Decision Support Systems and Education
J. Mantas et al. (Eds.)

doi:10.3233/978-1-61499-921-8-15

Secondary Use of Healthcare Structured Data: The Challenge of Domain-Knowledge Based Extraction of Features

Emmanuel CHAZARD[a,b,1], Grégoire FICHEUR[a,b], Alexandre CARON[a,b], Antoine LAMER[b], Julien LABREUCHE[b], Marc CUGGIA[c,d], Michaël GENIN[a], Guillaume BOUZILLE[c,d] and Alain DUHAMEL[a,b]
[a] *CERIM EA2694, Lille University, F-59000 Lille*
[b] *Public Health Department, CHU Lille, F-59000 Lille*
[c] *LTSI, Université de Rennes 1, F-35000 Rennes, France*
[d] *CIC & CIC-IT Inserm 1414, CHU Rennes, F-35000 Rennes, France*

Abstract. Secondary use of clinical structured data takes an important place in healthcare research. It was first described by Fayyad as "knowledge discovery in databases". Feature extraction is an important phase but received little attention. The objectives of this paper are: 1) to propose an updated representation of data reuse in healthcare, 2) to illustrate methods and objectives of feature extraction, and 3) to discuss the place of domain-specific knowledge. Material and methods: an updated representation is proposed. Then, a case study consists of automatically identifying acute renal failure and discovering risk factors, by secondary use of structured data. Finally, a literature review published par Meystre et al. is analyzed. Results: 1) we propose a description of data reuse in 5 phases. Phase 1 is data preprocessing (cleansing, linkage, terminological alignment, unit conversions, deidentification), it enables to construct a data warehouse. Phase 2 is feature extraction. Phase 3 is statistical and graphical mining. Phase 4 consists of expert filtering and reorganization of statistical results. Phase 5 is decision making. 2) The case study illustrates how time-dependent features can be extracted from laboratory results and drug administrations, using domain-specific knowledge. 3) Among the 200 papers cited by Meystre et al., the first and last authors were affiliated to health institutions in 74% (68% for methodological papers, and 79% for applied papers). Discussion: features extraction has a major impact on success of data reuse. Specific knowledge-based reasoning takes an important place in feature extraction, which requires tight collaboration between computer scientists, statisticians, and health professionals.

Keywords. Data reuse, feature extraction, data transformation.

1. Introduction

Secondary use of clinical data can be defined as "non-direct care use of personal health information" [1,2], notably for research purposes. This concept was first approached by Fayyad, under the term "knowledge discovery in databases". Five steps were then defined [3]: data selection, data preprocessing, data transformation, data mining, and interpretation. The term "data reuse" or "secondary use of data" denotes the same process,

[1] Emmanuel Chazard, CERIM EA2694, Medicine Faculty, 1 place de Verdun F-59045 Lille cedex, France; E-mail: emmanuel.chazard@univ-lille.fr.

but became popular later [4]. It focusses on the data source instead of the goal, as it appeared that new knowledge was not so easy to obtain.

Data reuse presents several advantages compared with traditional methodologies [1,5,6]: studies are faster because they don't imply specific data collection, they cost less, they enable to analyze high sample sizes (and sometimes big data [7]), and to obtain a high statistical power or to study rare events, and they enable to build historical cohorts. It also presents drawbacks [1,5,6]: less scientific questions are addressable than using traditional methodologies, data quality may be insufficient, and it is more difficult to manage confounding factors and indication bias. Finally, it is difficult to perform.

In our experience, the most critical part of structured-data reuse is the phase initially called "data transformation" [3] or "data aggregation" [8,9], and more recently "feature extraction". Despite its critical impact on scientific results, it has been discussed for signal, image and free-text processing, but sparsely for structured data analysis.

The objectives of this paper are: 1) to propose an updated representation of data reuse in healthcare, including feature extraction, 2) to discuss methods and objectives of feature extraction, and 3) to discuss the place of domain-specific knowledge.

2. Material and methods

To get an updated representation of structured data reuse in healthcare, the authors took profit from their experience (~150 studies performed) and scientific readings in that field.

To characterize objectives and methods of feature extraction, a case study taken from the PSIP Project [10] was analyzed. By reusing data from electronic health records (including laboratory results and administered drugs of 175,000 inpatient stays), the aim of the case study was to automatically detect acute failure, and identify risk factors.

To assess the importance of domain-specific knowledge to reuse healthcare data, a literature review published par Meystre *et al.* in 2017 [1] was analyzed. The 231 references cited by this review were searched for on Pubmed. Scientific papers were classified as methodological or applied papers. The affiliation of their authors was also analyzed, and their institutions were classified as "health" (for medicine faculties, hospitals, etc.), or "non-health" otherwise.

3. Results

3.1. Data reuse process in Healthcare

We propose the process illustrated on Figure 1. This figure also enables to compare traditional methodologies based on questionnaire data, and structured data reuse. The healthcare data reuse process consists of 5 phases. Previously, routine data collection is performed for instance for patient care. It is independent from the study and enables to obtain reusable databases.

Then, the first phase consists of data preprocessing. This includes data cleansing, data linkage, terminological alignment, unit conversions, data deidentification, and securitization. This phase enables to construct a data warehouse, which does not depend primarily on the further intended use, but rather on the available data.

The second phase consists of features extraction. Data from the warehouse are generally raw, not question-oriented, and complex (e.g. many tables). Those data are then

transformed into questionnaire-like data, often made of one table, with one row per statistical individual, and one column per variable. This phase is detailed in next section.

The third phase consists of statistical and graphical mining. It aims at producing more compact and useful information, such as distribution metrics (e.g. mean, conditional proportion), graphical representations, and statistical associations or models.

As the results obtained from the previous phase may be too abundant (e.g. thousands of association rules) and may suffer from various biases which may lead to false knowledge discovery, they generally need to be filtered, validated and reorganized by experts. This fourth phase enables to get new knowledge.

Finally, the fifth phase consists of decision making.

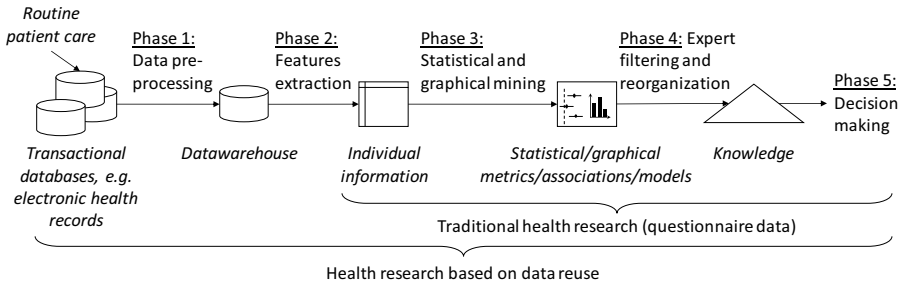

Figure 1. Process of secondary use of structured data in healthcare.

3.2. Methods and objectives of feature extraction

In this case study, according to the researchers who performed the analysis, the objectives of feature extraction were:

- To reduce data complexity to one single data table (with one row per statistical individual, and one column per variable). For instance, a multivalued categorical variable was transformed into a set of binary variables.
- To introduce domain specific thresholds for some quantitative variables (see example of laboratory results below), which enabled identifying abnormalities, and then to get precise start and stop dates for those abnormalities.
- To reduce data imbalance, by enabling knowledge-based consistent grouping (see example of drug administrations below). Such grouping made it easier to discover statistical associations, enabled predictive models to face new situations by being more inclusive, and facilitated the further process of expert-operated knowledge discovery from statistical associations.
- To handle heterogeneous data in the form of generic time-dependent events.
- To make results more acceptable for experts, as validated criteria were used (see example of acute renal failure detection below).

Some variables could directly be analyzed, such as age and gender. Other features had to be extracted. Laboratory results most often consist of functional data. A common way to extract features is to consider expert-defined normality thresholds. This enabled to define temporal events (Figure 2). Sometimes, missing data can also be imputed: it is commonly admitted that, if a patient had any symptom in relation with hypoglycemia or hyperglycemia, glycemia would immediately have been measured. In this case study, the absence of value was inferred as the absence of hypoglycemia and hyperglycemia.

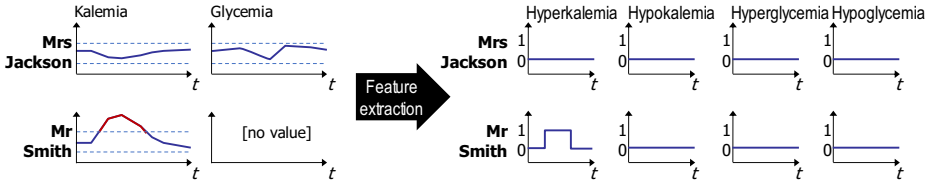

Figure 2. Example of feature extraction from laboratory values (left: raw data, right: features)

A similar process could be performed for administered drugs. In the case described on Figure 3, the 2 drugs enabled to extract 4 features. Such features were inferred from drug names using an expert-designed mapping. It is worth noting that both drugs been hepatic enzyme inductors, the dates of the corresponding feature took it into account.

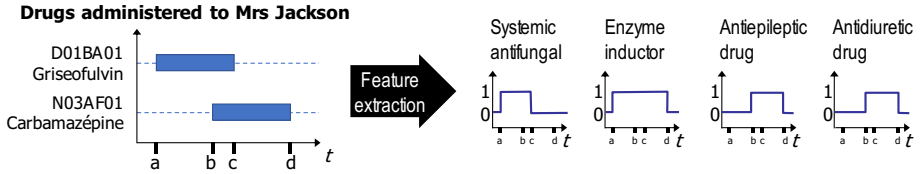

Figure 3. Example of feature extraction from administered drugs (left: raw data, right: features)

Finally, some feature extractions require specific domain-knowledge based reasonings. In our case study, to define the outcome (acute renal failure), we used the KDIGO criteria [11], and implemented a moving window to extract features (Figure 4). This feature extraction was specific and was not suitable for other laboratory parameters.

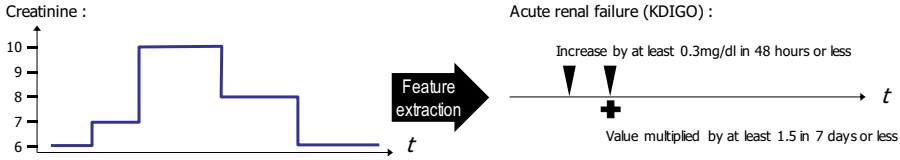

Figure 4. Example of feature extraction according to KDIGO criteria (left: raw data, right: features)

In our case study, we could then extract 666 types of feature: 48 from chronic diagnoses (~35,000 codes initially), 568 from drugs (~5,400 codes initially), 35 from laboratory results (~200 parameters initially), and 15 from administrative variables. Temporal associations could then be mined [10].

3.3. Place of domain-specific knowledge

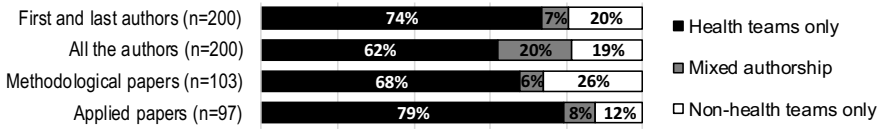

Figure 5. Classification of the scientific papers from Meystre et al.'s review [1].

Among the 231 references cited by Meystre et al. [1], 200 were scientific papers. All the authors were affiliated to health institutions for 62% papers. This proportion raised up to 74% when only the first and last authors were considered. Among those papers, 103 were

methodological papers, and 97 were applied papers. The proportion of papers with first and last authors both affiliated to health institutions was 68% for methodological papers, and 79% for applied papers. Those results are detailed on Figure 5.

4. Discussion

In the present papers, we proposed an updated schema of data reuse process for healthcare structured data. We illustrated objectives and methods of feature extraction. Finally, we observed that feature extraction was mainly based on domain knowledge, and we could confirm that most publications in the field of data reuse were written by health research teams, even for methodological papers.

Features extraction has a major impact on success of secondary use of structured data [3]. The complete data reuse process requires a synergistic collaboration of different skills, namely informatics, statistics, and health sciences [12,13]. Those skills are not required sequentially, but intertwine at each phase, which we have illustrated for the feature extraction step. This is even more important in data mining, when studies do not focus on a precise outcome and a precise exposure [10]. This highlights the need for training dedicated professionals, called data scientists [14].

References

[1] S.M. Meystre, C. Lovis, T. Bürkle, G. Tognola, A. Budrionis, C.U. Lehmann, Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress, *Yearb Med Inform*. **26** (2017). doi:10.15265/IY-2017-007.

[2] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, and null Expert Panel, Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper, *J Am Med Inform Assoc*. **14** (2007) 1–9. doi:10.1197/jamia.M2273.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery in Databases, *AI Magazine*. **17** (1996) 37.

[4] A. Dan Corlan, Medline trend: automated yearly statistics of PubMed results for any query, (2004). http://dan.corlan.net/medline-trend.html (accessed August 29, 2016).

[5] C. Safran, Using routinely collected data for clinical research, *Stat Med*. **10** (1991) 559–564.

[6] C. Safran, Reuse of clinical data, *Yearb Med Inform*. **9** (2014) 52–54. doi:10.15265/IY-2014-0013.

[7] E. Baro, S. Degoul, R. Beuscart, E. Chazard, Toward a Literature-Driven Definition of Big Data in Healthcare, *Biomed Res Int*. **2015** (2015). doi:10.1155/2015/639021.

[8] E. Chazard, G. Ficheur, B. Merlin, M. Genin, C. Preda, PSIP consortium, and R. Beuscart, Detection of adverse drug events detection: data aggregation and data mining, *Stud Health Technol Inform*. **148** (2009) 75–84.

[9] A. Lamer, M. Jeanne, G. Ficheur, R. Marcilly, Automated Data Aggregation for Time-Series Analysis: Study Case on Anaesthesia Data Warehouse, *Stud Health Technol Inform*. **221** (2016) 102–106.

[10] E. Chazard, G. Ficheur, S. Bernonville, M. Luyckx, R. Beuscart, Data mining to generate adverse drug events detection rules, *IEEE Trans Inf Technol Biomed.* **15** (2011), 823–830. doi:10.1109/TITB.2011.2165727.

[11] A. Khwaja, KDIGO clinical practice guidelines for acute kidney injury, *Nephron Clin Pract* **120** (2012), c179-184. doi:10.1159/000339789.

[12] Harnessing big data. How to achieve value, *Hosp Health Netw* **88** (2014), 61–71.

[13] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, R. Yu, U. Sikora, A. Ozcan, Crowd-sourced BioGames: managing the big data problem for next-generation lab-on-a-chip platforms, *Lab Chip*. **12** (2012), 4102–4106. doi:10.1039/c2lc40614d.

[14] T.H. Davenport, D.J. Patil, Data scientist: the sexiest job of the 21st century, *Harv Bus Rev* **90** (2012), 70–76, 128.