A Similarity Measure Based on Care Trajectories as Sequences of Sets

Yann Rivault, Nolwenn Le Meur, Olivier Dameron

A Similarity Measure Based on Care Trajectories as Sequences of Sets

Yann Rivault[1,3,4], Nolwenn Le Meur[1,4], Olivier Dameron[2,3,4]

[1] EHESP Rennes, Sorbonne Paris Cité, EA 7449 REPERES, Recherche en Pharmaco-Epidémiologie et Recours aux Soins, France
[2] Université de Rennes 1, 35000 Rennes, France
[3] IRISA équipe Dyliss, 35042 Rennes
[4] PEPS, Pharmacoepidemiology for health products safety

{Yann.Rivault, Nolwenn.LeMeur}@ehesp.fr
Olivier.Dameron@univ-rennes1.fr

Abstract. Comparing care trajectories helps improve health services. Medico-administrative databases are useful for automatically reconstructing the patients' history of care. Care trajectories can be compared by determining their overlapping parts. This comparison relies on both semantically-rich representation formalism for care trajectories and an adequate similarity measure. The longest common subsequence (LCS) approach could have been appropriate if representing complex care trajectories as simple sequences was expressive enough. Furthermore, by failing to take into account similarities between different but semantically close medical events, the LCS overestimates differences. We propose a generalization of the LCS to a more expressive representation of care trajectories as sequences of sets. A set represents a medical episode composed by one or several medical events, such as diagnosis, drug prescription or medical procedures. Moreover, we propose to take events' semantic similarity into account for comparing medical episodes. To assess our approach, we applied the method on a care trajectories' sample from patients who underwent a surgical act among three kinds of acts. The formalism reduced calculation time, and introducing semantic similarity made the three groups more homogeneous.

Keywords: Care trajectories · LCS-based similarity · Semantic similarity

1 Introduction

Medico-administrative databases are valuable data source for health research notably because of their large population coverage, as well as of their longitudinal properties [1]. In France, the French national health insurance inter-regime information system (SNIIRAM) records the reimbursements of health care covered by the main insurance funds for workers. These data include ambulatory care data and hospital discharge summaries issued from the French hospital discharge information systems (PMSI). Although their primary goals are essentially financial and managerial, these

databases make possible to explore and analyse patients' care trajectories [2]. Understanding and analysing these data are crucial for efficient healthcare planning and fair allocation of health care resources [3]. Moreover, care trajectory analysis can be an asset for epidemiology studies by providing statistical indicators to understand and explain care seeking behaviours [4]. Care trajectories' comparison, which is part of their analysis, relies (i) on their representation and (ii) on an adequate comparison method. If we consider medico-administrative data for composing care trajectories, such as diagnoses, medical procedures or drugs prescriptions, an intuitive way of representing it is to write it down as a sequence. However, such formalism could be too simplistic considering the complexity of a care trajectory. The main complexity could be that the temporal order between events is not always known or even meaningful, especially when they occur simultaneously or in a really short time range. The second challenge is to handle the complexity of the large alphabet of trajectories' components. They are medical concepts that belong to detailed taxonomies, and taking into account semantic similarities [5] between these codes could render the method more robust to small variations [6]. The goal of this article is to introduce a representation of care trajectories that better accounts for administrative data complexity and an associated similarity measure for comparing them.

2 Materials and Methods

2.1 Representing trajectories as sequences of sets

To take into account the uncertainty or simultaneity ignored with the simple sequence formalism, we proposed to group the events as unordered sets of events. Trajectories can then be seen as sequences of sets composed of simultaneous or related events. Similarity measures between sequences [7] could then be generalized to this formalism. To determine the overlapping part between two trajectories, we generalized the principle of the longest common subsequence (LCS) to this formalism.

2.2 Comparing Sequences of Sets

Longest Common Subsequence for Sequences of Sets

Definition 1: sequence of sets

A sequence of sets is a non-empty sequence composed of sets of elements.

Definition 2: size of a sequence of sets

Let $X = (x_1, x_2, \dots, x_m)$ be a sequence of sets. The size of a sequence of set is:

$$|X| = \sum_{i=1}^{m} |x_i| \quad (1)$$

Definition 3: subsequence of a sequence of sets

Given two sequences of sets $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, with $m \leq n$, X is a subsequence of Y if it exists the indexes $1 \leq j_1 < j_2 < \cdots < j_m \leq n$ such as $x_i \subseteq y_{j_i}$ is true for all $i = 1,2, \dots, m$.

Definition 4: longest common subsequence of two sequences of sets
Given two sequences of sets $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, Z is a LCS of X and Y if $|Z| \geq |Z'|$, for all other common subsequence Z' of X and Y.

Trajectories Structural Similarity. To get a similarity measure between sequences, it is intuitive and popular to normalize the size of the LCS by maximal size of the compared sequences.

Definition 5: similarity between trajectories
We define a similarity measure between X and Y sequences of sets as:

$$sim(X,Y) = \frac{|LCS(X,Y)|}{max(|X|,|Y|)} \quad (2)$$

Considering Semantic Similarity between Events. In order to take into account similarities between elements of two trajectories, we proposed a modification of the $|LCS(\mathrm{X},\mathrm{Y})|$ calculation. Instead of using intersection between sets to determine the common part of two trajectories, we used a similarity measure between sets, which requires a similarity between elements of these sets. These elements are issued from taxonomies, and several semantic similarities based on their hierarchical structures are conceivable [5]. This similarity allows us to compute similarity measure between two sets of concepts.

Definition 6: similarity measure between two sets of concepts
Given two sets of concepts $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$, we note C the set of all matchings between elements of X and Y, and $sem(,)$ a semantic similarity between concepts. The similarity measure between X and Y is defined as:

$$sim(X,Y) = \max_{c \in C} \sum_{(x,y) \in c} sem(x,y) \quad (3)$$

Due to this modification, the similarity measure is not anymore based on the longest common subsequence but on what we could call the longest similar subsequence. As for solving many algorithmic text problems, such as sequence alignment, both similarities can be computed using a dynamic programming algorithm [9].

2.3 Experimentations

We performed a retrospective analysis using the permanent sample of the SNIIRAM database (EGB). The EGB is a representative cross-sectional sample of the population covered by National Health Insurance. First, we extracted the hospital stay and reimbursed drug prescription information of the patients who underwent an ambulatory care surgery in 2012. All available medical codes, i.e. principal diagnoses, related diagnoses, associated diagnoses, clinical acts and drugs delivered in pharmacies were extracted from three months before to three months after the hospital stay to reconstruct the care trajectories. Next, to experiment the use of our method in performing cluster analysis, we selected three sub-groups of ambulatory surgeries, namely angioplasties, eye surgeries and breast surgeries, which constitute a sample of 287 patients. Elements of the care trajectories were drugs, diagnoses and medical acts, represented respectively by the Anatomical Therapeutic Chemical Classification System (ATC),

the International Statistical Classification of Diseases and Related Health Problems – 10th revision (ICD-10), and the Common Classification of Medical Procedures (CCAM). To compute semantic similarities between elements from a same classification, we used the Wu and Palmer's similarity [10] which is based on their hierarchical structure. We then computed equation (2) between each pair of the 287 trajectories, with and then without taking into account the semantic similarities. We performed a cluster analysis using the R software (version 3.1.1), with an ascending hierarchical classification and a Ward linkage. Three classes were identifed based on the highest drop of inertia between classes. Then we focused on intra-class and inter-class similarity. Because we worked on three sub-groups of patients, we had an *a priori* knowledge of the class a patient belongs to. We computed the ratio between the sum of patient's similarities with its own group and the sum of patient's similarities with the other groups, for both kind of similarities and for each patient.

3 Results

Before introducing semantic similarities in the method, the running time was twice faster when considering sequences of sets than sequences of atomic elements. It was no longer the case with their introduction, because it is more complicated to compute semantic similarities between sets of concepts than between atomic concepts.

Before evaluating the relevance of using semantic similarity, we ensured that a cluster analysis based on these similarities led to three distinct clusters associated to the three initial sub-groups. Only three patients were not classified in the correct cluster. Further analysis revealed that all three shared frequent comorbidities with patients from the other groups.

Because including semantic similarity to the care trajectory similarity measure could only increase the final similarity values, we focused on the ratios between intra-class and inter-class similarity. A statistical comparison has shown that they were significantly higher with the semantic similarities' introduction (Wilcoxon signed-rank test, p=0.002). Overall, with this enrichment, the similarity of a patient with the patients of its own group has thus more increased than its similarity with the patients of the other groups, which is desirable in a cluster analysis.

4 Discussion and Conclusion

Thanks to its expressivity, the formalism of sequence of sets is appropriate to represent care trajectories based on medico-administrative data, e.g. clinical acts, diagnosis and drug codes. We have proposed a method to compare two care trajectories written as sequences of sets, which relies on a generalization of the LCS problem, in order to identify the homologous parts of two patients' care trajectories. And to make this method less strict, more robust to small variations, we tried to take into account the possible similarity between the care trajectories' components.

Our next objective will be to study the potential of the method in a clinical context, specifically the hospital stays for an angioplasty, to see if the method could be useful

for predicting post hospital stay outcomes (e.g. rehospitalisation and adverse events), explaining disease conditions severity or a mode of taking in charge the patients (e.g. ambulatory or inpatient care), and discovering trends in the care consumptions.

We envision enriching the method by taking into account other kinds of similarity between events, such as delay between events or events' durations. It is also our objective to apply this method to patient and guideline comparison, to know the homologous part between a care trajectory and a guideline.

Funding

Doctoral fellowship funded by PEPS Research consortium, supported by Agence Nationale de Sécurité des Médicaments et produits de santé (ANSM).

References

1. Tuppin, P., de Roquefeuil, L., Weill, A., Ricordeau, P., Merlière, Y.: French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique. 58, 286–290 (2010).
2. Moulis, G., Lapeyre-Mestre, M., Palmaro, A., Pugnet, G., Montastruc, J.-L., Sailler, L.: French health insurance databases: What interest for medical research? Rev Med Interne. 36, 411–417 (2015).
3. Jay, N., Nuemi, G., Gadreau, M., Quantin, C.: A data mining approach for grouping and analyzing trajectories of care using claim data: the example of breast cancer. BMC Med Inform Decis Mak. 13, 130 (2013).
4. Le Meur, N., Gao, F., Bayat, S.: Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. BMC Health Serv Res. 15, 200 (2015).
5. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLoS Computational Biology. 5, e1000443 (2009).
6. Girardi, D., Wartner, S., Halmerbauer, G., Ehrenmüller, M., Kosorus, H., Dreiseitl, S.: Using concept hierarchies to improve calculation of patient similarity. Journal of Biomedical Informatics. 63, 66–73 (2016).
7. Studer, M., Ritschard, G., Ritschard, G., Ritschard, G.: A comparative review of sequence dissimilarity measures. LIVES Working Papers. 2014, 1–47 (2014).
8. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM. 18, 341–343 (1975).
9. Bellman, R.: The theory of dynamic programming. Bull. Amer. Math. Soc. 60, 503–515 (1954).
10. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. Presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics June 27 (1994).