrators to improve the quality of service. Data mining functions include clustering, classification, prediction and association rule discovery. Several studies have used clustering, classification and prediction data mining techniques [Soni, 2011; Ordonez, 2004; Antonie, 2001; Barati, 2011; Bellaachia, 2006; Subbalakshmi, 2011; Deepika, 2011].

One of the most important data mining utilization is association rules mining. Association rules, first introduced in 1993, are used to identify association between a set of items in transactional databases. Traditionally, Apriori algorithm has been successfully used for finding frequent item sets in retail data. More recently, it has been used for exploring EMR data to generate the association rules in medical billing data. For example [Abdullah, 2008] determines the frequency of diseases in particular geographical area at given time period with the support of association rules. Moreover, *"Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining; Umair Abdullah, Jamil Ahmad, Aftab Ahmed"* [Abdullah, 2008] finds associations among diagnosis and treatments. Affinity between medical bill and purchase bill is the motivation of using Apriori algorithm in this research project. Healthcare is a data rich field. Insurance companies, medical practices, and other health related organizations have collected large amount of data, thus attracting data mining researchers to examine it and find something beneficial from it. After having all the results and trying different plans of action, it is found that it is a priori property (i.e. a frequent set can only be generated from frequent subsets) which makes Apriori algorithm suitable for finding frequent item-sets from billing database. Even if several modifications have been made in the algorithm to overcome with the requirement, but basic rule generation process is same as that of Apriori algorithm.

2.7. Aim of our Research

As explained in the previous paragraph, actually the healthcare system is afflicted by different kinds of lack and gaps concerning integration and data sharing. In this research we try to find efficient solutions in particular with respect to the problems of data quality, the timing to access data, the lack of understanding physicians needs and the missing integration of different IT system in different healthcare structures. In particular, our goal is to investigate the relation among specific diagnosis and information requested, i.e., identifying the data most commonly requested by physicians to treat the most frequent diagnosis. This objective is significant because in this way we think that it is possible to improve the integration of information of healthcare services, reducing time for diagnosis and providing physicians with the data they need, by integrating different sources and avoiding "excessive clicking" phenomenon [Rudin, 2011]. For example, a patient in critical condition due to a severe infection is transferred to the IM of our hospital. Having access to useful health information about previous bacterial cultures and the workup performed at the transferring hospital might save time and resources at the admitting hospital. Discovering the patient-specific clinical information physicians need the most, will help to develop improved HIE networks. An improved HIE system should provide both quick access to the most useful information based on user and patient type and access to all outside patient-specific clinical information. Our expected results are a set of strong association rules that could be fundamental guidelines for the design and the implementation of new HIE, which should simplify the access at the most frequently requested data and incentivize the integration of data between different hospitals and clinics. In this way doctors will be motivated to use the new information system because problem like the "excessive clicking" will be reduced and solved. The other relevant result that we plan to achieve is the reduction of problem linked with the reliability of the data. In particular, if physicians will find the new system reliable and useful ,we expect that the total load of re-requested information will decrease quickly and this is a relevant efficiency increase both for the single hospital and for the total healthcare system.

3. METHODS

Starting from a data driven approach, we show how data mining could have a direct impact on data quality, timing to access, physician needs and an indirect though relevant impact on the general integration of HIE system. Following this idea, in this study we investigate the data of the most important teaching hospital in the Tampa Florida Area, trying to discover connections between the different types of information. Our goal is to improve the HIE system providing a new efficient tool that simplifies the access of physicians to the most requested information, in particular we want to track and understand the most significant hidden relation among health information exchanged about the most frequent health disease. The theory that we follow to achieve this goal is called affinity analysis and we use it to find co-occurrence relationships among patient-specific clinical information requests to outside healthcare providers.

Institutional review board approval was granted for this study by the Office of Clinical Research of the hospital and the University of South Florida (IRB Number: Pro00014574).

The Affinity Analysis.

Data mining is the process of discovering patterns and useful information in large datasets. Affinity analysis, widely used for market basket analysis, is a tool to discover co-occurrence relationships between events performed by specific agents. This technique is useful for processes where agents can be uniquely identified and their activities can be accurately recorded. Affinity analysis is commonly used in the retail sector to perform market basket analysis, where sellers try to predict the purchase behavior of customers. In our study, we analyze the information "purchase" behavior of physicians from outside health care facilities. Affinity analysis will tell HIE developers and software designers the set of data items requested the most by physicians to outside health care facilities. We assume that there is a linear positive relationship between request frequency and usefulness. These data items will be presented on the first screens of an improved HIE system to reduce the time and effort needed to access useful clinical information.

Affinity Analysis Steps.

Various tasks were performed to complete the affinity analysis. First, we went through a data preparation process including missing value and outlier analysis. Second, we analyzed the co-occurrence data using Apriori algorithm. As a result, a set of several association rules was generated. Third, we evaluated the strength of each association rule using confidence and support. Fourth, we conducted a pruning phase to select those association rules with high support and confidence. Finally, we assessed the set of generated association rules with field experts.

3.1. Study Setting and Research Team

The hospital considered in our study is a private not-for-profit hospital and one of the most comprehensive medical facilities in West Central Florida serving over 4 million people from 23 different counties. This medical structure with its 1,018 beds and 6,600 employees admitted 42,129 patients during Fiscal Year 2014. Additionally, this hospital is the primary teaching facility to the USF Health Morsani College of Medicine. Over 300 residents are assigned to this teaching institute for specialty training in areas ranging from internal medicine to neurosurgery. USF medical students, nurses and physical therapy students all receive part of their training at in this clinic . Faculty of the USF Health Morsani College of Medicine admit and care for patients at this hospital as do community physicians, many of whom also serve as adjunct clinical faculty. In the case of the IM of this medical structure, there is a team of 10 physicians affiliated to both hospital and USF.

The research team comprises two graduate students from Politecnico di Milano, three industrial engineering researchers from University of South Florida, and four physicians from the hospital studied. This multidisciplinary approach allowed us to analyze the system from different perspectives. During weekly research meetings, we checked, reviewed, and plan the tasks to be completed. This particular study was part of a bigger research project about health information exchange and its impact on clinical care lead by the USF's Industrial and Management Systems Engineering (Principal Investigator: Dr. Zayas-Castro) and co-investigators from USF-Health.

3.2. Dataset & data preparation

The dataset comprised demographic, clinical and operational EMR data for all admissions at the IM of the hospital from October 2011 to March 2014. The demographic data included randomly generated unique admission and patient identifiers, age, gender, language preference, marital status and payer class. The clinical data included admission and discharge dates, admission source, presence or not of primary care physicians, discharge disposition and the list of medical procedures and health problems. The operational data included date and time for OI (outside information) request, type of provider authorizing OI request and the type of OI received.

Dataset Dimension.

Fourteen percent (2,089/15,230) of the consecutive hospitalizations seen by the internal medicine department from October 2011 to March 2014 generated at least one request for OI. These 2,089 hospitalizations (corresponding to 2,002 unique patients) generated a total of 3,508 requests for OI to outside healthcare facilities.

Data Preparation.

Data preparation was a central and relevant part of our project. Since the operational data we requested is not typically used by hospital analysts, the initial database contained several useless data to reach our objective. To prepare the dataset, we followed four steps: record selection, attribute check and conversion, attribute selection, and the creation of tables.

Record selection.

The original dataset included every request for OI as a new patient admission. In other words, that admission that generated more than one request for OI were duplicated. The explanation of these re-request for OI is that either there was an excessive delay to receive OI so the hospital personnel had to re-request or the OI received was not useful so more OI request were placed. A total of 824 out of 2,913 (28.3%) were re-request for OI so the linked admissions were eliminated.

Attribute check and conversion.

We analyzed each attribute in the dataset to clean and convert them if needed. The aim of this conversion part was transform the original data categories into new ones without loss of generality or information. To do this, we reduce the initial number of attributes where there was a category with several different values and we maintain the original for those with few possible choice. In both of case we generate new dummy variable to count if a patient presents or nor a specific attribute. All the originals and new categories are reported in the Appendix B.

Attribute Selection & Tables creation.

This was the last phase of data preparation. At first we selected only the variable essential to be implemented with the Apriori Algorithm: the first 30 referred to the presence or not of a specific symptom and the last 9 referred to the presence or not of a specific Outside Information requested. In this way we obtained a new dataset composed by 2089 records (rows) and 39 dummy variables (columns).

In the attempt to obtain significant and useful rules, we tested three different kind of input table. The first one was composed by all Diagnosis and Outside Information columns, the second was structured with only one Outside Information and all the Diagnosis, the third was the dual of the second, with only one Diagnosis and all Outside Information types.

After these test we chose to work only with the first configuration cause the Diagnosis low-frequency level did not allow us to obtain strong results, so we prepared the General Case Table (Figure 3.1).

Figure 3.1) General Case Table

	Anemia	**Chest Pain**	**Other Diagnosis**	**Outside Medical Record**	**Outside Labs**	**Outside Imaging**
1	1	0	0	1	1	0
2	0	1	0	0	0	0
3	0	0	1	0	1	1
….	1	1	0	1	0	1
2089	0	0	1	0	0	0

3.3. The Apriori Algorithm

To better understand what the Apriori Algorithm is, it is necessary to introduce some concepts from the Association Rule Theory. We want to introduce this kind of theory starting from the definition of *itemset* and *association rule* proposed by *Pang, Steinbach, Kumar* [Adams, 2008]. Let $I = \{i_{1,}i_{2,} \ldots,\ i_d\}$ is the set of d different items in a market basket data and $T = \{t_1, t_2 \ldots, t_N\}$ be the set of N transactions. Each transaction t_i contains a subset of items chosen from I. A *k-itemset* is defined as a collection of k items, where$k\ = \{1, \ldots, K\}$. Additionally, an *association rule* is an implication expression of the form $X \rightarrow Y$, where X and Y are disjoint itemsets and $\sigma()$ is a count function. The strength of an association rule can be measured in terms of its *support* and *confidence***.** *Support* denotes how often a rule is applicable to a given data set while confidence indicates how frequently items in Y appear in transactions that contain X. The correct definitions of these metrics are:

$$\textbf{Support}, \boldsymbol{s}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\sigma(X \cup Y)}{N}$$

$$\textbf{Confidence}, \mathbf{c}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

To measure the validity of the association rules found by the Apriori algorithm, we used support, confidence and lift performance measures. Association rules having low support may occur simply by random case. A low level of support rule is also likely to be unattractive from the HIE developer or implementer perspective, because it may not be effective to promote items that physicians seldom request together. For this these reasons, support is often used to eliminate dull rules and has the desirable property that can be exploited for the efficient discovery of association rules. Confidence, on the other hand, is the conditional probability of requesting itemset Y given that itemset X has been requested. In other words, confidence measures the accuracy of the association rules because reports the probability to have Y (the consequent) if it is already present X (the antecedent) and shows the real prediction power of the rule.

Support and confidence are used to determine if a rule is valid. However, there are occasions in which both of these performance values may be high, and yet still have as outcome a rule that is not valuable. Look at this example:

“Store customers who buy cookies also buy tea with a 70% confidence. Tea and cookies combination has a support of 25%.”

This rule, sounds like an optimal rule, and in most cases, effectively it would be. It has high levels of confidence and good support. Conversely, what if convenience store customers in general buy tea 85% of the time? In that circumstance, cookies customers are actually *less* likely to buy tea than customers, in the general case. Starting by this view results necessary a third measure to evaluate the rule quality and values. This measure is called Lift and indicates if a rule is strength over than the possible random co-occurrence of antecedent and consequent, using their individual support. Lift provides additional information about the improvement and the increase in the probability of the consequent, having already fixed the antecedent. Lift is definite as here:

$$\textbf{Lift}, \mathbf{l}(\boldsymbol{X} \to \boldsymbol{Y}) = \frac{s(X \cup Y)}{s(X)s(Y)}$$

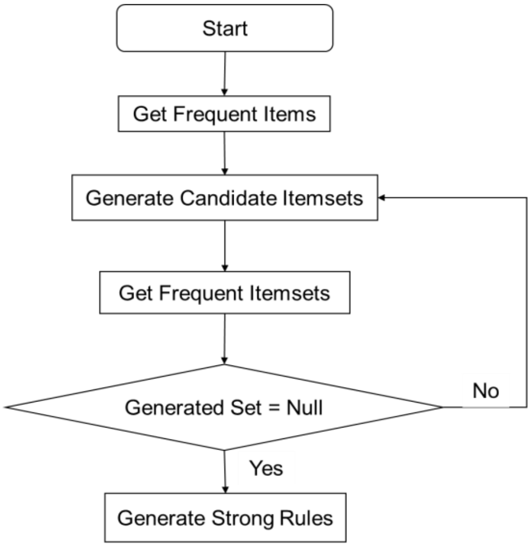
It can also be defined as the confidence of the combination of the different items divided by the support of the consequent. So in the tea example, assuming that 45% of the customers buy cookies, the improvement would be: 25% / (45% * 85%) = **65%<100%.**

Each rule with improvement lower than 1 indicates a not real "cross-selling" opportunity since it offers less ability to predict customers' behavior, so the results are less reliable.

Association Rule Discovery.

Given a set of patient-specific clinical information transactions with outside healthcare facilities, called *T*, our goal is to find those association rules such that $s \geq minsup$ and $c \geq minconf$, where $minsup$ and $minconf$ are arbitrarily selected support and confidence thresholds. A common strategy adopted by many associations rule mining algorithms is to decompose the problem into two major subtasks: Frequent Itemset Generation and Rule Generation. Frequent Itemset Generation finds all the itemsets satisfying the $minsup$ threshold. These itemsets are called *frequent itemsets*. Rule Generation, on the other hand extracts all the high-confidence rules from the frequent itemsets. These extracted rules are termed *strong rules*. In our study, Apriori algorithm has been used because it is one of the most widely used association rule mining tools in healthcare [Abdullah, 2008; Srikant, 1997; Sharma, 2014; Ilayaraja, 2013], in fact is one of the most intuitive algorithm and exists several numbers of modified version applied in different case studies. Moreover, using the support measure it is possible to reduce the number of candidate itemsets explored during Frequent Itemset Generation.

Figure 3.2) Apriori Flowchart



Why Apriori?

Apriori Algorithm is the most commonly used association rule mining methods [61]. The intuitive structure is easy to implement characteristics makes it one of the most studied rule mining algorithms by the data mining research community.

Apriori algorithm follows this principle: If an itemset is frequent, then all of its subsets must also be frequent, and vice versa [Adams, 2008].

Another important characteristic of Apriori is it Monotonicity Property.

Monotonicity Property: Let I be a set of items, and $J = 2^I$ be the power set of *I.*

A measure *f* is monotone (or upward closed) if:

$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(X) \leq f(Y),$$

which means that if *X* is a subset *of Y*, then *f(X)* must not exceed *f(Y)*.

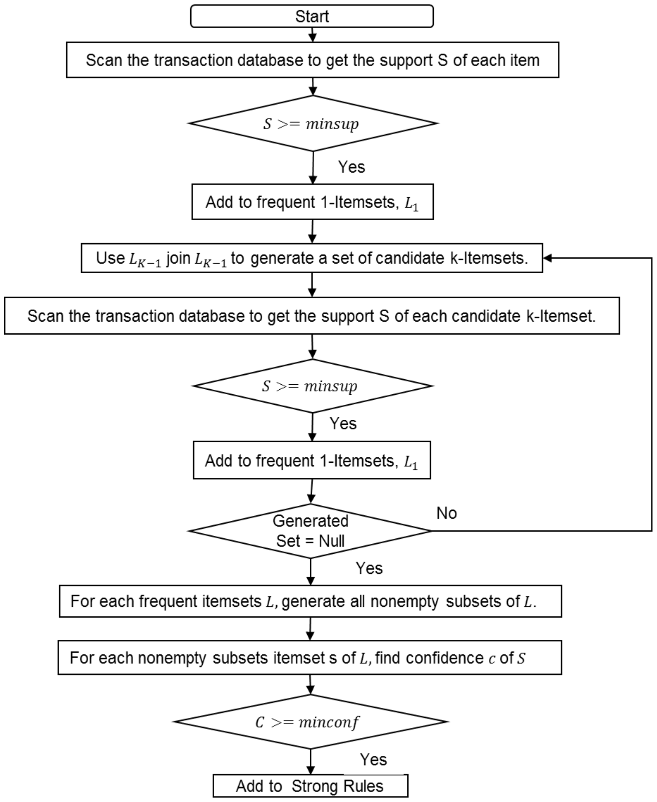
However, *f* is anti-monotone (or downward closed) if

$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(Y) \leq f(X),$$

which means that if *X* is a subset of *Y*, then *f(Y)* must not exceed *f(X).*

These lasts properties are really relevant in our cases because we work with a huge database and thank to these properties we are able to reduce the timing and the calculation capacity of the generation process, and this is one of the main motive of our Apriori choice.

Figure 3.3) Apriori Stages


Start
Scan the transaction database to get the support S of each item
$S >= minsup$
Yes
Add to frequent 1-Itemsets, L_1
Use L_{K-1} join L_{K-1} to generate a set of candidate k-Itemsets.
Scan the transaction database to get the support S of each candidate k-Itemset.
$S >= minsup$
Yes
Add to frequent 1-Itemsets, L_1
Generated Set = Null
No
Yes
For each frequent itemsets L, generate all nonempty subsets of L.
For each nonempty subsets itemset s of L, find confidence c of S
$C >= minconf$
Yes
Add to Strong Rules

Frequent Itemset Generation in the Apriori Algorithm.

Let C_k denote the set of candidate k-itemsets and F_k the set of frequent k-itemsets.

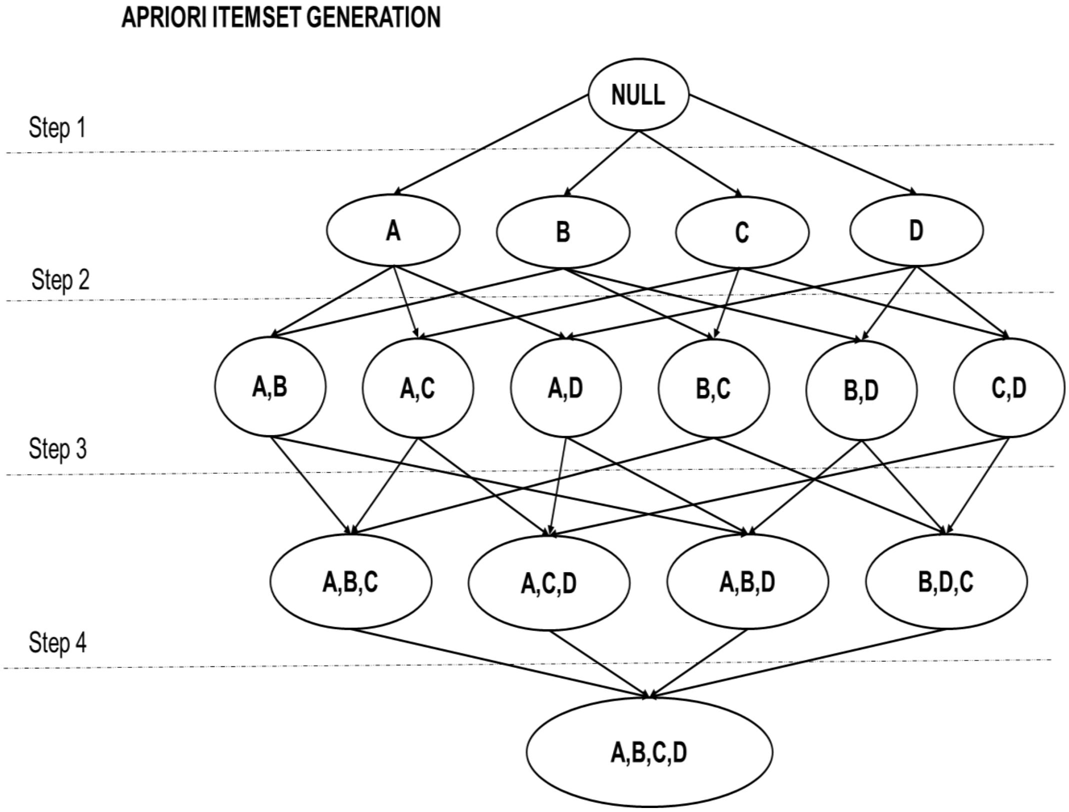
First of all, the algorithm makes a single revision over the dataset to determine the support of each 1-itemset. Upon completion of this step, the algorithm determines F_1 using the selected $minsup$. Second, the algorithm will iteratively generate new candidate k-itemsets using the frequent $(k-1)$ – itemsets. The candidate generation is implemented using a function called Apriori-gen. To count the support of the new candidates, the algorithm needs to make an additional pass over the dataset. The Apriori-gen function is used to determine all the candidates in C_k that are contained in each transaction t. After determining their supports, Apriori algorithm removes all the candidate itemsets whose support counts are less than $minsup$. The algorithm terminates if $F_k = \emptyset$, i.e., when there are no new candidate frequent itemsets.

The frequent itemset generation has two important characteristics.

First, it is a level-wise algorithm that traverses the itemset lattice (i.e., the itemset structure shown in Figure 3.4 on level at a time, from frequent 1-itemsets to the maximum size of frequent itemsets). Second, it employs a generate-and-test strategy for finding frequent itemsets. At each iteration, new candidate itemsets are generated from the frequent itemsets found in the previous iteration. The support for each candidate is then counted and tested against the $minsup$ threshold.

The total number of iterations needed by the algorithm is $k_{max} + 1$, where $k_{\max}$ the maximum size of the frequent itemsets is.

Figure 3.4) Apriori Lattice Structure



Candidate Generation and Pruning.

The Apriori-gen generates candidate itemsets by performing the following two operations; Candidate Generation and Candidate Pruning. During Candidate Generation, new candidate k-itemsets are generated based on the frequent $(k-1)$ - itemsets.

During Candidate Pruning, the new candidate k-itemsets are discarded using a support-based pruning strategy. To illustrate the candidate pruning operation, consider a candidate k-itemset, $X = \{i_1, \dots, i_k\}$. The algorithm must determine whether all of its subsets,$X - \{i_j\}, (\forall j = 1, \dots, k)$, are frequent or not. If one or more of its subsets is infrequent, then the candidate itemset X is immediately pruned. This approach can effectively reduce the number of candidate itemsets instead of consider the total combinations number. The complexity of this operation is $\Theta(k)$, where $\Theta()$ represent the total complexity function and k the passages number. For each candidate k-itemset. However, there is no need to evaluate all k subsets of a given candidate itemset. If m subsets, where$m \subseteq k$, were used to generate a candidate, we only need to check the remaining $k - m$ subsets during candidate pruning operation. Add a sentence explaining how much (approximately) the complexity is reduced by this fact.

Apriori Computational Complexity: key criteria for support level choice.

Lowering the **support threshold**, $minsup$, often results in more itemsets being declared as frequent. This has an adverse effect on the computational complexity of the algorithm because more candidate itemsets must be generated and counted. Additionally, the maximum size of frequent itemsets tends to increase with lower $minsup$. As the maximum size of the frequent itemsets increases, the algorithm needs to make more iterations over the dataset.

Another factor influencing the complexity of Apriori is the **data dimensionality**. As the number of item types increases, more space will be needed to store the support counts. If the number of frequent items also grows, with the dimensionality of the data, the computation and input/output costs will increase because of the larger number of candidate itemsets generated by the algorithm. Also, and since the Apriori algorithm makes repeated passes over the dataset, its complexity increases with a larger **number of transactions**. Furthermore, for **dense datasets** the average transactions width can be very large. This affects the complexity of the Apriori algorithm in two ways. First, the maximum size of frequent itemsets tends to increase as the average transaction width increases. Like a consequence, more candidate itemsets must be examined during candidate generation and support counting. Second, as the transaction size increases, more itemsets are enclosed in the transaction. This will increase the number of candidate itemsets to be examined during support counting.

Generation of frequent 1-itemsets: For each transaction, we need to update the support count for every item present in the transaction. Assuming that ω is the average transaction width, this operation requires O(N ω) time, where N is the total number or transactions.

Candidate Generation: To generate candidate k-itemsets, pairs of frequent $(K-1)$-itemsets are merged to determine whether they have at least $k-2$ items in common. Each merging operation requires at most $k-2$ equality comparisons. In the best-case scenario, the algorithm must merge every pair of frequent $(K-1)$-itemsets found in the previous iteration.

Hence, the overall cost of merging frequent itemsets is

$$\sum_{k=2}^{\omega}(k-2)|C_k| < Cost\ of\ merging < \sum_{k=2}^{\omega}(k-2)\ |F_{k-1}|^2$$

A hash tree is also constructed during candidate generation to store the candidate itemsets. Because the maximum depth of the tree is *k*, the cost for populating the hash tree with candidate itemsets is $O(\sum_{k=2}^{\omega} k|C_k|)$. During candidate pruning, we need to verify that the $k-2$ subsets of every candidate k-itemset are frequent. Since the cost of looking up a candidate in a hash tree is $O(k)$, the candidate pruning step requires $O(\sum_{k=2}^{\omega} k(k-2)|C_k|)$ time.

Support Counting: Each transaction of length $|t|$ produces $\binom{|t|}{k}$ itemsets of size *k.* This is also the effective number of hash tree traversals performed for each transaction. The cost for support counting is $O\left(N \sum_k \binom{\omega}{k} \alpha_k\right)$, where ω the maximum transaction width is and α_k is the cost for updating the support count of a candidate *k*-itemset in the hash tree.

Rule Generation in Apriori Algorithm.

The Apriori uses a level-wise approach for generating association rules, so following this scheme each level corresponds to the number of items that belong to the rule consequent. Initially, all the high-confidence rules that have only one item in the rule consequent are selected. These association rules are then used to generate new candidate rules.

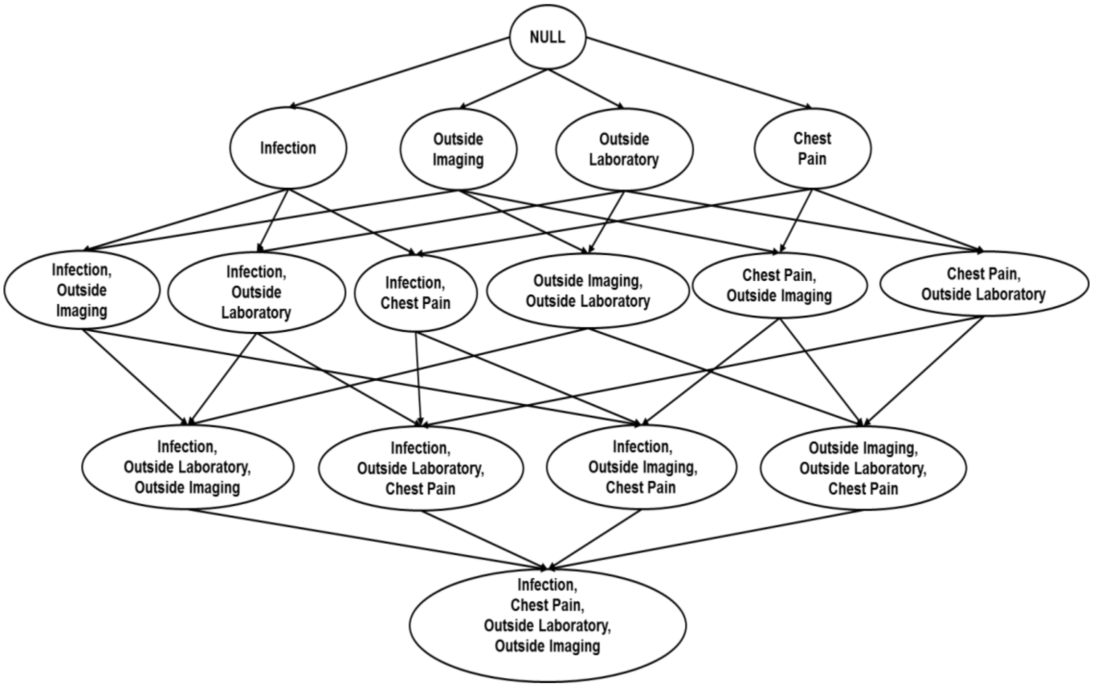
Apriori: a HIE example.

Consider the following three patient admissions to the IM of our medical structure. Patient 1 exhibits an acute infection. The attending physician decides to request previous lab results from an outside laboratory, as well as imaging results performed in a previous hospitalization in a near hospital. Patient 2 is admitted with an infection, and its medical team decides to obtain imaging and laboratory tests results performed at an outside clinic. Finally, patient 3 is admitted with severe chest pain. The attending physician requests to her resident physicians to obtain the results of a previous lab performed in an external laboratory. In this hypothetical example, outside patient-specific clinical information resides in 4 information silos; two external laboratories, a hospital and an outside clinic. Accessing this OI through a fax-based information exchange system would require several hours and lot of effort from health professionals. Accessing this information through an HIE system would require fewer hours and effort, but still clinicians will have to navigate through several screens and performing lot of clicking around. Instead, an improved HIE system would present the most useful patient-specific information in its first screens. To identify the most useful information, we proposed to use Apriori algorithm in the transactional data described in Figure 3.5. In this simple example, we will show how Apriori algorithm is capable of identifying the most useful information out of the transactional data.

Figure 3.5) Table Example

Patient Admission Identifier	Chest Pain	Infection	Outside Laboratory	Outside Imaging
1	0	1	1	1
2	0	1	1	0
3	1	0	1	0

Figure 3.6) Lattice Example


NULL
Infection
Outside Imaging
Outside Laboratory
Chest Pain
Infection, Outside Imaging
Infection, Outside Laboratory
Infection, Chest Pain
Outside Imaging, Outside Laboratory
Chest Pain, Outside Imaging
Chest Pain, Outside Laboratory
Infection, Outside Laboratory, Outside Imaging
Infection, Outside Laboratory, Chest Pain
Infection, Outside Imaging, Chest Pain
Outside Imaging, Outside Laboratory, Chest Pain
Infection, Chest Pain, Outside Laboratory, Outside Imaging

In the Figure 3.6 clearly appear how Apriori proceed: in the first step the algorithm generate an empty list., then it will consider all the different attributes present and step by step will combine these only if their support level results higher than the thresholds values obtaining itemsets always more numerous till the end of possible combinations.

This example might sound simple but useful to understand how Apriori algorithm is able to find useful association rules between data-items request to outside healthcare facilities. For instance, considering the values of the example in Table 3.5, we can evaluate the relation between the voices Infection and Outside Imaging:

Support: Supp(Infection→Outside imaging) $= \boldsymbol{s}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{\sigma}(\boldsymbol{X} \cup \boldsymbol{Y})}{\boldsymbol{N}} = \frac{\mathbf{1}}{\mathbf{3}}$

Supp(Outside Imaging→Infection) $= \boldsymbol{s}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{\sigma}(\boldsymbol{X} \cup \boldsymbol{Y})}{\boldsymbol{N}} = \frac{\mathbf{1}}{\mathbf{3}}$

Confidence: Conf(Infection→Outside imaging)= $\boldsymbol{c}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{\sigma}(\boldsymbol{X} \cup \boldsymbol{Y})}{\boldsymbol{\sigma}(\boldsymbol{X})} = \frac{\mathbf{1}}{\mathbf{2}}$

Conf(Outside imaging→Infection)= $\boldsymbol{c}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) = \frac{\boldsymbol{\sigma}(\boldsymbol{X} \cup \boldsymbol{Y})}{\boldsymbol{\sigma}(\boldsymbol{X})} = \frac{\mathbf{1}}{\mathbf{1}} = \mathbf{1}$

As showed by this example, in the support the items order is not relevant, but in the confidence it is. Considering the total lattice structure, Apriori is able to calculate all support and confidence automatically:

Figure 3.7) Support & Confidence Example Table

Itemset = 1	**support**	**confidence**
Chest pain	1/3	1
Infection	2/3	1
Outside Imaging	1/3	1
Outside Laboratory	1	1
Itemset = 2		
Chest pain, Infection	0	0
Chest pain, Outside Laboratory	1/3	1
Chest pain, Outside Imaging	0	0
Infection, Outside Imaging	1/3	1/2
Outside Imaging, Outside Laboratory	1/3	1
Infection, Outside Laboratory	2/3	1
Itemset = 3		
Chest Pain, Infection, Outside Laboratory	0	0
Chest Pain, Infection, Outside Imaging	0	0
Outside Laboratory, Infection, Outside Imaging	1/3	1
Outside Laboratory, Chest Pain, Outside Imaging	0	0
Itemset = 4		
Chest Pain, Infection, Outside Laboratory, Outside Imaging	0	0

Fixing for example minsup = 0.33 and minconf = 1 Apriori reports these rules:

Figure 3.8) Rules ExampleTable

Antecedent	**Consequent**	**Support**	**Confidence**
-	Chest Pain	1/3	1
-	Infection	2/3	1
-	Outside Imaging	1/3	1
-	Outside Laboratory	1	1
Chest pain	Outside Laboratory	1/3	1
Outside Imaging	Outside Laboratory	1/3	1
Infection	Outside Laboratory	2/3	1
Outside Laboratory, Infection	Outside Imaging	1/3	1

The previous Rules Example Table 3.8 well showed which the real contribute of this kind of analysis. In the first column is reported the name of the antecedent, in the second the consequent. The rules with empty space in Antecedent column describe just the probability to have a specific element (as for example “Chest Pain” has an occurrence of 1/3), while all the other rules explain the occurrence of itemsets with both antecedent and consequent when we look at support value and the probability to find the consequent if we already have the antecedent if we check the confidence. This means that for istance, looking at the fourth rule, that we have 1/3 of probability to find “Chest Pain” and “Out lab” in the same itemset and we have the 100% of probability to have “Out lab” in a itemset that we know contain “Chest Pain”.

In synthesis, this simple example shows how Apriori works with the real data, first generating al the possible itemsets and the pruning results with support and confidence lower than *minsup* and *minconf*. The final result is a set of strong rules, as shown above.

Outcome of the Study.

The proposed study will identify the frequent outside patient-specific medical information in a large dataset. This outcome will help developers and implementers in making HIE networks more useful. Improved HIE networks should provide quick access to useful information for medical-decision making.

4. RESULTS

4.1. Population

Our dataset contains information of 2089 number of patient admissions from October 2011 to March 2013 to an internal medicine department of a general hospital. These information are classified by Gender, Age, Language, Marital Status, Primary Care Physician, Payer Class, Admission Source, Payer Class, Provider Type, Length of Stay and Discharge Disposition. The meaning of some of these categories is explicit and intuitive, as for Gender, Age, Language, Marital Status, Length of Stay, but for the others it is necessary add a simple explanation. Primary Care Physician represents the presence or not of a relation between the patient and a primary doctor, Payer Class shows the different types of patient insurance, Admission Source reports the way through that the patient was admitted, Provider Type is a classification of the diverse kinds of personnel requesting Outside Information and Discharge Disposition reports the possible discharge ways of patient. All these characteristics of the teaching hospital patient data used for this study are reported in the next Figure 4.1.

Figure 4.1) Population Characteristics Table - Demographic and clinical factors of hospitalizations with a request for outside information. Abbreviations: HCHCP, Hillsborough Country Health Care Plan.

Population Characteristics (N=2,089)		**No. (%)**
GENDER	Male	1030 (49,3)
	Female	1059 (50,7)
AGE (average= 54.012)	0-20	27 (1,3)
	21-40	452 (21,6)
	41-60	855 (40,9)
	0ver 60	755 (36,1)
LANGUAGE	English	1948 (93,3)
	Spanish	94 (4,5)
	Others	47 (2,2)
MARITAL STATUS	Married	650 (31,1)
	Single	1361 (65,2)
	Others	78 (3,7)
PRIMARY CARE PHYSICIAN	Yes	1290 (61,8)
	No	799 (38,2)
PAYER CLASS	Medicaid	465 (22,3)
	Medicare	804 (38,5)
	Others	820 (39,3)
ADMISSION SOURCE	Self-Referral	1916 (91,7)
	Physician or Clinical Referral	84 (4)
	Outside Hospital	84 (4)
	Emergency Room	3 (0,1)
	Others	2 (0,1)
PROVIDER TYPE	Resident	1718 (82,2)
	Physician	298 (14,3)
	Nurse Practitioner	36 (1,7)
	Physician Assistant	17 (0,8)
	Physician Assistant/Physical Therapy	13 (0,6)
	Anesthesiologist	4 (0,2)
	Null	2 (0,1)
	Dentist	1 (0)
LENGTH OF STAY (average=6.788 days)	0 days	4 (0,2)
	1 days	280 (13,4)
	2 days	372 (17,8)
	3 days	292 (14)
	4 days	211 (10,1)
	5 days	162 (7,8)
	6 days	143 (6,8)
	7 days	108 (5,2)
	8 days	76 (3,6)
	9 days	61 (2,9)

	10 days	52 (2,5)
	Over 10	328 (15,7)
DISCHARGE DISPOSITION	Transfer at Home	1512 (72,4)
	Expired	65 (3,1)
	Transfer another health facility	452 (21,6)
	Others	60 (2,9)

4.2. Diagnosis & Outside Information

Using the software R, we mined the data to obtain a set of relationships between transactions of patient-specific medical information with outside health care facilities.

We obtain different possible output just changing the two main parameters, which are support and confidence. Starting from our dataset, we drew several rules using different constraints, for example fixing minimum levels of support, confidence and lift or asking to obtain a set of rules with a specific Diagnosis or Outside Information inside. After these experiments, we chose to work selecting only the most significant twenty association rules which present at least one diagnosis in the antecedent side. This choice follows the attempt to find the most valuable relations between each kind of diagnosis and an outside information document in order to provide rules useful. Following are present Figure 4.2 and Figure 4.3 that report respectively the 30 diagnosis diseases and the OI selected for our analysis with Apriori Algorithm.

Figure 4.2) Diagnosis List Table: this table reports the 30 diagnosis adopted in our analysis, the last two columns on the right represent respectively number of cases and the percentage on the total number of admission considered.

	Diagnosis	**N**	**%**
1	Chest pain	387	19%
2	Others Diagnosis	325	16%
3	Abdominal pain	315	15%
4	Anemia	261	12%
5	Dyspnea	206	10%
6	Hypertension	199	10%
7	Diabetes mellitus	195	9%
8	Leukocytosis	182	9%
9	Renal Failure	177	8%
10	Vomiting	152	7%
11	Nausea	150	7%
12	Altered mental status	133	6%
13	Fever	122	6%
14	Cancer	109	5%
15	Tachycardia	107	5%
16	Hypotension	100	5%
17	Lower urinary tract infection	97	5%
18	Hypokalemia	96	5%
19	Hyponatremia	92	4%
20	Back pain	88	4%
21	Syncope	88	4%
22	Coronary artery disease	84	4%
23	Pneumonia	81	4%
24	COPD	78	4%
25	CHF	76	4%
26	GI bleed	75	4%
27	Cellulitis	73	3%
28	Headache	69	3%
29	Alcohol	69	3%
30	Weakness	66	3%

Figure 4.3) Outside Information List Table: these table reports a list of the categories in which are classified all the different types of Outside Information requested. Each following voice could represent only a specific document or a mix of diverse information grouped in a more general class.

	Outside Information	**n**	**%**
1	Outside Medical Record	1635	78%
2	Outside Labs	389	19%
3	Outside Imaging	382	18%
4	Outside History and Physical	255	12%
5	Outside Note	206	10%
6	Outside Consultation	173	8%
7	Outside Discharge Summary	164	8%
8	Outside EKG	153	7%
9	Outside Surgery/Procedure	151	7%

4.3. Association Rules between Health Problems and Outside Information Type

We generated our results starting from the most general and simple data conditions, that is without strong constraints or specific request. In particular we fixed a minimum level of support over 2%, for confidence at least of 75% and lift greater than 1. These values were chosen considering data properties and others academic paper based on data mining tool. Therefore, the output has a little of unusual values and it appears generic because the consequent columns presents only one type of Outside Information, Outside Medical Records, as shown in the Figure 4.4. We expected a similar result, indeed Outside Medical Record is requested in the 78% of cases of Outside Information.

Figure 4.4) Apriori Rules Results Table

	Antecedent	Consequent	Support	Confidence	Lift	N adm
1	Abdominal pain	Outside Medical Record	0.12	0.83	1.059	261
2	Anemia	Outside Medical Record	0.10	0.80	1.028	210
3	Dyspnea	Outside Medical Record	0.08	0.79	1.011	163
4	Hypertension	Outside Medical Record	0.08	0.81	1.040	162
5	Diabetes mellitus	Outside Medical Record	0.08	0.82	1.042	159
6	Renal Failure	Outside Medical Record	0.07	0.83	1.061	147
7	Cancer	Outside Medical Record	0.05	0.88	1.125	96
8	Lower urinary tract infection	Outside Medical Record	0.04	0.86	1.093	83
9	Hypotension	Outside Medical Record	0.04	0.83	1.060	83
10	Back pain	Outside Medical Record	0.04	0.85	1.089	75
11	Pneumonia	Outside Medical Record	0.03	0.89	1.136	72
12	Chest pain, Outside Imaging	Outside Medical Record	0.03	0.93	1.194	71
13	Anemia, Outside Labs	Outside Medical Record	0.03	0.93	1.190	68
14	Abdominal pain, Nausea	Outside Medical Record	0.03	0.83	1.059	63
15	Abdominal pain, Vomiting	Outside Medical Record	0.03	0.85	1.088	63
16	CHF	Outside Medical Record	0.03	0.82	1.042	62
17	Anemia, Outside Imaging	Outside Medical Record	0.03	0.94	1.195	58
18	Hypertension, Diabetes mellitus	Outside Medical Record	0.03	0.85	1.087	57
19	Abdominal pain, Vomiting, Nausea	Outside Medical Record	0.03	0.85	1.081	55
20	Chest pain, Outside EKG	Outside Medical Record	0.02	0.98	1.252	48

In the Figure 4.4 twenty lines of connection between items are presented and each of them represents a rule; indeed we chose to collect only the twenty most correlated relations. More rules can be generated; however, we decided to focus on those with a support level higher than 2%. Association rules with a support lower than 2% will have a limited impact of HIE. The output as shown in table 4.4 is organized from the rule with the highest support to the lowest. As shown in table 4.4 we got top level of support equal to 12% and the least equal to 2%. On the other hand, confidence ranges from 79 to 98%. On the antecedent side we have the twenty most recurrent diagnosis and at least one diagnosis for each rule. On the consequent side we drew the most common outside information documents, that in this case is outside medical record. In the left column we always obtained values greater than 1 and that is an independence indicator, meaning that the consequent item occurrence is not affected by the occurrence of the item in the antecedent. The last column shows the number of patient admissions in which a rule occurs. Outside medical records are more frequently requested for abdominal pain and anemia patients with a frequency of 12 and 10% respectively. Furthermore, when admitting an abdominal pain patient there is an 83% chance of requesting outside medical records. Similarly for anemia patients, there is an 80% chance of requesting outside medical records. The internal medicine department usually serves people carrying several chronic conditions as comorbidities for an acute condition. Therefore, we note that most requests for outside medical records were for chronically ill patients. However, the data shows that acute cases such as lower urinary tract infections usually trigger requests for outside medical records as well. For this particular patient cohort, there is an 86% likelihood of requesting outside medical records when admitted to the internal medicine unit. No other acute conditions were found among the 20 most frequent association rules.

4.4. HIE's Apriori matrix

Following the idea of understanding which is the real contribution of each one of the twenty rules generated, we created a matrix that shows the robustness of each association rules. This need was generated by the evidence that the simple list of rules is not fully exhaustive by itself and should be investigated in order to discover the real predictive value of each rule, establishing different kinds priority. In order to achieve this goal, we introduce a new classification of the discovered rules based on support and confidence levels. This matrix allow us to select most significant rules, erase the less interesting and understand which are the critical relations.

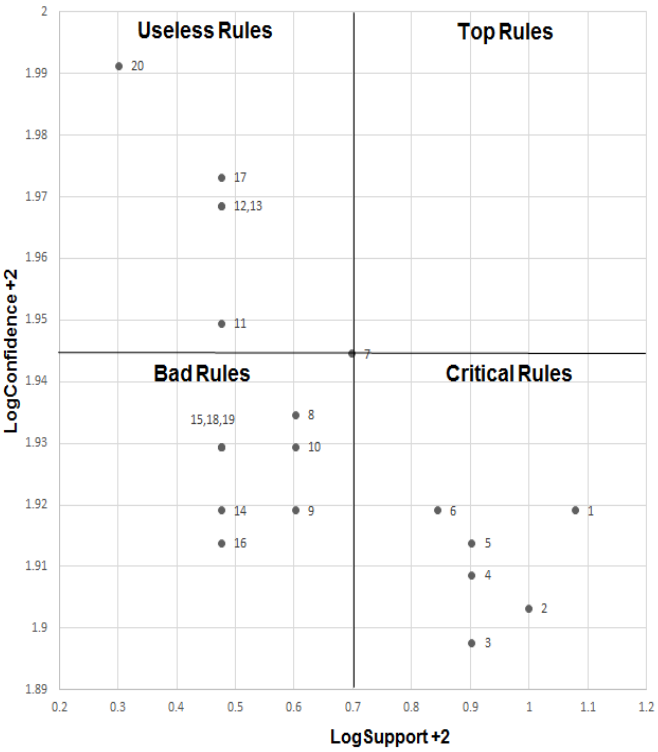
Generation and results.

First, we transformed the 20 rules support and confidence using a logarithmic transformation in order to obtain values easier to plot. Then, we fix LogConfidence+2 as vertical axis and LogSupport+2 as horizontal axis. In this way all the 20 rules should be classified respectively in 4 categories:

- ***Useless Rules*** *(low support, high confidence):* belong at this class rules 11,12,13,17, 20.
- ***Top Rules*** *(high support, high confidence) .*
- ***Bad Rules*** *(low support, low confidence):* belong at this class rules 8,9,10,14,15,16,18,19.
- ***Critical Rules*** *(high support, low confidence):* belong at this class rules *1,2,3,4,5,6.*

Analyzing the HIE matrix of Figure 4.5 we are furthermore able to understand the hidden nature of the twenty rules. The ideal result of the Apriori method is represented by the category Top Rules with a high level of support and confidence. Unfortunately in our case this category is empty because none of the selected association rules has sufficient values to be included here. The opposite type is Bad Rules. We found in this group all those categories with the lowest prediction power and the lowest occurrences level so these rules are can be considered with a lower priority respect to the others. In Useless Rules are contained those relations with a strong correlation between consequent and antecedent, so with an optimal predictive power, but that appear in few cases. In the end we have critical Rules that in our case represent the most interesting group. In fact we consider critical all those relations with a high percentage of occurrences but with a low confidence. About these, a choice will be necessary. In fact the high probability suggests a high priority to access, but the limited confidence constrains us to take a decision about the priority level. To effectuate this, a consultation between the IT developers and the medical users will be necessary.

Figure 4.5) Apriori HIE Matrix: in this figure is reported the classification of the 20 rules selected in the four categories Useless, Top, Bad and Critical Rules divided by confidence and support level.



	Antecedent	Consequent	LogSupport+2	LogConfidence+2
1	Abdominal pain	OMR	1.079181	1.919078
2	Anemia	OMR	1	1.90309
3	Dyspnea	OMR	0.90309	1.897627
4	Hypertension	OMR	0.90309	1.908485
5	Diabetes mellitus	OMR	0.90309	1.913814
6	Renal Failure	OMR	0.845098	1.919078
7	Cancer	OMR	0.69897	1.944483
8	Lower urinary tract infection	OMR	0.60206	1.934498
9	Hypotension	OMR	0.60206	1.919078
10	Back pain	OMR	0.60206	1.929419
11	Pneumonia	OMR	0.477121	1.94939
12	Chest pain, Outside Imaging	OMR	0.477121	1.968483
13	Anemia, Outside Labs	OMR	0.477121	1.968483
14	Abdominal pain, Nausea	OMR	0.477121	1.919078
15	Abdominal pain, Vomiting	OMR	0.477121	1.929419
16	CHF	OMR	0.477121	1.913814
17	Anemia, Outside Imaging	OMR	0.477121	1.973128
18	Hypertension, Diabetes mellitus	OMR	0.477121	1.929419
19	Abdominal pain, Vomiting, Nausea	OMR	0.477121	1.929419
20	Chest pain, Outside EKG	OMR	0.30103	1.991226

4.5. Data limits

Our first concern about results is the data quality: we realized that the data we obtained do not have a uniform distribution of frequency. This fact means that we have few very frequent items, as for example Outside Medical Record, which hide the other less frequent items in the selection phase of principal rules, as shown in the Figure 4.4.

Another important aspect is the lack of the International Classification of Diseases code (ICD) in the dataset. The ICD allows to standardize the different diagnosis, so when someone records an admission he would be able to match the patient diagnosis with a preset code. In our case we have not available this information and at the same time the dataset presents a high number of unclassified admissions. We confronted this limitation by scrubbing and verifying the accuracy of the entire dataset.

5. DISCUSSION

5.1. Overview of the results

The United States health care system us driven by competition, where most hospitals are profit oriented companies to generate revenue or ensure financial sustainability. Information about patients (i.e., customers) is a valuable asset to provide good quality service, as well as to maintain market power. In other words, there might be no incentives to share patient information, despite the possible lack of service for patients. The result of this situation it is a healthcare system not integrated with lack of communication and coordination, and consequently full of potential problems. In our study, we focused in the particular issue of informing HIE design. In the sense of discovering physicians information needs in a general hospital in the Tampa Bay region in Florida. More specifically, we shed light on new solution for problems as data quality of the information exchanged, the timing to access data by hospital personnel, the lack of understanding of the physician needs and the missing integration of different IT system in different healthcare structures of the same Country.

Starting from these problems, in this research we used data mining tools to discover useful and significant information to restructure and redesign a new possible HIE systems, providing new guidelines for the IT developers. The specific aim that we followed consists in providing a new efficient tool that simplifies the access of physicians to the most requested information. To achieve this goal, we examined the actual state of HIE technologies analyzing physician behaviors during the diagnosis phase, in particular searching hidden relation between the diverse types of information requested by them. In that way, we described how HIE performs into the clinical workflow in the hospital, identifying the general HIE-related workflow patterns and an exploration of exchange use across clinical contexts. Our work addresses a significant gap in the knowledge about patient-specific information needs during the admission and treatment of patients in the inpatient setting. We recognized the importance of evaluating and understanding the information needs of the different subjects involved in the HIE system. The rules generated in our case study are valid for the context previously described, but our approach is replicable in other hospitals with different characteristics.

HIE implementers can in fact adapt the HIE technology to meet the information and workflow needs corresponding to their clinical context. Also, users may need assistance in integrating HIE into clinical workflow and in understanding how HIE can directly benefit healthcare delivery [Unertl, 2012]. The delivery and the quality of discharge summaries can be improved substantially through health information technology [van Walraven, 1999]. Apriori results show a clear example of how technology offers the potential to quickly extract information about diagnoses, medications, and test results into a structured discharge document that can be reviewed for accuracy by the hospital physician and enriched with specific instructions about pending test results of outpatients. The current design of the information exchange is a ‘one-size fits-all’ model; and it is expected that allowing for user and role-specific customization may increase adoption and use [Unertl, 2012]. Understanding why and how clinicians use HIE data can assist with designing in function of user needs. Despite the good possible outcomes, it is necessary to consider all the single details that could influence the workflow processes in order to avoid erroneous information which sometimes enters into discharge communications, and it is rarely questioned once documented as part of the medical record [Adhiyaman, 2000]. An electronic medical record integrated system can ensure integrity and speed in the data capture process.

5.2. Results Analysis

First of all we need to consider that the algorithm has worked as we expected in fact the probability levels are similar to work previously done. As shown in the Results section, we obtained 2 output useful to be synthesized in smart guidelines: 20 association rules Diagnosis-Outside Medical Record, those represent the most relevant associations between the most requested Outside Information

(OMR) and the most frequent diagnosis present in our dataset , and the HIE Apriori Matrix, where we introduce a new classification of the discovered rules based on support and confidence levels.

5.2.1. 20 association rules: “Diagnosis-Outside Medical Record” Analysis

These rules are the main contribution of our research. These rules guide the design of new health IT systems. The first immediate observation is that “outside medical record” is the most requested information. This document was present in 1635 medical requests (78%) and it was found in the first 20 rules selected by the algorithm. This finding is interesting because we did not imagine before of this analysis a similar level of occurrence for a single outside information type. In our study, this strong occurrence has represented a crucial decision point: once obtained this result, we tried to work without this type of information but the support and confidence level necessary for Apriori analysis without Outside Medical Record were too low to be acceptable. This limitation is caused by the Apriori algorithm way to work that select only those items with the highest level of frequency. With this kind of limitation we formulated two hypotheses that could explain this situation despite the lack of information available. We only know, in fact, that the voice “outside medical record” collects different subtypes of document but we do not know what kind of information are inside. The first hypothesis suggests that the high level of request is the direct consequence of several subtypes of information classified under OMR but this explanation, even if could be possible, looks improper: is not clear which could be the interest for a hospital of having a similar classification. The second hypothesis, more realistic in our view, is related with the nature of the information: we think that in this category fall those medical information considered necessary and essential for the diagnosis of patient coming from outside hospital. Probably the real reason could be a mix of these two visions, anyway, starting from all these aspects, we chose to study the output with OMR as consequent, leaving the investigation about its characteristics for future research. Of course this outcome suggests that it is necessary to take a decision about the implementation of the new IT systems: actually the access at OMR should be the first priority during the design phase of the new HIE system, but for the future it will be necessary a redefinition of this information category, in order to provide data more with specifics on time without excessive clicking problem. Actually, many times physicians request Outside Medical Record because it is a generic document in which are present the most meaningful and different patient information. This type of outcome could change significantly if the health care system will adopt new more specific and standardized documents. In our sample we have other different outside documents but they have a low frequency of usage; with a new classification of information the results could be really different, with rules acceptable according to the diverse kinds of outside information.

Analyzing the set of relations, we can see that the first 11 rules are composed by only one diagnosis on the antecedent and this is a correct operation of the Apriori algorithm. Indeed, its processing way is to analyze simple itemsets before examining composed itemsets, so it generated rules with better support for that first case of itemset. The evident consequence of these results is the suggestion, for the developers, to prioritize the links between these diseases and the OMR in the new IT systems. There are various rules that have surprised us because of their nature. One of these is the number 12, composed by “chest pain” and “outside imaging” on the antecedent. Chest Pain is the most recurrent item among diagnosis with 387 cases (19%), but it was not found until the 12th position since the lift level is lower than 1 (it is assumed that a lift value >1 is an acceptable independence level). Consequently, about Chest Pain, they are present only rules composed by its sub-itemsets, because increasing the number of items in an itemset the level of independence increases. Another surprising aspect is the absence of the second most requested outside information document on the consequent, the “Outside Labs”. It occurred in 389 (19%) cases of the total number of admissions (2089). It appears rare to have no rule with this item on one side and on the other side one of the frequent diagnosis, like chest pain, other diagnosis, abdominal pain or anemia. After investigations, we discovered that the first excluded rule, containing one of the most frequent diagnosis, is composed by anemia and outside labs, and it has support of 3.5%, confidence of 28%, and lift equal to 1.5, so we can state the confidence level is unsatisfying. As for the Chest Pain case, we did not expect that the lift value could be a discriminating factor for inclusion or exclusion in a rule. This rules analysis well shows the pros and cons of Apriori tool: despite a lack of generality in the output, the algorithm provides strong relations, considering the independence as a necessary condition between the different items. Could be interesting in future research work utilizing other more complex algorithms in order to improve further the knowledge between disease and outside information.

The results show that the diagnosis diseases order of frequency analyzed, before implementing the algorithm, is respected by the selected rules.

5.2.2. HIE's Apriori Matrix Analysis

The second relevant result that we found is the HIE Apriori Matrix. Through this classification, we were able to understand which is the real potential value of each rule. In this way we could select the most significant rules, erase the less interesting and understand which are the critical relations. We chose to introduce this kind of scheme to provide a simple graphic instrument that easily explains why rules should have different level of priority to access.

Analyzing the HIE matrix of Figure 4.5, we are so furthermore able to understand the hide nature of all twenty rules. The first relevant aspect to consider is the absence of rules in the Top Rules category. This fact that in a first approach could appear as bad result, was an expected result. Indeed those that could be considered Top Rules were pruned by Apriori during the selection cause the lift level was lower than 1, so we can say that Top Rules do not exist because the items of these rules are afflicted by a form of dependence and Apriori necessarily has to erase these associations. For Bad Rules the situation is very different; belong at this class the major part of our rules but for different reasons. Rules 8, 9, 10, 16 are composed by only one item as antecedent but this is not so frequent, in particular the occurrences of “Low urinary tract infection” and “Hypotension”, those compose respectively rules 8, 9 with the best support of this category, are lower than 5 % so result natural obtaining rules with low support and confidence. For association rules 14,15,18,19 the condition is a little bit different but with the same output. In fact these are formed by frequent objects as “Abdominal Pain”, “Hypertension” and “Diabetes Mellitus” but the itemsets count three or four items inside so the final support and confidence result not sufficient for the rules to be included in other classes. A positive aspect of Bad Rules is the distribution of the all relations: they are bad but close to Useless and Critical Rules so these should have a lower level of priority to access respect the others classes but of course they should be considered in the new HIE hierarchy. The lasts two categories, Useless and Critical Rules, present a sort of duality between themselves and could be considered the most interesting. Useless Rules has a really strong predictive power with the highest levels of confidence in our case, that report the robustness of the relation between objects of the same itemsets, but the few occurrences of these make the rules not so useful: we can say that these associations should be considered rare certainties in our case. Furthermore, it is interesting that all the rules except 11 contain Outside Information in the antecedent part, which suggest us to provide an ulterior link between OMR and the other kinds of Outside Information present in these rules. Critical Rules include those associations with high support level despite a low confidence. We found here all the firsts rules selected, with the diseases present in the top part of the Disease List Table, and this element was expected: the most frequent diseases usually do not have a strong predictive power, or because the lift value is under than 1 or because they result too general to make a certain prediction. This category represent a crucial point respect the priority level of access to data: these symptoms occur often but is not sure that the physician need OMR. As conclusion HIE Apriori Matrix analysis we can say that unfortunately are not present Top Rules, but the aim of this classification was achieved with success in fact now we are able to establish a sort of hierarchy to data access based on the tradeoff between support and confidence. The entire hierarchical scheme will be decided by physicians and developers together, but we can state that Top Rules should have the priority in an ideal scheme, followed by Critical or Useless Rules (will depend by physician opinion), and as last Bad Rules.

5.3. Providing guidelines for possible involved figures

The results of this work want to provide insights for three HIE stakeholders: implementers, software developers and researchers. These potential readers are interested in different aspects about our study, so following we will explain for each one of these the most relevant characteristics where they are involved.

With implementers we refer to all the users included in the administrative hospital process which are interested about the new HIE technologies. This because they are the possible customers and direct users about new IT systems which contain strong relations, like Apriori rules, and so will be essential, for them, understanding the way of work of these new technologies. Developers are all people who work to improve and design the new HIE informatics software.

Our results might represent for them fundamental guidelines to consider for creating new efficient system being able to save time for users accessing to data.

The last category of stakeholders is researchers and all the scientists and academics which study HIE theme belong to this one. They could be interested at this research because represent a new application of data mining algorithm in this new context.

Starting from the obtained rules, it will possible realizing more efficient and time saving network. The Apriori technique offers advantages in terms of speed and time, so people are able in this way to analyze a large amount of data in a fair period of time. Another relevant contribute is the decision support in diagnosing process. This is an old research area, but up till now it has been constrained by the dependency on human medical experts for the population of knowledge in the system. Now data mining features embedded in such systems, will add up another dimension view in diagnostic decision making process.

In order to create a simpler IT system for hospital personnel, our study shows that it will be efficient to design a software that includes X diagnosis categories (in our case we considered 30 of these) directly linked with the Y most requested Outside Information (in this study were 9) because in this way it is possible classifying and having a direct access at the useful data on time and without an “excessive clicking”. This choice evidences a trade-off between the accuracy degree and the amount of information submitted to the physicians. Our new study shows how it is possible create a new system simpler, user friendly and complete. This new idea of software in fact does not exclude the less frequent voices by the entire system but put the user focus only on the most requests categories. A key factor to consider during the data cleaning process is the quality of the data and the measures needed to solve possible problems. In particular our work show how can be possible create a new classification for those data not well standardize and suggest an easy and effective way to order in a Pareto approach the most part of information.

Only a part of the total results in this field could be really useful information, so the future researchers shall be able to select only those rules that they make sense for healthcare application that is creating a function that highlights all the rules which connect a diagnosis with an outside document.

Critical and related at the final outcome of data mining application is the choice of threshold levels. In particular, it needs to consider two aspects to effectuate an optimal choice. The first is seeking similar studies to understand which could be the initial range of support and confidence level. Then, it is necessary look at the dataset characteristics in order to verify if those thresholds values are still acceptable or need to be changed. Indeed below certain percentages the rules generated could not be assumed as general assumption.

As conclusion of this section, we can say that this document has not the presumption to offer a definitive and complete solution at HIE problems, but provide smart guidelines to improve the actual network and develop future systems, showed by the following Figure 5.1 Smart Guidelines table and explained in this last paragraph.

Figure 5.1) Smart Guidelines & Research Results Table

1	Data Mining as innovative approach to obtain significant healthcare information
2	Set of 20 strong association rules: these represent the most related diagnosis between diagnosis and OI in a teaching hospital; should be consider the first guideline to follow during the design of IT systems
3	Relevance of OMR: is the most requested information. In the new HIE informatics system personnel should be able to access rapidly at these data.
4	Human review phase: only through this is possible select meaningful results despite of the others not interesting
5	Significance of dataset characteristics: data distribution influences strongly the final result so it is better worked with heterogeneous data, with classes of similar dimensions and well standardized.
6	Data Driven approach to create a new Pareto classification of medical data
7	HIE Apriori Matrix as useful classification that explains clearly the real nature of each association rules

5.4. Data driven versus qualitative approach

A relevant aspect to consider is the data driven approach that we followed to develop our elaboration. Despite of qualitative way, the data driven approach allows us to show with quantified data our results and make us able to understand with a concrete vision how the system could be implemented. Classic qualitative methods used in other cases, for example questionnaires, sometimes are too much influenced by the personal experienced the interviewers and the data collected results not so valid or truthful. With our data driven approach we start from real data so we did not have this kind of problems about data reliability and results are quantified in a technical way.

This new kind of exploring process needs also a human review. In fact, in our work, Apriori generated numerous association rules but the most part of this was not interesting for our scope, so we effectuated another selection phase in order to obtain only useful rules at our goal. The data distribution structure has been an important factor in our project as well, because working with not heterogeneous data makes harder mining the different attributes and the most common items, as for example OMR, covered the less frequents so the resulting rules from these will not be considered. Thinking on this factor, for future research will be helpful trying to analyze dataset with a more heterogeneous level of distribution. Another critical aspect linked to this research was the classification of clinical data cleaning. Not always the data analyzed are standardizing in unique categories and our case was one of this. We received a data collection were was not include the ICD-9 code, the key diagnosis code classification, so we were constrained to invent a new method to order all the different patient diagnosis considering that we did not have medical competence and the time to re-request others data. We decided to count all the health problems listed for each admission in order to obtain the most frequent categories and then we selected those with the higher occurrence level. After this process, we review all those words not classified in the first 30th to check if there were synonymous or equivalent. At the end of this process we obtained a final classification that included over the 84% of cases considering only 30 pathologies. This could be considered a new heuristic way that allowed us to achieve a Pareto satisfying classification.

5.5. Limitations

This new kind of exploring process needs also a human review. In fact, in our work, Apriori generated numerous association rules but the most part of this was not interesting for our scope, so we effectuated another selection phase in order to obtain only useful rules at our goal. The data distribution structure has been an important factor in our project as well, because working with not heterogeneous data makes harder mining the different attributes and the most common items, as for example OMR, covered the less frequents so the resulting rules from these will not be considered. Thinking on this factor, for future research will be helpful trying to analyze dataset with a more heterogeneous level of distribution. Another critical aspect linked to this research was the classification of clinical data cleaning. Not always the data analyzed are standardizing in unique categories and our case was one of this. We received a data collection were was not include the ICD-9 code, the key diagnosis code classification, so we were constrained to invent a new method to order all the different patient diagnosis considering that we did not have medical competence and the time to re-request others data. We decided to count all the health problems listed for each admission in order to obtain the most frequent categories and then we selected those with the higher occurrence level. After this process, we review all those words not classified in the first 30th to check if there were synonymous or equivalent. At the end of this process we obtained a final classification that included over the 84% of cases considering only 30 pathologies. This could be considered a new heuristic way that allowed us to achieve a Pareto satisfying classification.

6. CONCLUSION

All these are the reasons why we tried to create a tool that may be helpful in the HIE process. Indeed only improving the system from the technical point of view, healthcare system could effectively begin to save time and cost in the future, generating a real well integrated system. We believe that this technical tool supplies a new kind of manner for supporting information exchange design, but the research needs to continue to be developed in order to be applied in the healthcare network. The Apriori Algorithm, with the association rules, is only a starting point to satisfy the clinician behaviors.

In summary, we have used a new approach, association rule mining, for understanding the type of patient-level clinical information requested by internists to outside healthcare providers via Apriori algorithm. Analysis of the internists' information needs by Apriori indicates that chronically ill patients usually require more outside information than patients with acute conditions. By focusing on the most frequent health conditions generating requests for outside medical records, it was found that patients with abdominal pain, anemia, dyspnea, hypertension, and diabetes are target of HIE utilization. On the other hand, patients with acute conditions such as cancer and lower urinary tract infection were found to be more prone to have medical records requested. By using data from other care settings and contexts, association rule mining can be applied to discover health professionals' information needs for specific types of patients. We expect our approach open up further research on information needs in the context of health IT and on the development of better HIE systems.

6.1. Future Research

An added value of this work is the application in a new context such HIE of a data mining algorithm. In this specific case, Apriori is used to discover and extract meaningful information from clinics data.

Starting from this kind of study, it might be interesting, in future research projects, acquiring new knowledge in several different contests related with healthcare and explore those details of this research still unknown, as for example the classification criteria and the data included in the different kinds of Outside Information and OMR.

In this sense, our suggestion is to find different algorithms to mine information across the healthcare system in a more effective way. Indeed, other data mining techniques or an adaptation of a modified Apriori algorithm might provide better outputs for HIE area. We state that basing on our experience of applying this simple version of Apriori.

REFERENCES

Abdullah U., Ahmad J., Ahmed A., “Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining”, ICET 2008. 4th International Conference on Emerging Technologies, Pakistan, pages 327-331.

Adams V., “Introduction to Data Analysis”, JSAP, Vol. 49 no. 8, 2008, pages 375-376.

Adhiyaman V., Oke A., White A., Shah I., “Diagnoses in discharge communications: how far are they reliable?” IJCP, Vol. 54 no. 7, 2000, pages 457-458.

Adler-Milstein J., DesRoches C., Jha A., “Health information exchange among U.S. hospitals”, AJMC, Vol.17 no. 11, 2011, pages 761-768.

Adler-Milstein J., Landefeld J., Jha A., “Characteristics associated with regional health information organization viability”, JAMIA, Vol. 17, 2010, pages 61-65.

Adler-Milstein J., Sarma N., Woskie L., Jha A., Health Affairs, Vol. 33 no. 9, 2014, pages 1559-1566.

Antonie M. et al., “Application of Data Mining Techniques for Medical Image Classification”, Proceedings of the second international workshop on multimedia Data, San Francisco, USA, 2001.

Bailey J., Wan J., Mabry L., Landy S., Pope R., Waters T., Frisse M., “Does Health Information Exchange Reduce Unnecessary Neuroimaging and Improve Quality of Headache Care in the Emergency Department?”, Journal of general internal medicine, May 2012.

Barati E. et al., “A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction”, Journal of Selected Areas in Health Informatics (JSHI), Vol. 2 no. 3, 2011, pages 1-11.

Bellaachia A, Guven E., “Predicting Breast Cancer Survivability Using Data Mining Techniques”,Department of Computer Science The George Washington University, Vol. 58 no. 13, 2006.

Bernstein A., Health Care in America: Trends in Utilization, USA, National Center for Health Statistics (U.S.), Centers for Disease Control and Prevention (U.S.), Dept. of Health and Human Services, 2004.

Bourgeois F., Valim C., Wei J., McAdam A., Mandl K., “Influenza and other respiratory virus-related emergency department visits among young children”, Pediatrics, Vol. 118 no. 1, pages e1–8, 2006.

Callen J., Paoloni R., Li J., Stewart M., Gibson K., Georgiou A., Braithwaite J., Westbrook J., “Perceptions of the effect of information and communication technology on the quality of care delivered in emergency departments: a cross-site qualitative study”, Annals of emergency medicine, vol. 61 no. 2, 2013, pages 131–44.

Campion T., Edwards A., Johnson S., Kaushal R., “Health information exchange system usage patterns in three communities: Practice sites, users, patients, and data”, IJMI, Vol. 82 no. 9, 2013, pages 810-820.

Chen Y., Avery A., Neil K., Johnson C., Dewey M., Stockley I., “Incidence and possible causes of prescribing potentially hazardous/ contraindicated drug combinations in general practice”, Drug Safety, 2005, Vol. 28 no. 1, 2005, pages 67-80.

Choksi V., Marn C., Bell Y., Carlos R., “Efficiency of a semiautomated coding and review process for notification of critical findings in diagnostic imaging”, AJR, Vol. 186 no. 4, 2006, pages 933-936.

Committee J., British National Formulary 49th Edition, British Medical Association and Royal Pharmaceutical Society of Great Britian: London, 2005.

Cullen, Ambulatory Care Data from Health, United States, National Center for Health Statistics, 2005.

Deepika N. et al., “Association rule for classification of Heart-attack patients”, International Journal of Advanced Engineering Sciences and Technologies, Vol. 11 No. 2, 2011, pages 253-257.

Foundation M., “The Personal Health Working Group. Final Report”, The Markle Foundation, 2003, pages 1-58.

Furukawa M., Patel V., Charles D., Swain M., Mostashari F., “Hospital electronic health information exchange grew substantially in 2008-12”, Health Affairs, Vol. 32 no. 8, 2013, pages 1346-1354

Hadley J., Agola J., Wong P., “Potential impact of the American College of Radiology appropriateness criteria on CT for trauma” AJR, Vol. 186 no. 4, 2006, pages 937-942.

Hincapie A., Warholak T., Murcko A., Slack M., Malone D., “Physicians' opinions of a health information exchange”, Journal of the American Medical Informatics Association, 2011, pages 60-65.

Ilayaraja M., Meyyappan T., “Mining Medical Data to Identify Frequent Diseases Using Apriori Algorithm”, Pattern Recognition, Informatics and Mobile Engineering (PRIME), International Conference on, 2013, pages, 194-199.

Jasperson J., Carter P., Zmud R., “A comprehensive conceptualization of post-adoptive behaviors associated with information technology enabled work systems”, Journal MIS Quarterly, Vol.29 no. 3, 2005, pages 525-557.

Johnson K., Unertl K., Chen Q. et al., “Health information exchange usage in emergency departments and clinics: the who, what, and why”, JAMIA, Vol. 18, 2011, pages 690-697.

Kaelber D., Bates D., “Health information exchange and patient safety”, Journal of biomedical informatics, Vol. 40 no. 6, 2007, pages S40-S45.

Kern L., Barron Y., Abramson E., Patel V., Kaushal R., “HEAL NY: Promoting interoperable health information technology in New York State”, Health Affairs, Vol. 28 no.2, 2009, pages 493-504.

Kleinke J., “Dot-gov: market failure and the creation of a national health information technology system”, Health Affairs, Vol. 24 no. 5, 2005, pages1246-1262.

Kouroubali A., Starren J., “Costs and Benefits of Connecting Community Physicians”, AMIA, 1998, pages 205–209.

Kralewski J., Zink T., Boyle R., “Factors Influencing Electronic Clinical Information Exchange in Small Medical Group Practices”, Journal of Rural Health, Vol. 28 no. 1, 2012, pages 28-33.

Kripalani S., LeFevre F., Phillips C., Williams M., Basaviah P., Baker D., “Deficits in Communication and Information Transfer Between Hospital-Based and Primary Care Physicians, Implications for Patient Safety and Continuity of Care”, JAMA, Vol. 297 no.8, 2007, pages 831-841.

Lapointe L., Rivard S., “Getting physicians to accept new information technology: insights from case studies”, CMAJ, Vol. 174, 2006, pages 1573-1578.

Lazarou J., Pomeranz B., Corey P., “Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies”, JAMA, Vol. 279 no.15, 1998, pages 1200-1205.

Leape L., Bates D., Cullen D., Cooper J., Demonaco H., Gallivan T. et al., “Systems analysis of adverse drug events”, JAMA, Vol. 274 no.1, 1995, pages 35-43.

Levy G., Blachar A., Goldstein L., Paz I., Olsha S., Atar E. et al., “Non radiologist utilization of American College of Radiology Appropriateness Criteria in a preauthorization center for MRI requests: applicability and effects”, AJR, Vol. 187 no. 4, 2006, pages 855-858.

Mann K., Hiemke C., Lotz J., Schmidt L., Lackner K., Bates D., “Appropriateness of plasma level determinations for lithium and valproate in routine care of psychiatric inpatients with affective disorders”, Journal of Clinical Psychopharmacology; Vol. 26 no. 6, 2006, pages 671-673

Miller A., Tucker C., “Health information exchange, system size and information silos”, Journal of Health Economics, Vol. 33, 2014, pages 28-42.

Miller R., Miller B., “The Santa Barbara County Care Data Exchange: what happened?” Health Affairs, Vol. 26 no. 5, 2007, pages w568–w580.

Nicholson C., Jackson C., Tweeddale M.,Holliday D., “Electronic patient records: achieving best practice in information transfer between hospital and community providers – an integration success story”, Quality in Primary Care, 2003, pages : 233-240.

Ordonez C., "Improving Heart Disease Prediction Using Constrained Association Rules”, Seminar Presentation at University of Tokyo, 2004.

Ozkaynak M., Brennan P., “Revisiting sociotechnical systems in a case of unreported use of health information exchange system in three hospital emergency departments”, JECP, Vol. 19 no. 2, 2013, pages 370-373.

Pujari A., Data Mining Techniques, India, Universities Press, 2001.

Ross S., Schilling L., Fernald D. , Davidson A., West D., “Health information exchange in small-to-medium sized family medicine practices: Motivators, barriers, and potential facilitators of adoption”, international journal of medical informatics, 2010, pages 123–129.

Rudin R., Motala A., Goldzweig C., Shekelle P., “Usage and Effect of Health Information Exchange A Systematic Review”, AIM, Vol. 161 no. 11, 2014, pages 803-811.

Rudin R., Volk L., Simon S., Bates D., “What Affects Clinicians’ Usage of Health Information Exchange?”, ACI, Vol. 2 no. 3, 2011, pages 250-262.

Rudin R., Volk L., Simon S., Bates D., “What Affects Clinicians’ Usage of Health Information Exchange?”, Appl clin Inform, 2011.

Shapiro J., Kannry J., Kushniruk A., Kuperman G., “Emergency Physicians’ Perceptions of Health Information Exchange”, Journal of the American Medical Informatics Association, Vol. 14 no. 6, 2007, pages 700-705.

Shapiro J., Kannry J., Lipton M. et al., “Approaches to patient health information exchange and their impact on emergency medicine”, Annals of Emergency Medicine, Vol. 48 no. 4, 2006, pages 426-432.

Shapiro J., Kuperman G., “Health Information Exchange”, Ong K., medical informatics, edited by MD,MPH, 2007.

Sharma N., Om H., “Extracting Significant Patterns for Oral Cancer Detection Using Apriori Algorithm”, Intelligent Information Management, Vol. 6 no. 2, 2014.

Sicotte C., Pare G., “Success in health information exchange projects: Solving the implementation puzzle”, Social Science Medicine, Vol. 70 no. 8, 2010, pages 1159-1165.

Soni J. et al., “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications, Vol. 17 no. 8, 2011.

Srikant R., Vu Q., Agrawal R., “Mining Association Rules with Item Constraints” KDD, 1997.

Subbalakshmi G. et al., “Decision Support in Heart Disease Prediction System using Naive Bayes”, Indian Journal of Computer Science and Engineering, 2011.

Sutcliffe K., Lewton E., Rosenthal M., “Communication failures: an insidious contributor to medical mishaps”, Academic Medicine, Vol. 79 no. 2, 2004, pages 186-194.

Thorn S., Carter M., Bailey J., “Emergency Physicians’ Perspectives on Their Use of Health Information Exchange”, Annals of Emergency Medicine, Vol. 63 no. 3, 2014, pages 329-337.

Unertl K., Johnson K., Lorenzi N., “Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use”, JAMIA, Vol. 19 no. 3, 2012, pages 392-400.

Unertl K., Johnson K., Lorenzi N., “Health information exchange technology on the front lines of healthcare: workflow factors and patterns of use”, JAMIA, Vol. 19, 2012, pages 392-400.

van Walraven C., Laupacis A., Seth R., Wells G., “Dictated versus database-generated discharge summaries: a randomized clinical trial”, CMAJ, Vol. 160 no. 3, 1999, pages 319-326.

Vest J., “More than just a question of technology: factors related to hospitals’ adoption and implementation of health information exchange”, IJMI, Vol. 79 no. 12, 2010, pages 797-806.

Vest J., Jasperson J., “How are health professionals using health information exchange systems? Measuring usage for evaluation and system improvement”, JMS, Vol. 36 no.5, 2012, pages 3195-3204.

Vest J., Jasperson J., “How are Health Professionals Using Health Information Exchange Systems? Measuring Usage for Evaluation and System Improvement”, JMS, Vol. 36 no. 5, 2012, pages 3195-3204.

Vest J., Zhao H., Jaspserson J., Gamm L., Ohsfeldt R., “Factors motivating and affecting health information exchange usage” , JAMIA, 2011.

Wilcox A., Kuperman G., Dorr D. et al., “Architectural strategies and issues with health information exchange”, AMIA, 2006, pages 814-818.

Ye Y., Chiang C., “A Parallel Apriori Algorithm for Frequent Itemsets Mining”, Software Engineering Research, Management and Applications, Fourth International Conference on, 2006, pages 87-94.

Zwaanswijk M., Verheij R., Wiesman F., Friele R., “Benefits and problems of electronic information exchange as perceived by health care professionals: an interview study”, BMC Health Service Research, Vol. 11, 2011, page 256.

APPENDIX

Appendix A – Apriori Code

In this appendix, we explain how implement the Apriori Algorithm using R software.

The first step is to import the text data into the program in a format readable by R.

To accomplish this,we use the command "read.delim" to read the data:

```
tableresults <- read.delim("C: /tableresults.txt")
```

Then we create an empty matrix of the same size of our table:

```
tableresults2=matrix(rep(0,2089*39),nrow=2089,ncol=39)
```

After of this passage, we insert a for cycle to convert the original data format into a numerical format with the command "as.numeric" and fill the empty table with the value results:

```
for(i in 1:ncol(tableresults))

{

  tableresults[,1]=as.numeric(tableresults[,1])

  for(j in 1:nrow(tableresults))
  {if(tableresults[j,i]==1){tableresults2[j,i]=1}}
}
```

So, we fix the new columns names with the original names:

```
colnames(tableresults2)=colnames(tableresults)
```

and then we coerce the new matrix with the "as.(,"transaction")" command into a new variable:

```
tableresults3=as(tableresults2,"transactions")
```

In this way now R could be able to implement Apriori Algorithm.

We use then the "apriori()" function to generate all the itemsets and the rules among these, with values better than the minisup and minconf, and with subset(lift>1) we select only the rules with a lift value greater than 1.

```
rules=apriori(tableresults3,parameter=list(support=0.02,confidence=0.79))
ruleslift <- subset(rules, subset = lift > 1)
```

This measure was necessary to obtain rules with a real new contribution.

In the end, we use "inspect()" to print the first association order by support and "write()" to export these last in another file format.

```
inspect(head(sort(ruleslift, by="support"),100))

write(ruleslift,file="results table 1.txt",sep=" ",col.names=NA)
```

Appendix B – Attributes check and conversion

Here are reported all the original data classification and the transformation of these into new synthetic voice. The parenthesis report the old categories included in the new ones.

AGE: Patient age was calculated by calculating the difference between the date of birth and admission date.

MARITAL STATUS: Single (“Single”,” Divorced”,” Widowed”,” Separated”), Married and Others (“Significant Other”, “Unknown”)

PRIMARY CARE PHYSICIAN: Yes, No

PAYER CLASS: Medicare (“Medicare”,” Medicare Advantage”), Medicaid (“Medicaid”,” Medicaid Managed Care”,” Medicaid Pending”), and Others (“Other”, ”Self-Pay”,” TPL/Auto”, ”Tricare”, “Worker’s Comp”, “Commercial”, “HCHCP”, “HCHCP Pending”)

DISCHARGE DISPOSITION: Transfer to another Healthcare Facility (”Skilled Nursing Facility”, “Home-Health Care Svc”, “Psychiatric Hospital”, “Inpatient Rehab Facility”, “Intermediate Care Facility Hospice/Medical Facility”, “Rehab Facility”, “Long Term Care”, “Another Health Care Institution Not Defined”, “Federal Hospital”, “Cancer Center or Children's Hospital,Short Term Hospital”, “Disch/Trans to a SNF with MCARE Certification with a Planned Readmission”, “Disch/Trans to an Inpt Rehab Facility/Unit with a Planned Readmission”), Transfer to Home (“Home or Self Care”, “Hospice/Home”), Expired (“Expired – No Autopsy-No Organ Donation”, “Expired”, “Expired-Surgical Death within 3-10 days post-surgery-Autopsy”, “Expired- Autopsy No Organ Donation”), and Other (“Left Against Medical Advice”, ”Court/Law Enforcement”).

ADMISSION SOURCE: Self-Referral, Physician or Clinical Referral, Outside Hospital (“Outside Health Care Facility”, “Outside Hospital”), Emergency Room, Other (“Court/Law Enforcement”, “Skilled Nursing Facility”)

PROVIDER TYPE: Resident, Physician, Nurse Practitioner, Physician Assistant, Physician Assistant/Physical Therapy, Anesthesiologist, NULL, Dentist.

30 DIAGNOSIS LIST: Abdominal pain, Anemia, Chest pain, Leukocytosis, Fever, Hypertension, Hypotension, Hypokalemia, Hyponatremia, Tachycardia, Altered mental status, Diabetes mellitus, Coronary artery disease, Dyspnea, Back pain, Lower urinary tract infection, COPD, Headache, Renal Failure, Vomiting, Nausea, Cancer ,Syncope, Pneumonia, CHF, GI Bleed, Cellulitis, Alcohol, Weakness and Other Diagnosis. With the first 29 health problems we are able to classify over the 84% of the total admissions, so we introduce the voice “Other Diagnosis” to consider the remaining 16%

that could be consider the sum of the other rare health disease, with less than 3% of cases for each one of the total hospitalizations.

OUTSIDE INFORMATION TYPE: Outside Medical Record, Outside Labs, Outside Imaging, Outside History and Physical, Outside Clinical Note, Outside Discharge Summary, Outside Consultation, Outside Surgery/Procedure, Outside EKG, Outside Radiology, Outside Medications and Outside Exercise Stress Test.

RINGRAZIAMENTI

Al Prof. Zayas-Castro, grazie al quale abbiamo avuto la possibilità di svolgere la nostra tesi nell'ambito della sanità americana.

Ai Dr. Martinez e Garcia, che ci hanno affiancato in questi mesi di lavoro insegnandoci le basi della ricerca universitaria.

Al Prof. Lettieri, per averci messo nelle condizioni e reso possibile svolgere il nostro lavoro di tesi al termine della carriera universitaria, dimostrando disponibilità e cortesia.

Alla Dott.ssa Segato, per averci seguito e supportato costantemente durante l'intero svolgimento del lavoro.

Alle nostre famiglie e ai nostri amici, per averci sostenuto e aiutato durante il nostro percorso formativo.