EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Towards early and automatic detection of Urinary Infection during pregnancy

Lizbeth Escobedo[1,*], Adán Hirales-Carbajal[1]

[1]CETYS Universidad, Calz Cetys 813, Lago Sur, Tijuana, B.C. Mexico

Abstract

INTRODUCTION: Worldwide Fetal-Maternal morbidity and mortality is frightfully high. Most of these diseases occur in developing countries. One of the main reasons for this problem, after gestational hypertension and complications in childbirth, is infections. Urinary Tract Infections (UTI) during pregnancy is one of the main causes for fetal-maternal morbidity and mortality in Mexico. Among others, the pervasiveness and heterogeneity of data in Electronic Medical Records (EMR) complicates early diagnosis and treatment of UTI.
OBJECTIVES: Our goal is extract empirical knowledge, in the form of association rules, that generalize symptomatology and treatment of UTI patients with positive and nagative diagnosis.
METHODS: In this study, we developed a criterion to extract words and expressions that uniquely characterize each patient class. We extracted association rules from EMRs and evaluated its level of correspondence between the rules and the extracted word sets.
RESULTS: By defining a bound on word frequency usage and evaluating the positive to negative word ratio we were able to identify word sets that uniquely characterize each patient class. A bound of 47 enabled extraction of 25 unique words and expressions for each patient class. Further, approximately 17% and 27% of association rules drew terms from each word set correspondingly.
CONCLUSION: This work seeks to promote the creation of more effective criterions to extract features, from EMRs, that improve characterization of patients and that ultimately lead to a more accurate diagnosis of UTIs.

Received on 15 July 2019; accepted on 10 August 2019; published on 23 August 2019

Keywords: Pregnancy health, pervasive health, digitization of healthcare, data analysis, fetal-maternal morbi-mortality, natural language processing, urinary tract infections

doi:10.4108/eai.13-7-2018.162810

1. Introduction

Approximately 830 women die each day around the world from preventable causes related to pregnancy and childbirth [1]. Most of these complications occur during pregnancy and most are preventable or treatable. Maternal and neonatal health are closely related. Around 2.7 million newborns died in 2015 [2], and another 2.6 million were born dead [3].

According to the general direction of epidemiology report, during the first week of January 2018 there were 15 deaths in Mexico, [4]. Public health institutions in Mexico offer a wide variety of services for this population to access to prenatal care. The National Epidemiological Surveillance System reported that in 2014 the Urinary Tract Infections (UTIs) occupied third place within the main causes of fetal-maternal morbidity and mortality [4, 5].

The Federal Government published the Official Mexican Norm NOM-007-SSA2-2016, which defines the care of women during pregnancy, delivery, puerperium, and the newborn, as well as criterion's and procedures for providing the service. Among others the norm defines prenatal control, and procedures for detection, prevention, and treatment of urinary tract infections.

The percentage of daily clinical visits, related to UTI, ranges between 10% to 20% [6]. The supervision of

*Corresponding author. Email: Lizbeth.escobedo@cetys.mx

pregnancy according to the norm, specifies a minimum of 5 clinical visits with the gynecologist during the pregnancy length. In 2015, a preliminary analysis at the Family Medicine Unit from the Mexican Social Security Institute (MSSI, e.g., the largest provider of medical services in Mexico), informed that 119,740 clinical visits were given to women; 20,584 were related to any stage of pregnancy of which 3,120 were UTI related. Of these, 13 resulted in hospitalization due to a complication of UTI, which resulted in an operational cost of $187,866.00 Mexican Pesos [7].

The main instrument for detection of fetal and maternal risk are preventive clinical visits. Clinical visits may be assigned to different medical staff. Often, physicians provide care base on empirical reasoning gain by treating many cases. Expertise that finds its ways to patients EMRs and aggregates different diagnosis and treatment styles.

During the clinical visit, the medical provider conducts auscultation, diagnosis, and treatment. All information generated from the clinical visit is added to the patient's medical record that could be electronic or not depending on the hospital's area (e.g., x-rays, laboratories, dentist, etc). This information is produced in the form of notes that the medical provider writes according to her/his experience, which can be an important risk factor for early detection and treatment. In addition, the quantity and quality of information generated by the hundreds of thousands of daily clinical visits make the task of discovering valuable and timely knowledge complex. In Mexico, the health sector is not exempt from this reality and has identified in IT (e.g., information technologies) as an ally to increase efficiency through the increasingly frequent use of electronic medical records (EMR) [8]. Despite the advantage implied by the use of IT, this medium is only used as an electronic capture and access of data, notes are still made based on the medical provider's experience. In addition, medical providers are full of work because of the amount of daily clinical visits, and the complicated work environment add together another negative factor. In a typical day of the Institute, gynecologist can have more than 29.5 clinical visits per 8-hours workday [9], with an average of 8.13 minutes per clinical visit. Even though according to the Mexican Social Security Institute a clinical visit should take no less than 20 minutes [10].

Digitalization of healthcare of pregnant women is key for early detection of problems, improving the diagnosis process, and to train healthcare provides. This work advances general understanding in two ways: in health information systems context by developing a criterion that extracts a set of unique words and expressions, from EMRs, that correspond to patients with a UTI positive or negative diagnosis; and by analysing the correlation between the extracted word set and association rules extracted for each patient class from the EMRs. Our intent is to evaluate the level of correspondence between the classification rules and the unique words sets. Thus the scope of this work is to evaluate this relation in an exploratory manner. In subsequent studies, we will either improve existing learning heuristics or propose new ones.

2. Related work

Quality assurance of healthcare has been analysed from different points of view. First efforts were in creating standards to activities and programs intended to assure or improve the quality of care in either a defined medical setting or a program. The World Health Organizations (WHO) is leading the standards worldwide through different institutions in each country.

2.1. Semantic Indexing

Modern healthcare critically depends on data analysis [11] and processing. Semantic technologies offer a possible solution, analysing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

In the US, the National Committee for Quality Assurance (NCQA) specifies a wide range of quality measures through the Healthcare Effectiveness Data and Information Set (HEDIS). For example, the HEDIS specifies, among other diseases, in the Diabetic Care that a diabetic patients in the age range of 18 to 75; must receive an eye exam during a relevant period of time. Piro et. al. [11], applied Semantic Technologies to compute part of the HEDIS measures, this is encoding it from natural language to a set of rules, then computing data from a whole Kaiser Permanente clinical notes looking for an eye exam request in diabetic patients.

Others have used Latent Semantic Analysis to identify patterns in electronic medical records [12, 13] when based in writing standards. But there are several challenges when there are no writing standards. Medical providers use to capture a clinical notes as fast as s/he can using different words to refer to the same term in different way patient to patient (e.g., different abbreviations, terms of compound words, etc.). As we have seen LSI has been used for data analysis for example to find words, identify patterns, encode standards from natural language to a set of rules, and many more. But there are still efforts to process data before and after data analysis; for example, 'before' when data depends on writing style and experience of doctors, data is way different from doctor to doctor, and 'after' to identify compound words of medical records.

2.2. Association rules

Association rules are used to create knowledge bases quickly and automatically in healthcare. They are useful, as they offer the possibility to conduct intelligent diagnosis and extract information [14–17].

Healthcare organization widely used association rules for discovering relationships between various diseases and drugs [18], discover disease associations [17], predict the patient's next drug [19], etc. Stilou et.al. [14] created a system that offer an efficient and effective tool in the management of diabetes through data mining in databases. Association rules can also reveal biologically relevant associations between different genes or between environmental effects and gene expression as Creighton et. al. [15] proved. Overall Doctor's produce daily large amount of data through prescriptions in clinical notes and treatment materials. In this way, this approach can contribute to take advantages of electronic medical records and identify the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals, etc.

Invaluable knowledge can be extracted from large amount of data stored in databases, but what happens when the data comes from unclassified and unorganized databases (e.g., a clinical note field text). In this paper we process and analyse data from clinical notes from different gynaecologists in order to identify and early detect and treat urinary tract infections during pregnancy as a way to decrease fetal-maternal morbi-mortality.

2.3. Word embedding

Word embeddings are used to calculate or predict semantic similarity between words, terms or sentences [20]. Word embedding had been used in several fields to create cohorts, identify similarities among groups, classification, etc. Garg, N. et. al. [21] used word embedding to analyse a century of gender and ethnic stereotype to quantify historical trends and social change. In the medical field, Zhu, Z. et. al. [22] measured patient similarities from electronic medical records. Found the correct similarities and classification could ensure a cohort study and treatment. In this study they found significant classification improvements compared to the traditional classification by humans. Glicksberg BS et. al. [23], automated cohort disease selection of 5 disease using word embedding and compared with a traditional phenotyping algorithms. They found significant improvements in classification but with a variation between diseases, thus requiring further investigation.

3. Methods

Two hundred and seven, randomly chosen, medical records corresponding to pregnant patients with and without UTI were use in our analysis. Out of which 77 cases correspond to UTI negative patients and 130 cases corresponded to UTI positive patients. Patients show different classes of symptoms and developed the disease at different levels of severity. Selected patient records included five or more records in their EMR, stating with the first doctor visit.

Knowledge regarding the month of gestation was unclear. All data correspond to consultations that occurred from December 2014 to December 2015 in a Mexican public clinic. Due to patient protection regulations the location and name of the public clinic is confidential.

The EMR are based on the Mexican norm NOM-168-SSA1-1998, which establish scientific, technological, and administrative criterions for preparation, integration, and storage of medical records. The norm defines five classes of records see [24]. We analyse general class records. Such are composed by 40 fields, with knowledge corresponding to: timestamp, clinical summary, diagnosis, patient demographic data (name, size, weight, glucose, blood pressure, state, etc.), and physician name, among other features. Fields corresponding to the patient clinical history, diagnosis, and state (UTI positive or negative) are analysed.

EMRs are maintain primarily by the family doctor and by one or more specialists (i.e. gynaecologist-obstetrician). The family doctor treats primary illnesses. Specialist monitor and provide prenatal care and may attend delivery. Medical practitioners conduct pre and post consultation care. During a regular appointment, the pregnant patient is subject to a visual inspection and auscultation. Questions are often made to gain rapport and understand the patient context. Demographic information, diagnosis, treatment, prescription drugs, physiologic data, and laboratory tests (optional) data are included in the EMR. Currently, the hospital EMR system does not correct orthographic error in online manner.

In our previous study [25], we focused greatly on automating spelling correction and on extracting rules from empirical data. In this work, we move away from automatic spell correction since context dependent word disambiguation is complex and it is not the primary objective of our work. We instead focus on answering the following fundamental questions: can patients, with and without the disease, be characterized by a unique set of words? How does such word set relates to the rule set of each patient class?

4. Data normalization

Despite efforts to standardize the structure of EMR and to normalize the usage of medical terms, the language complexity in EMR is high. Physicians often convey knowledge in mixed style, both in descriptive

and narrative forms. Further, they make extensive use of scientific terms, abbreviations, acronyms, and initialisms. In our dataset, we have seen extensive usage of acronyms followed by numeric literals, i.e. Gluc. 74.0 mg/dl, and adverbs of time, which are often used to express temporal events and the state of the patient or disease. Finding similarities between patients, even among those sharing the same class of disease, is hampered due to differences in phenotypic data, i.e. it is less likely to find similarities between patients if compared by their period date. It may be more insightful to compare them by the frequency by which they experience of the event. Such linguistic nuances make analysis of EMR complex. To reduce the linguistic complexity in EMR we apply normalization. Our hypothesis is that reducing complexity and ambiguity, in health records, leads to a better generalization of rules.

Text normalization was carried out as follows: a list of 30 acronyms and expressions used to describe gynecologic problems and procedures was built. We scanned EMRs in search of grammatical variations and built the set of regular expressions. Each expression class was associated with a unique tag. We define a tag by the unique character # followed by a unique label name, i.e. #fup. See Table 1 UTI regular expressions. We only show 14 of 30 expressions we built*. The list is by no means complete. However, we consider it sufficient since it allowed us to identify nearly 90% of the gynecologic terms found in the EMRs. Spelling errors of non-technical words were corrected. Spell correction of technical terms was complex. For each term identified as erroneous, we searched for sentences containing similar words in clinical obstetrics and gynecology collections. A term was corrected only if we observe high similarity between the term usage in the EMR and queried expressions; Expression with form acronym + numerical literal; acronym + adverb of time; and containing prescription, appointment date, and the patient age (e.g. 500 mg, 8 hrs, 30 days, 22 years, 1x10x10, etc.) data were automatically tagged. We further automatically tag data corresponding to patient given name and place of residence data; Surnames and workspace data were manually tag. Finally, stop-words and punctuation characters were filtered-out.

To address the question, can patients, with and without the disease, be characterized by a unique set of words? We analyse the log positive to negative word frequency ratio. Words that have similar usage in both classes of patients will result in values close to zero. Words that are more frequently used in positively diagnosed patients will produce a positive value, otherwise, a negative value. To address the question How does such word set relates to the rule set of each patient class? we analyse the relation between the extracted rule set (association rules) and the unique and shared word sets produced by the previous analysis.

Extracted rules may enhance our understanding on what factors physicians evaluate when they assess patients with and without the medical condition; what factors are present in patients that deny having the condition; what procedures are applied in practice; what factors occur in patients with potential miscarriage; among other knowledge. Association rules are extracted by applying Agrawal et. al. [26] heuristic. Rule extraction hyperparameters include support of 0.6 and a minimum rule length of 1. In this early stage of results, we expect discovered knowledge to be noisy and perhaps redundant. In future work we will evaluate more robust methods for rule extraction, such as those proposed by Ait-Mlouk et. al. [27] and Bayardo & Agrawal [28]. For interpretation purposes, the support constraint represents the number of EMRs that support the rule. A support of 1% indicates the fraction of EMRs contains the rule. Since strong associations rules may not provide valuable knowledge [29]. Maximizing or minimizing such value may only be evaluated by the problem domain expert.

Let P be the word set that characterize patients with positive diagnosis, N the word set that characterize patients with negative diagnosis, S the set of words that characterize both positive and negatively diagnosed patients, such that $S = P \cap Q$. The sets of unique words corresponding patients with negative and positive diagnosis are defined by $N' = N - S$ and $P' = P - S$ correspondingly. Recall, sets P and N contain words whose is at least a given bound. Such bound is estimated in the results section.

Let R_P and P_N the rule sets corresponding to patients with positive and negative diagnosis. Further, let r be an association rule, with w_r its word set. The relation between an extracted rule set $(R_P \vee P_N)$ and a unique word set $(N' \vee P')$ is analysed as follows. For patients with positive diagnosis, for each rule $r \in R_P$ the ratio between $|w_r|$ and $|w_r \cap P'|$ is analysed. Such ratio accounts for the fraction of words in the rule that are unique to the patient class. The procedure is repeated for negatively diagnosed patients.

To give a better idea about how the proposed technology should work in the context of a ubiquitous solution, we present a scenario of use with a gynecologist, Edgar, and 19-year-old 6-months-pregnant women, Karla, which is still a student. Karla is the last clinic-visit of the day for Edgar, who has been attending 29 women before Karla. Karla is a low-income woman who lives in a rural community far from medical services. This is the 2nd clinical visit of Karla since she realized is pregnant. Edgar is tired because the long day journey. Karla has been sat an hour ago in the waiting room full of other people waiting. Edgar finally calls Karla. When Karla is sited in front of Edgar, Edgar starts

Table 1. UTI regular expressions

Item	Spanish Acronym	REGEX
Uterine fundus height	#*AFU*	`r'afu\s*\d{2}\s*(cm)?'`
Dosage	#DO	`r'\d+\s*x\s*\d+\s*x\s*\d+'`
Age	#EDAD	`r'\d{2,3}\s+a(?:n|ñ)os'`
Date	#SFECHA	`r'\b\d{1,2}\s*(?:de)?\s*(?:e(?:nero|ne|n.|n)?|f(?:ebrero|ebr.|eb.|eb|b|e)?|m(?:arzo|zo.|ar|r|z|a)?|a(?:bril|br.|bl.|br|b)?|m(?:ayo|ay|y.|y)?|j(?:unio|un.|un|n)?|j(?:ulio|ul.|ul|l)?a(?:go|sto|gto.|gt.|go|g.|g)?|s(?:eptiembre|etbre.|ept.|et.|ep|et|p|t|e)?|o(?:ctubre|ctbre.|ct.|ct|c|t)?|n(?:oviembre|ovbre.|ov.|ov|v|o)?|d(?:iciembre|icbre.|dic.|ic|c|i)?)\s*(?:de|del)?\s*(?:19[5-9]\d|2\d{3}|\d|\d{2})(?=\D|$)'`
	#NFECHA	`r'\b(?:0[1-9]|[12][0-9]|3[01])(?:\D|$)(?:0[1-9]\|1[012])(?:\D\|$)(?:19[5-9]\d|2\d{3}|\d{2})'`
Date of last period	#FUP	`r'\b(?:fup|fum|fur)\s*:?\s*(?:#NFECHA|#SFECHA)\b'`
Probable date of birth	#FPP	`r'\bfpp\s*:?\s*(?:#NFECHA|#SFECHA)\b'`
Fetal heart rate	#FCF	`r'fcf\s*\d{3}\s*(?:x\s*min|lat\s*/\s*min)?'`
Glucose	#GLUC	`r'\b(?:Glc.|Gluc.)\s*\d{2,3}(?:.\d)?\s*(?:mg/dL)?\b'`
Hematocrit	#HCT	`r'h[c]?t(?:.|o|o.)?\s*\d{1,2}(?:.\d{1,2}%?)?'`
Hemoglobin	#HB	`r'hb[.]?\s*(?:de)?\s*\d{1,2}(?:\.\d{1,2})*'`
Body mass index	#IMC	`r'imc\s*\d{2}(?:.\d+)?'`
Starts sex life	#IVSA	`r'ivsa[. :]?\s*#EDAD'`
Plaquets	#PL	`r'\b(?:pl.|pl|plaq.|plaq|plt.|plt)\s*\d{3}(?:.\d{2,3})?\s*(?:mm3|mil)?\b'`
Weeks of gestation	#SDG	`r'\d+\s*x\s*\d+\s*x\s*\d+'`

interviewing Karla about her pregnancy. Edgar uses the system who has the system embedded to capture Karla's answers. Karla stilly articulate she is filling good and sometimes she feels a low-back pain and thinks it's because the weight she is gaining. While Edgar types 'pv pain' the system automatically completes the phrase to 'pelvic pain' and identifies Karla had the same keyword 'pelvic pain' in her first clinic visit. The system alerts Edgar of a positive UTI. The systems suggests he inquiries and analyse other symptoms and produces a more accurate prescription for Karla.

5. Results

To address the question, can patients, with and without the disease, be characterized by a unique set of words? We first study the effect of the word frequency bound on the log positive to negative ratio and select a bound that balances the amount of shared and unique words. We explain the rationale for this criterion shortly.

Fig. 1 is interpreted as follows. The lower bound of 20 on word frequency allows less common words be included in to the sets. Assume analysis of EMRs of positively diagnosed patients and the previous bound, words with a low frequency count, perhaps corresponding to rare medical conditions, are included in the unique positive word set. We refer to it as positive set for short. Note, set sizes are also bounded 50 words.

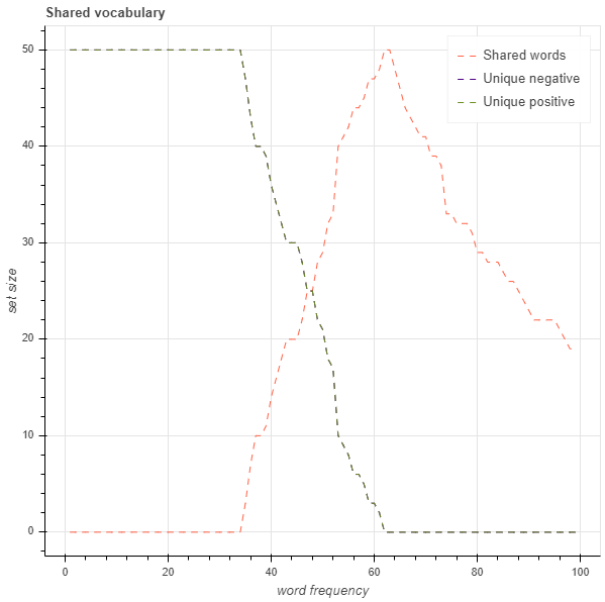

Figure 1. Shared word set vs unique word set size. Note unique positive and negative plots overlap.

Recall, words are in increasing or decreasing order of ratio. Both bounds limit the amount of words that can be used to diagnose and/or threat an illness. We define

Table 2. UTI word sets

Word class	Word set
Shared words	movimientos, resumen, rítmicos, datos, normales, pérdidas, globoso, negados, #periodo, campos, #sfecha, fetales, clínico, orientada, consciente, ego, pulmonares, #umcm, salida, embarazo, extremidades,tranquila, abdomen, exploración, física
Negatively diagnosed	niega, intensidad, piel, prenatal, control, femenina, expensas, vasoespasmo, cita, #edad, tono, #imc, #nombrepropio, normal, negadas, coloración, al, útero, diagnóstico, alergias, buen, tegumentos, bien, buena
Positively diagnosed	dolor, actividad, #do, integras, vías, uterina, #prescripcion, giordano, #localidad, diagnostico, sangrado, #fcf, usg, #afu, negativo, compromiso, urinarias, infección, #sdg, #umd, positivo, ivu, edema, #nfecha, #fup

the set size to 50 based on previous results. In [25] we found that the mean length of medical notes is 50 terms.

The positive and negative sets share words when the bound is within the range of 35-62. Half of words are shared at values of 47 and 48. Intuitively, shared words should correspond to those that originate during the auscultation phase. While the remaining ones, characterize the patient illness and treatment. We chose a bound of 47 to construct the positive, negative, and shared word sets.

Table 2 list the extracted sets. The share word set contain words that describe the patient state and patients responses the physician inquiries, i.e. the fetus rhythmic heart sounds and movements (movimientos ritmicos), lung fields (pulmonares), general urinary exam results (ego), calm (tranquila), conscious (conciente), and the patient's menstrual period dates (#periodo), etc.

The negative word set contain terms associated to the visual inspection of the skin, i.e. good coloration and teguments (buena coloración tegumentos). Physicians often inspect the skin looking for varicose veins, melasma, increased thickness, distended teguments, among other skin conditions; visual inspection of limbs, i.e. vasospasm (vasoespasmo) is often related with vibration white finger syndrome. As observed in our previous study [25], the term deny (niega) is often expressed by the patient when asked if she has a particular condition, disease, or sufferers of adverse reactions to prescription drugs.

Words corresponding to the positive set refer to: the exploration of edemas, which lead to swollen ankles and feet; the state of the fetus, its heart rate (fcf) and movement; exploration for signs of renal pathologies such as Murphy, Mc Burney, and Giordano; inspection of ultrasonography data (usg); exploration of the uterine bottom height (afu); and assessment of the pregnancy stages. The term negative (negativo) qualitatively describe the result of a laboratory test.

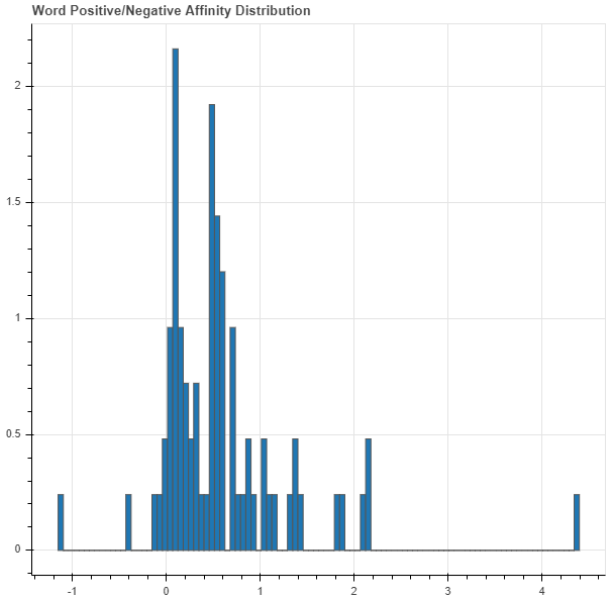

Figure 2. Word positive to negative affinity distribution

In summary, rare words, which show up in patients with unique conditions, are filtered-out, as we require word frequency usage to be greater than a given bound. Thus, intuitively, the tails in Fig. 2 are cut-off. Note that despite the distribution is multimodal, there is a greater amount of words which describe positively diagnosed patients. This is explained by the use of a richer vocabulary required to describe different variants or conditions related to a disease. Recall, words with positive-negative ratio close to zero represent words with similar frequency of usage in both classes of patients. Since there is almost twice as much EMRs corresponding to positively diagnosed patients than negative ones, the application of a large bound may lead to reduction of words in the negative set.

Fig. 3 analyses the relation between the positive rule set R_P and set P', which correspond to words with greater affinity to positively diagnosed patients.

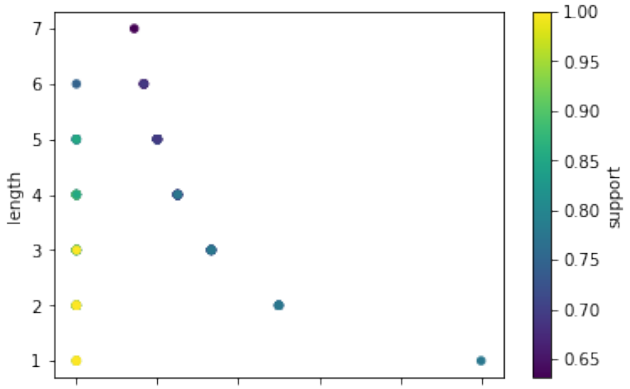

Figure 3. Correlation between R_P and P' representing Words with greater affinity to positive diagnosed patients

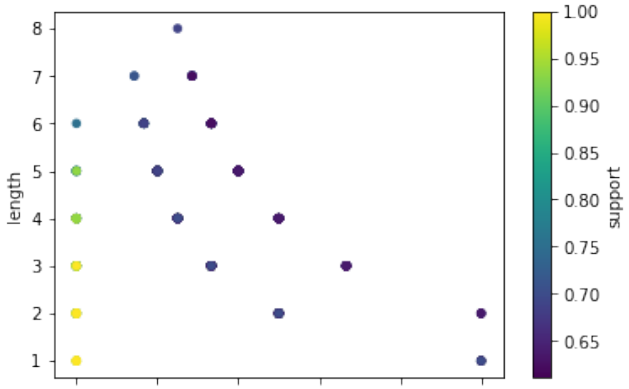

Figure 4. Correlation between R_P and N' representing Words with greater affinity to negative diagnosed patients

Table 3. Rule frequency per ratio for positive cases

Ratio	Mean support	Mean length	Numb. Rules	%
0.1	0.86	3	63	49.61
0.2	0.66	5.36	22	17.32
0.3	0.7	4	20	15.75
0.4	0.72	3	15	11.81
0.5	0.75	2	6	4.72
0.6-0.9	0	0	0	0.00
1	0.77	1	1	0.79

Table 4. Rule frequency per ratio for negative cases

Ratio	Mean support	Mean length	Numb. Rules	%
0.1	0	0	0	0.0
0.2	0.72	5.3	63	16.4
0.3	0.72	4.33	83	21.6
0.4	0.7	4.22	136	35.4
0.5	0.69	3.13	72	18.8
0.6	0	0	0	0.0
0.7	0.67	3	20	5.2
0.8-0.9	0	0	0	0.0
1	0.7	1.4	10	2.6

Set sizes for R_P and P' where of 127 and 25, correspondingly. The ration between $|w_r|$ and $|w_r \cap P'|$ represented by the figure abscissa. Rules with the highest ratio will have a value of one. The correspond to rules furthest to the right of the figure. Rules with lowest ratio will have a value of zero (left). Rules with highest ratio will have a value of one (right).

Approximately, 17% of rules corresponding to patients with a positive diagnosis have between 40% and 100% of their terms in P' (see table 3). This number is much lower than we expected. This result might be explained due to filtering of terms or acronyms with low utilization and/or there might be different types of diseases that branch from UTI positive patients. For the moment, we do not explore such hypothesis

Fig. 4 analyses the relation between the negative rule set R_N and set N', which correspond to words with greater affinity to negatively diagnosed patients. Set sizes for R_N and N' where of 384 and 25, correspondingly. The ration between $|w_r|$ and $|w_r \cap N'|$ represented by the figure abscissa. Support of rules with highest ratio where short, between 4 and 1 term, and support values was less than 0.8. Thus, they can be found in near 80% of EMRs of the negatively diagnosed patients.

Approximately 27% of the rules corresponding to patients with negative diagnosis have between 40% and 100% of their terms in N' (see table 4). Although we found a larger number of rules for this class, they are highly redundant. We have observed EMRs of negatively diagnosed patients include data that are less disease oriented and instead focus on gathering of demographic data. Thus, the resulting vocabulary is less diversified and words in it will have higher usage frequency.

6. Discussion

From the scientific point of view, this work is appealing because it addresses a real healthcare problem and it provides an initial approximation to find patterns in EMR data.

This work shows how to model the medical provider-patient interaction produced during auscultation. At this moment, input data source is in textual form, but as this work matures, we plan to augment data input from different sources (e.g., patients and other medical providers) and forms (e.g., voice conversations and patient behavioural data). Our work does not close the data context acquisition – modelling – reasoning

– and dissemination loop, nevertheless it examines how to extract and generalize knowledge needed for reasoning. As the rule set is improved, we are faced with the challenge applying them so that patient-caregiver interaction is empowered and the risk to patients is minimized.

The relation between the extracted rule set (association rules) and the unique word sets produced by the previous criterion is analysed. We found that 17% and 27% of rules, corresponding to patients with a positive and negative diagnosis, share between 40% to 100% of terms its corresponding unique word set.

The limitation of our results acknowledge the level of correspondence between the classification rules and its corresponding set of unique terms. However, the innovating impact of the proposed work include: strategies for text clinical data normalization; a criterion that facilitates the extraction of words that discriminate patients with and without the medical condition, and the generalization of shared procedure; the quality of the rules produced by a set of heuristics. The proposed method needs further research for it to be used as a diagnostic aid. To improve the quality of rules, the learning process must also pay attention to knowledge that facilitates the discrimination of patients (the unique word sets). We will address this issue in subsequent research.

The predictive value of the word sets and rules might seem to be limited, since data come from already taken diagnostics. However the quality and amount of data used to create the model will steadily improve as we complement it with new classes of data from other forms of interaction, i.e. physician-patient, physician-community, patient-community, UTI related papers, and online UTI forums or blogs. After the model predictive capability is improved, its effectiveness must be validated by domain experts to be used as a diagnostic aid. We are aware of such issues and will address them in continuing studies.

Lastly, the learning phase of the proposed method can easy be applied to EMRs with knowledge of other illnesses. However, the normalization phase must be tailored to meet the needs of the domain language.

7. Conclusions and future work

In this study, two hundred and seven, randomly chosen, medical records corresponding to pregnant patients with and without UTI were analysed. We analyse the log positive to negative word frequency ratio and extract word sets: words that characterize positively diagnosed patients; words that characterize negatively diagnosed patients; and words that are shared and describe similar features in both classes of patients. We propose a criterion for selecting the set sizes. A cut-off point of 47 (word frequency) is proposed so that rare words, which show up in patients with unique conditions, are filtered-out and discovered sets, which are unique for both patient classes, share the same number of terms. Our hypothesis is that such word sets generalize, adequately, each patient class. Thus, a learning algorithm should pay close attention to such word sets.

In future studies, we will seek to improve text normalization by grouping words, acronyms, and sentences with related knowledge in its immediate context (i.e. qualifiers). We will also increase our data set sizes as they small, as well as explores other sources of data that assess UTI related diseases. In long-term future we would like to support medical providers' training and support by suggesting prescriptions and treatment.

While we present some interesting results, we lack knowledge if these generalize biases that may be present in data, i.e. when physicians intentionally favour prescription of a pharmaceutical; when physicians conduct racial profile. The effect of social bias in data has been identified by Bolukbasi et al. [30] and Caliskan et al. [31]. In future research, we plan to analyse social biases in healthcare data.

References

[1] Alkema L, Chou D, Hogan D, Zhang S, Moller AB, Gemmill A, et al. Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. The Lancet. 2016 Jan;387(10017):462–474.

[2] You D, Hug L, Ejdemyr S, Idele P, Hogan D, Mathers C, et al. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation. The Lancet. 2015 Dec;386(10010):2275–2286.

[3] Blencowe H, Cousens S, Jassir FB, Say L, Chou D, Mathers C, et al. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. The Lancet Global Health. 2016 Feb;4(2):98–108.

[4] de Salud S. Informes Semanales para la Vigilancia Epidemiológica de Muertes Maternas 2019; 2019. Accessed: 14-Oct-2019. https://www.gob.mx/salud.

[5] Soto-Estrada G, Moreno-Altamirano L, Díaz DP. Panorama epidemiológico de México, principales causas de morbilidad y mortalidad. SciELO. 2016;59(6):8–22.

[6] Medic CV, del Rosario López Villegas M, Ángel Enríquez Guerra M, Valverde BR. Prevalencia de infecciones de vías urinarias en embarazadas atendidas en el Hospital Universitario de Puebla. Enfermedades Infecciosas y Microbiología. 2010 Oct;30(4):118–122.

[7] SEGOB. Diario Oficial de la Federación; 2019. Accessed: 14-Oct-2019. https://www.dof.gob.mx/.

[8] Frenk J, González-Pier E, Gómez-Dantés O, Lezana MA, Knaul FM. Comprehensive reform to improve health system performance in México. The Lancet. 2006 Oct;368(9546):1524–1534.
[9] Cantón SF. El IMSS en cifras: la mortalidad en la población derechohabiente, 2003. Revista Médica del Instituto Mexicano del Seguro Social. 2004 Apr;42(4):353–364.
[10] Castañón-González JA, Barrientos-Fortes T, Polanco-González C. Reflections concerning the care process in the emergency medical services. Revista Médica del Instituto Mexicano del Seguro Social. 2016;54(3):376–379.
[11] Piro R, Nenov Y, Motik B, Horrocks I, Hendler P, Kimberly S, et al. Semantic technologies for data analysis in health care. In: Groth P, Simperl E, Gray A, Sabou M, Krötzsch M, Lecue F, et al., editors. The Semantic Web – ISWC 2016. Cham: Springer International Publishing; 2016. p. 400–417.
[12] Gefen D, Miller J, Armstrong JK, Cornelius FH, Robertson N, Smith-McLallen A, et al. Identifying patterns in medical records through latent semantic analysis. Communications of the ACM. 2018 May;61(6):72–77. Available from: http://doi.acm.org/10.1145/3209086.
[13] Fernández-Breis JT, Maldonado JA, Marcos M, del Carmen Legaz-García M, Moner D, Torres-Sospedra J, et al. Leveraging electronic healthcare record standards and semantic web technologies for the identification of patient cohorts. Journal of the American Medical Informatics Association. 2013 Dec;20(e2):288–296.
[14] Stilou S, Bamidis PD, Maglaveras N, Pappas C. Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. Studies in health technology and informatics. 2001;84:1399–1403.
[15] Creighton C, Hanash S. Mining gene expression databases for association rules. Bioinformatics. 2003 Jan;19(1):79–86.
[16] Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent advances and emerging applications in text and data mining for biomedical discovery. Briefings in Bioinformatics. 2015 Sep;17(1):33–42.
[17] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems. 2014 Feb;2(1).
[18] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011 May;7(3):263–269.
[19] Wright AP, Wright AT, McCoy AB, Sittig DF. The use of sequential pattern mining to predict next prescribed medications. Journal of Biomedical Informatics. 2015 Feb;53:73–80.
[20] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data. 2019 May;6(1).
[21] Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences. 2018 Apr;115(16):3635–3644.
[22] Zhu Z, Yin C, Qian B, Cheng Y, Wei J, Wang F. Measuring patient similarities via a deep architecture with medical concept embedding. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE; 2016. .
[23] Glicksberg BS, Miotto R, Johnson KW, Shameer K, Li L, Chen R, et al. Automated disease cohort selection using word embeddings from electronic health records. PubMed. 2018;23:145–156.
[24] de Salud S. Normal Oficial Mexicana NOM-168-SSA1-1998 del expediente clínico; 2016. Accessed: 14-Oct-2019. http://www.salud.gob.mx.
[25] Hirales-Carbajal A, Escobedo L. UTIW: Urinary Tract Infection Workflow system towards early and automatic detection of urinary infection during pregnancy. In: Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare. PervasiveHealth'19. New York, NY, USA: ACM; 2019. p. 462–468. Available from: http://doi.acm.org/10.1145/3329189.3329238.
[26] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. SIGMOD Rec. 1993 Jun;22(2):207–216. Available from: http://doi.acm.org/10.1145/170036.170072.
[27] Ait-Mlouk A, Agouti T, Gharnati F. Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. Journal of Big Data. 2017 Nov;4(1):42. Available from: https://doi.org/10.1186/s40537-017-0105-4.
[28] Bayardo RJ Jr, Agrawal R. Mining the most interesting rules. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '99. New York, NY, USA: ACM; 1999. p. 145–154. Available from: http://doi.acm.org/10.1145/312129.312219.
[29] Han J. Data Mining: Concepts and Techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2005.
[30] Bolukbasi T, Chang K, Zou JY, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Clinical Orthopaedics and Related Research. 2016;abs/1607.06520. Available from: http://arxiv.org/abs/1607.06520.
[31] Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017;356(6334):183–186. Available from: https://science.sciencemag.org/content/356/6334/183.