Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220469

Data Element Mapping in the Data Privacy Era

Romain GRIFFIER[a,b,1], Sébastien COSSIN[a,b], François KONSCHELLE[a,b], Fleur MOUGIN[b] and Vianney JOUHET[a,b]
[a] *Bordeaux University Hospital, Public health, 33000 Bordeaux, France*
[b] *Bordeaux University, Inserm U1219, Bordeaux Population Health, ERIAS team, 33000 Bordeaux, France*

Abstract. Secondary use of health data is made difficult in part because of large semantic heterogeneity. Many efforts are being made to align local terminologies with international standards. With increasing concerns about data privacy, we focused here on the use of machine learning methods to align biological data elements using aggregated features that could be shared as open data. A 3-step methodology (features engineering, blocking strategy and supervised learning) was proposed. The first results, although modest, are encouraging for the future development of this approach.

Keywords. data element, mapping, machine learning, LOINC

1. Introduction

Health data produced in the context of care can be reused for many purposes (phenotyping, research, etc.): this is the field of secondary use of health data. However, this reuse is complex (large volumes of data, compartmentalized data, etc.). One of the difficulties lies in the heterogeneity of the representation of medical concepts, called semantic heterogeneity. Different approaches can be used to reduce this heterogeneity. In particular, many efforts are focused on the alignment of local terminologies to international standards [1], such as the Logical Observation Identifiers Names and Codes [2] (LOINC®) used in many countries to encode biological data.

The Bordeaux University Hospital has implemented a clinical data warehouse (CDW) based on the i2b2 technology. The CDW integrates the data of patients who came at least once to the hospital since 2010, representing more than 2 million patients, 13 million hospital admissions and 2 billion observations. The CDW contains biology data integrated from two different biology software (TD-Synergy® until 2018 and Glims® since then) whose data are mostly centralized in the hospital's computerized patient record (DxCare®) resulting in a total of three biology sources. Each of the biology source software has its own local terminology and, within a biology source, several data elements (i.e. codes in the local terminology) may encode the same concepts, resulting in a high degree of semantic heterogeneity. One of the sources is partially aligned to LOINC. Thus, mapping these biological sources to each other would result in mapping local codes to LOINC.

[1] Corresponding author, Romain GRIFFIER ; E-mail: romain.griffier@chu-bordeaux.fr

With the development of machine learning methods, and in the context of strengthening personal data protection with the General Data Protection Regulation[2] (GDPR), many studies have raised privacy and security issues in AI methods [3].

Here, we propose to study the machine learning alignment of clinical datas through the example of biology records. Moreover, we aim at evaluating the performances of machine learning methods using aggregated features, thereby limiting the risk of compromising data privacy.

2. Methods

The proposed alignment methodology consists of three successive steps (Figure 1): 1) the data element selection and feature extraction, 2) the blocking strategy, and 3) a supervised classification.

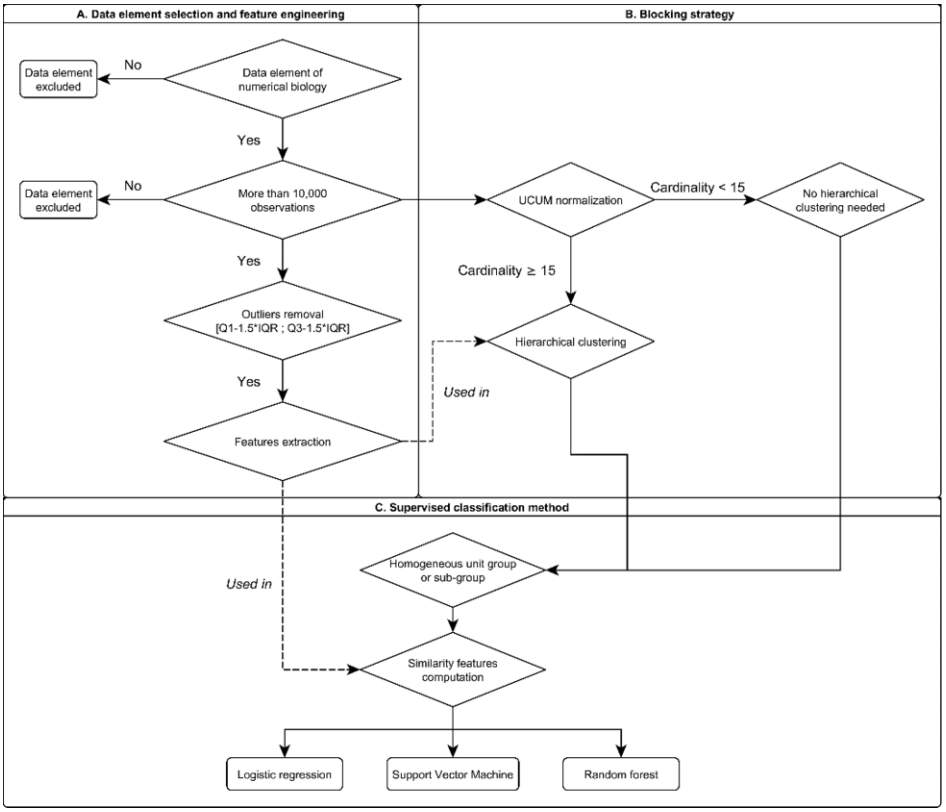

Figure 1. The 3-step alignment methodology

[2] https://gdpr.eu/

Data element selection and feature engineering. This first step consists in selecting the data elements[3] corresponding to the numerical biology with more than 10,000 observations available in the CDW (data-driven approach). For each selected data element, a data cleaning step with outlier removal was performed before computing the features. The following features were calculated: mean, standard deviation, median, quartiles, minimum, maximum, deciles, number of patients and stays, number of results by time of day (day or night), number of results above and below the norm. In addition, the sample distribution of each data element has been determined (using 1024 bins).

Blocking strategy. The final objective is to link similar identities (data elements) within the same data source and between several data sources. The number of possible comparisons is $\frac{n(n-1)}{2}$ where n is the sum of the cardinality of all data sources. To diminish the computational cost, it is necessary to have a blocking strategy [4] which limits the number of comparisons. The objective of this second step was therefore to constitute sub-groups of data elements in order to reduce the number of similarity features to be computed in step 3. This blocking strategy was based on:

1. The constitution of groups of homogeneous units on the basis of a standardization of units according to the Unified Code for Units of Measure[4] (UCUM) terminology.
2. Within the groups of data elements with a high cardinality (i.e. containing 15 or more data elements), an unsupervised clustering step using the hierarchical clustering (HC) method in order to form subgroups with lower cardinality.

Supervised classification. The third step was to compute the similarity features between all the data elements in each group resulting from the two previous steps. The following similarity features were calculated: difference in mean, difference in minimum, difference in maximum, difference in median, difference in quartile, difference in range and percent overlap of distributions. These similarity features were then used to train different supervised classification models: logistic regression, support vector machine (SVM) and random forest. The classification models were trained on a training sample (70%) and evaluated on a test sample (30%) using a gold standard of hand-crafted alignments by two experts in medical informatics (SC and RG).

Concerning the evaluation of the proposed method:

1. For the blocking strategy step, the evaluation only included gold standard concepts related to two or more data elements. Of these concepts, the percentage of those contained in a single homogeneous subgroup was assessed.
2. For the supervised classification method step, the evaluation was performed on the test set with recall, precision, F-measure, AUC and AUC_{PR}.

3. Results

Biological data integrated in the CDW represented 591,410,461 observations encoded in 170,933 data elements. The numerical biology corresponded to 475,117,464 observations (80.34%) encoded in 140,135 data elements (81.98%). After filtering out

[3] As defined by the ISO/IEC11179-3 standard. In our case, the data element corresponds to a code of a local terminology encoding a particular biological concept. Several data elements can encode the same concept (e.g. “HB001” and “HB002” are data elements that both correspond to the concept of Hemoglobin).

[4] https://ucum.org/ucum.html

Table 1. Results of the supervised classification models

	Threshold	Precision	Recall	F-measure	AUC_{PR}	AUC
Logistic regression	0.725	0.361	0.618	0.456	0.403	0.870
Support Vector Machine	0.710	0.391	0.476	0.430	0.393	0.800
Random forest	0.235	0.545	0.845	**0.663**	0.662	0.955

data elements with at least 10,000 observations available, 4,580 data elements (3.27%) remained representing 436,623,181 numerical observations (91.90%). Among these data elements, 1,421 (31.02%) were not associated with a unit (representing 89,465,534 observations). The others were associated with 153 different units.

After normalizing the units with UCUM, the data elements were grouped into 72 homogeneous unit groups (70 data elements could not be normalized with UCUM, representing 249,510 observations). The mean cardinality of these homogeneous unit groups was 44.29 data elements (sd=106.97). 42 homogeneous unit groups had cardinality less than or equal to 15 data elements and 30 homogeneous unit groups had cardinality greater than 15 data elements (mean=97.90; sd=151.32; maximum=709 for the percent unit).

The hierarchical clustering (HC) performed next on the 30 homogeneous unit groups with cardinality greater than or equal to 15 generated 277 clusters, with an average cardinality of 10.60 data elements (sd=30.94; maximum=336). Among the gold standard, considering biological concepts with at least 2 data elements, 95% of them were associated with data elements belonging to a single cluster.

Similarity features were computed in each of the 277 clusters (obtained by HC) and 42 groups of homogeneous unit groups with low cardinality, resulting in a total of 35,606 data element pairs. 3,756 (10.55%) of the data element pairs were associated with the same biological concept.

The results of the three supervised classification models are presented in Table 1. The best performing model was the random forest with an F-measure of 0.663, a recall of 0.845 and a precision of 0.545. Concerning the logistic regression and SVM models, the F-measures were respectively 0.456 and 0.430. The recall-precision curves for each model are presented in Figure 2 and found better overall performance for the random forest with AUC_{PR} of 0.662.

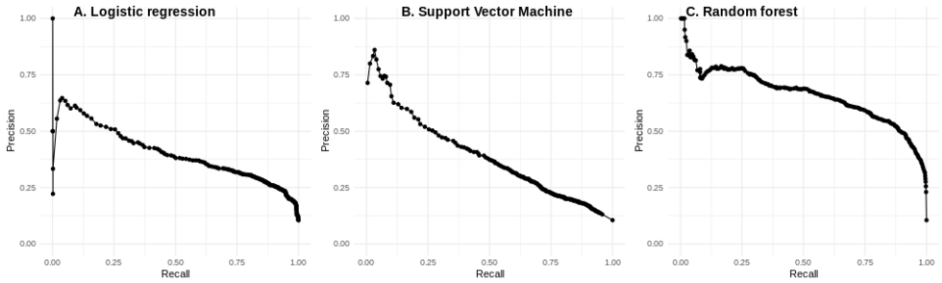

Figure 2. Recall-precision curves for the three classification methods

4. Discussion

We proposed a 3-step data-driven methodology to help achieve alignments between numerical biology data elements. The particularity of this work is the use of unsupervised

learning methods to implement a blocking strategy before training supervised classification models based on aggregated features that limit the privacy risk of re-identification based on trained algorithms.

Using HC to generate a blocking strategy reduced the cardinality of the homogeneous unit groups from 97.9 to 10.6 data elements without separating data elements of a same concept into different clusters. The performance of the HC was better than those obtained with the k-means method.

The supervised classification step yielded modest results. Using a random forest model gave the best results with an F-measure of 0.66 associated with a recall of 0.845 and a precision of 0.545 with a low threshold. The AUC was 0.955 in the context of a highly unbalanced data set (90/10). These results are slightly less good than those found in the literature [5], [6].

5. Conclusions

This preliminary work presents a 3-step method to align biological data elements using aggregated features that has obtained encouraging results. Further feature engineering work, including the addition of co-occurrence features, combined with semantic approaches, could optimize the performance of the proposed method, especially in the supervised learning step. An external validation step, using data from other healthcare institutions, will also be necessary to assess the generalizability of the method. Initiatives such as EHDEN[5] could provide a framework for implementing such an evaluation. Since this alignment method relies only on aggregated data at a very high level, sharing aggregate features related to LOINC concepts could help healthcare facilities to align their own local terminology with LOINC.

References

[1] Wade G, Rosenbloom ST. Experiences mapping a legacy interface terminology to SNOMED CT. BMC Med Inform Decis Mak. 2008 Oct;8 Suppl 1(Suppl 1):S3-3. doi: 10.1186/1472-6947-8-S1-S3.

[2] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. Clin Chem. 2003 04;49(4):624-33. doi: 10.1373/49.4.624.

[3] Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Med Ethics. 2021 Sep;22(1):122-2. doi: 10.1186/s12910-021-00687-3.

[4] Giang PH. A machine learning approach to create blocking criteria for record linkage. Health Care Manag Sci. 2015 Mar;18(1):93-105. doi: 10.1007/s10729-014-9276-0.

[5] Parr SK, Shotwell MS, Jeffery AD, Lasko TA, Matheny ME. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. J Am Med Inform Assoc. 2018 Oct;25(10):1292-300. doi: 10.1093/jamia/ocy110.

[6] Nikiema JN, Griffier R, Jouhet V, Mougin F. Aligning an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®). JAMIA Open. 2021 Jun;4(2):ooab035-5. doi: 10.1093/jamiaopen/ooab035.

[5] https://www.ehden.eu/