Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-85

A methodology for mining clinical data: experiences from TRANSFoRm project

Roxana DANGER[a,1], Derek CORRIGAN[b], Jean K. SOLER[c], Przemyslaw KAZIENKO[d], Tomasz KAJDANOWICZ[d], Azeem MAJEED[a], Vasa CURCIN[e]

[a] *Imperial College London, London, UK*
[b] *Royal College of Surgeons in Ireland, Dublin, Ireland.*
[c] *Mediterranean Institute for Primary Care, Attard, Malta*
[d] *Wroclaw University of Technology, Wroclaw, Poland*
[e] *King's College London, London, UK*

Abstract. Data mining of electronic health records (eHRs) allows us to identify patterns of patient data that characterize diseases and their progress and learn best practices for treatment and diagnosis. Clinical Prediction Rules (CPRs) are a form of clinical evidence that quantifies the contribution of different clinical data to a particular clinical outcome and help clinicians to decide the diagnosis, prognosis or therapeutic conduct for any given patient. The TRANSFoRm diagnostic support system (DSS) is based on the construction of an ontological repository of CPRs for diagnosis prediction in which clinical evidence is expressed using a unified vocabulary. This paper explains the proposed methodology for constructing this CPR repository, addressing algorithms and quality measures for filtering relevant rules. Some preliminary application results are also presented.

Keywords. E05.245, L01.700.508.190- Decision Support Techniques, Clinical Prediction rules; L01.470.625 – Data mining

Introduction

Clinical prediction rules (CPRs) quantify the contribution of symptoms, signs, diagnostic tests, demographic features and risk factors to a particular clinical outcome[1]. The outcome of interest can be diverse and range across the diagnostic, prognostic, and therapeutic spectrum[2]. A healthcare decision support system (DSS) can make use of high quality CPRs to support clinician's decision-making during a medical consultation. In this way, diagnosis, test requests, interventions, etc. can be reinforced by previous research. The number and quality of validated CPRs in the biomedical literature is low in comparison with other types of health care study rules formulation (such as, test discrimination value). Fahey and van der Lei have only found 60 articles summarizing CPRs in diverse clinical domains[2]. However, these rules use different terminologies and lack sufficient validation across different populations, making many of them unacceptable for use by clinicians out of the original research context.

One of the goals of the EU FP7 TRANSFoRm project is to provide a framework for describing CPRs to populate a rule repository through a set of evidence extracted

[1] Corresponding Author.

from electronic primary care datasets. Rules are expressed as axioms of an ontology[3] for diagnosis prediction in which clinical evidence is expressed using a unified vocabulary and therefore can be shared, validated in different populations and then published through both the rule repository and the Diagnostic Support System software. Reliable CPRs need to be supported by real clinical data, so data mining[4] tools are used to extract empirically quantified knowledge behind these CPRs. The clinical interpretation of what the quantified data means for CPRs then needs to be provided by clinical review using the experience, common sense and standardized language used by the clinicians.

Current methodologies for constructing CPRs are based on probabilistic or Bayesian reasoning, and have been defined with only one dichotomous variable in the antecedent, with quality measures (QM) such as apriori-probability, sensitivity, specificity, posterior probability, likelihood ratio (LR) and odds ratio (OR) used to describe the quality of the CPRs[2, 6]. In this paper we propose to use data mining algorithms for frequent pattern extraction to provide the empirically quantified variables upon which CPRs can be defined. We select a set of measures that can capture different quality aspects of the rules, while retaining standard measures as well.

1. Methods

To differentiate CPRs from the rule facts extracted directly from the data mining processes, we define Measured Clinical Evidence Rules (MCERs). An MCER has the structure *antecedent → consequent (QM)* and describes a dependence between clinical data (antecedent and the consequent of the rule) together with a set of QM tuples *(measure, value).* An optional *condition* can be added to the tuple to denote that a measure refers to a subset of conditions in the rule. Data in MCERs use the same set of concepts previously identified for CPRs[3]; namely reason for encounter (RfE), diagnosis, diagnostic cue, quantification and population. Thus, a typical rule would be:

Rfe_Y02 (Pain in testis/scrotum) →Y74 (Orchitis/epididymitis) (Support=89, Confidence=32.4, Lift=548.63, LR+=810.67, LR-=0.55, OR=1465.77)

Both of the elements (Rfe_Y02 and Y74) of the above rule were found in 89 episodes of care of the database. However, from all the cases referring to testis or scrotum pain, almost a third (32.4%) were effectively diagnosed with Orchitis. That makes the symptom a strong predictor for the disease, also supported by the large values of the LR+, lift and OR (significantly greater than 1).

MCERs are ideally obtained from longitudinal data, which describe the cause-effect dependencies for any combination of variables; and also consider the correlation factors between all the analysed elements, pre- and post- probabilities of each cause and associate a set of measures that guarantee certain level of rule quality. The problem of computing MCERs is formulated as: given the dataset containing the clinical records (including demographic features, RfEs, symptoms, signs, risk factors, tests performed, diagnosis) of patients, extract the rules to associate these factors. We limit this study to the computation of two types of MCERs for determining the etiology of diseases:

- **[RULE PATTERN 1], when details of only one consultation are used:** Demographic features, RfEs, Symptoms, Signs, Risk factors, Tests performed → Diagnosis(QM) or
- **[RULE PATTERN 2], when more than one consultation is used to produce a pattern:** (Demographic Features, RfEs, Symptoms, Signs, Risk factors, Tests performed, time since previous episode)* → Diagnosis(QM).

Two different timings of the patient evolution have to be considered independently:

- **First consultation of an Episode of Care (EoC)**[7]. Rules for this timing schema describe the first diagnosis impression in the first consultation, using the [RULE PATTERN 1]. In most cases the first diagnosis is the one maintained in the complete clinical history of a patient; so, mining MCERs at this moment is extremely important.
- **Consequent consultations of an EoC**. In this case, performed tests and consultation date variables also have to be included.

Such MCER rule patterns can be used to generate executable rules in a procedural language such as the Arden Syntax [8].

Algorithms and quality measures for MCERs

We propose the use of decision trees, dependence rules extraction and statistically significant sequential patterns algorithm types for extracting MCERs. Associated with each MCER, is a set of quality measures based either on the variables in the entire rule, or on its subset (condition), including confidence intervals guaranteeing minimal statistical significance. These algorithms provide outputs with high characterizing power, without independence assumption, while being human-readable and easy to understand. For assessing the quality of the rules we have selected the following set of measures (full formulas are omitted, but can be found in [2,4,5,6]):

- to characterize the consequent (disease) interest: apriori-probality;
- to characterize each variable in the antecedent: posterior probability and positive and negative likelihood ratios (LR+, LR-)
- to characterize the whole set of variables in the rule: support;
- to characterize the rule interest: lift, confidence, conviction, sensitivity, specificity, error rate, posterior probability, positive and negative likelihood ratio (LR+, LR-) and odds ratio (OR).

2. Results

To validate our proposed technique we tested and compared the results of obtaining MCERs following [RULE PATTERN 1] for patients' first consultations using three data mining algorithms[6]: Apriori for positive association rules, KingFisher for positive and negative association rules and C4.5 for decision trees. The dataset used was TRANSHIS[7],,a collection of clinical details from 126931 Netherlands patients (*age, gender, RfE* and *diagnosis*) that were collected by family doctors who participated in the Transition Project. Diagnosis and RfE were coded in ICPC2, and the age was converted into a categorical variable, splitting the age values into 5-year groups up until the age of 84, with a single group for ages 85 or over.

While Apriori is a straightforward algorithm that requires few parameters to be configured, KingFisher and C4.5 require some parameter adjustments to obtain an optimal solution. Table 1 describes the final, optimal parameters selected for each algorithm and the number of rules extracted, before and after filtering.

Table 1. Optimal parameters and number of generated rules per algorithm

Algorithm	Optimal parameters	No. rules	
		Before filtering	after filtering
Apriori	Min. support=10	33302	17381
KingFisher	M=50	3370	824
C4.5	cp=0.01	63	63

The MCERs obtained were then filtered, leaving only those strong or weak predictor rules, according to[7]. As the next step we analysed the rules extracted for Apriori and KingFisher algorithms, as those generated by C4.5 were the same as the top 63 rules generated by Apriori, in terms of LR+.

The 15% of the obtained rules with Apriori related RfE with the same diagnosis. This is the case of: *Eye infection, Nose symptom, Depressive disorder, Neurological symptom, Infected finger/toe, Pregnancy, Rectal bleeding,* and *Asthma,* which appear amongst the 30 rules with highest LR+, and *Rectal bleeding, Knee symptom, Hand/finger symptom, Ear pain, Wrist symptom,* and *ringing/buzzing ear* amongst the 30 rules with lowest LR-. In the case of KingFisher the 5% of the resulting rules related RfE with the same diagnosis, as in the cases of: *Vertigo/dizziness, Pain/tenderness of skin, Urinary frequency/urgency, Swollen ankles/oedema, Ear pain/earache,* and *Vaginal discharge* which appears amongst the first 30 rules with highest LR+; and *Contraception oral, Low back symptom, Neck symptom, Swollen ankles/oedema, Chest symptom, Weakness/tiredness general, Heartburn, Feeling anxious/nervous/tense* and *Lymph gland(s) enlarged/painful* amongst the 30 rules with lowest LR-.

More interesting are those rules where RfE and diagnosis differ. We have chosen the best 5 rules, according to LR+, LR-, and combined LR+ and itemset support, as shown in Table 2. The coincident rules between the two algorithms have been highlighted in the table. The large majority of rules shown are between an RfE and the diagnosis with 13% of the rules having more than two elements in the antecedent. Despite our interest in obtaining rules with negative factors amongst the antecedent elements, KingFisher could not find any useful association rules with negative variables.

Table 2. MCERs obtained per algorithm: best five sorted by the criterion on the first column. The values between parenthesis are referred to rule support, confidence, lift, LR+, LR- and OR.

	Apriori	
LR+	Rfe_L15 (Knee symptom);AgeGroup=[10-14];gender=Masculine → L94 (Osteochondrosis)	(32,18.8,774.4,953.74,0.61,1563.49)
	Rfe_Y02 (Pain in testis/scrotum) →Y74 (Orchitis/epididymitis)	**(89,32.4,548.63,810.67,0.55,1465.77)**
	Rfe_W03 (Antepartum bleeding) →W82 (Abortion spontaneous)	(55,30.1,530.83,758.49,0.71,1064.83)
	Rfe_F03(Eye discharge);Rfe_F02 (; Red eye) →F70	(136,87.2,91.28,705.17,0.96,736.20)
	Rfe_X12 (Malignant neoplasm genital female other)→ X77 (Postmenopausal bleeding)	(12,5.3,660.49,697.30,0.56,1254.35)
LR-	Rfe_L15 (Knee Symptom/complain) →L96(Acute internal damage knee)	(307,6.4,58.67,62.59,0.16,383.07)
	Rfe_L08 (Shoulder symptom/complaint) →L92 (Shoulder synd.)	**(1758,35,54.36,83.10,0.193,430.90)**
	Rfe_R05 (Cough) →R71 (Whooping cough)	(237,1.2,14.29,14.45,0.194,74.61)
	Rfe_R21 (Throat symptom)→R76 (Tonsillitis acute)	**(1802,22.5,32.60,41.76,0.23,181.68)**
	Rfe_L15 (Knee Symptom/complain)→L78 (Sprain/strain of knee)	(231,4.8,52.80,55.41,0.25,222.98)
Support	**Rfe_R05(Cough) → R78 (Acute bronchitis)**	**(4717,24.4,12.46,16.17,0.3053.87)**
	Rfe_R05 (Cough)→ R74 (Upper respiratory infection acute)	**(3288,17,7.19,8.46,0.62,13.68)**
	Rfe_S06 (Rash localized)→ S88 (Dermatitis contact/allergic)	**(2066,23,15.28,19.56,0.61,32.27)**
	Rfe_H02 (Hearing complain)→ H81 (Excessive ear wax)	(2034,46.5,25.28,46.41,0.68,68.53)
	Rfe_H13 (Plugged feeling ear)→H81 (Excessive ear wax)	(1952,74.8,40.63,158.19,0.69,230.25)
	KingFisher	
LR+	**Rfe_Y02 (Pain in testis/scrotum) → Y74 (Orchitis/ epididymitis)**	**(89, 32.4,548.63,991.72,0.68,1465.77)**
	Rfe_Y04 (Penis symptom/complaint other) → Y75 (Balanitis)	(224,53.97,275.05,596.46,0.66,900.99)
	Rfe_X19 (Breast lump/mass female) → X88 (Fibrocystic disease breast)	(56,46.67,270.96,507.18,0.90,561.17)
	Rfe_X19 (Breast lump/mass female) → X76 (Malignant neoplasm breast female)	(41,43.16,250.59,440.09,0.93,473.43)
	Rfe_D04 (Rectal/anal pain) → D95 (Anal fissure/ perianal abscess)	(116,30.29,258.67,370.61,0.71,524.28)

LR-	Rfe_L10 (Elbow symptom/complaint) → L93 (Tennis elbow)	(583,54.84,139.63,308.02,0.56,549.25)
	Rfe_L16 (Ankle symptom/complaint) → L77 (Sprain/strain of ankle)	(618,45.98,100.99,186.10,0.60,310.72)
	Rfe_D11 (Diarrhoea) → D73 (Gastroenteritis presumed infection)	(1268,52.48,50.38,104.93,0.64,163.61)
	Rfe_L08 (Shoulder symptom/complaint) → L92 (Shoulder synd.)	**(1758,35.54,54.36,280.38,0.65,430.90)**
	Rfe_Y04 (Penis symptom/complaint other) → Y75 (Balanitis)	(224,53.97,275.05,596.46,0.66,900.99)
Support	**Rfe_R05 (Cough) → R78 (Acute bronchitis)**	**(4717,71.30,12.46,40.94,0.76,53.87)**
	Rfe_R05 (Cough) → R74 (Upper respiratory infection acute)	**(3288,41.13,7.19,11.52,0.84,13.68)**
	Rfe_S06 (Rash localized) → S88 (Dermatitis contact/allergic)	**(2066,40.65,15.28,25.07,0.78,32.27)**
	Rfe_H01 (Ear pain/earache) → H71 (Acute otitis media/ myringitis)	(1898,60.68,35.81,89.53,0.67,133.53)
	Rfe_R21 (Throat symptom/) → R76 (Tonsillitis acute	**(1802,77.44,32.60,141.05,0.78,181.68)**

3. Discussion

In this paper we have described a methodology for extracting CPRs based on MCERs extracted by data mining algorithms. We have defined the most important factors that should be included in the rules and explained the patterns that interesting rules should follow. We have developed an analytical workflow to explore the resulting MCERs from Association rules, Decision Trees and Sequential patterns, and some preliminary results have proven the usefulness of using data mining algorithms for computing MCERs. Extracted rules are clinically reviewed, grouped by target condition, and compared against evidence based clinical guidelines describing that condition as found in literature (national guidrlines, JAMA reviews etc.). The literature indicates the generated rules agree with and are 'clinically sensible' based on that gold standard.
With MCERs, medical researchers can select, organize and group rules with the same resulting disease and associate normalized scores to each factor or combination of factors in such a way that diagnostic recommendations can be formulated.

In the future, we aim to extend the experiments with other datasets, goals of analysis and publish our tools to the biomedical research community.

Acknowledgment
The TRANSFoRm project (www.transformproject.eu) is partially funded by the European Commission - DG INFSO (FP7 247787).

References

[1] McGinn T, et. al. Diagnosis: clinical prediction rules, in: Guyatt G, R.D.e. (Ed.), Users' guides to the medical literature. Chicago: American Medical Association; 2004.
[2] Fahey T, van der Lei J. Producing and Using Clinical Prediction Rules. In: Knottnerus J, Buntinx F, eds. *The Evidence Base of Clinical Diagnosis*: Wiley-Blackwell; 2009:213-236.
[3] Corrigan D, Soler JK, Delaney B. Development of an ontological model of evidence for transform utilizing transition project data. Proc. of the Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine; 2012.
[4] Fayyad U et. al.. From data mining to knowledge discovery in databases. AI Magazine 1996;17:37–54.
[5] Duan L, Street WN, Liu Y, Xu S, Wu B. Right Correlation Measure for Binary Data. ACM Transactions on Knowledge Discovery from Data 2014;9(2):Art.1.
[6] Aggarwal Ch., Han J. (Eds) Frequent Pattern Mining. Springer, 2014.
[7] Soler JK, et al. An international comparative family medicine study of the transition project data from the Netherlands, Malta and Serbia. Is family medicine an international discipline? comparing diagnostic odds ratios across populations. Fam Pract 2012;29:299–314.
[8] HL7. Health Level Seven Arden Syntax for Medical Logic Systems, Version 2.10, http://www.hl7.org/implement/standards/product_brief.cfm?product_id=372. HL7 International 2014.