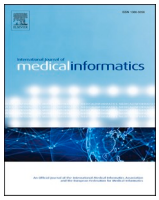
International Journal of Medical Informatics 179 (2023) 105233

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Augmented intelligence facilitates concept mapping across different electronic health records

Tariq A. Dam [a,b,*], Lucas M. Fleuren [a], Luca F. Roggeveen [a], Martijn Otten [a], Laurens Biesheuvel [a], Ameet R. Jagesar [a], Robbert C.A. Lalisang [b], Robert F.J. Kullberg [b], Tom Hendriks [b], Armand R.J. Girbes [a], Mark Hoogendoorn [c], Patrick J. Thoral [a], Paul W.G. Elbers [a], on behalf of The Dutch ICU Data Sharing Against COVID-19 Collaborators

[a] *Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science (AMDS), Amsterdam Public Health (APH), Amsterdam Cardiovascular Science (ACS), Amsterdam Institute for Infection and Immunity (AII), Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands*

[b] *Pacmed, Amsterdam, the Netherlands*

[c] *Quantitative Data Analytics Group, Department of Computer Science, Faculty of Science, Vrije Universiteit, Amsterdam, the Netherlands*

A R T I C L E I N F O

Keywords:
Concept mapping
Ontology
Ontology alignment
Augmented Intelligence
Intensive care

A B S T R A C T

Introduction: With the advent of artificial intelligence, the secondary use of routinely collected medical data from electronic healthcare records (EHR) has become increasingly popular. However, different EHR systems typically use different names for the same medical concepts. This obviously hampers scalable model development and subsequent clinical implementation for decision support. Therefore, converting original parameter names to a so-called ontology, a standardized set of predefined concepts, is necessary but time-consuming and labor-intensive. We therefore propose an augmented intelligence approach to facilitate ontology alignment by predicting correct concepts based on parameter names from raw electronic health record data exports.

Methods: We used the manually mapped parameter names from the multicenter "Dutch ICU data warehouse against COVID-19" sourced from three types of EHR systems to train machine learning models for concept mapping. Data from 29 intensive care units on 38,824 parameters mapped to 1,679 relevant and unique concepts and 38,069 parameters labeled as irrelevant were used for model development and validation. We used the Natural Language Toolkit (NLTK) to preprocess the parameter names based on WordNet cognitive synonyms transformed by term-frequency inverse document frequency (TF-IDF), yielding numeric features. We then trained linear classifiers using stochastic gradient descent for multi-class prediction. Finally, we fine-tuned these predictions using information on distributions of the data associated with each parameter name through similarity score and skewness comparisons.

Results: The initial model, trained using data from one hospital organization for each of three EHR systems, scored an overall top 1 precision of 0.744, recall of 0.771, and F1-score of 0.737 on a total of 58,804 parameters. Leave-one-hospital-out analysis returned an average top 1 recall of 0.680 for relevant parameters, which increased to 0.905 for the top 5 predictions. When reducing the training dataset to only include relevant parameters, top 1 recall was 0.811 and top 5 recall was 0.914 for relevant parameters. Performance improvement based on similarity score or skewness comparisons affected at most 5.23% of numeric parameters.

Conclusion: Augmented intelligence is a promising method to improve concept mapping of parameter names from raw electronic health record data exports. We propose a robust method for mapping data across various domains, facilitating the integration of diverse data sources. However, recall is not perfect, and therefore manual validation of mapping remains essential.

* Corresponding author at: Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science (AMDS), Amsterdam Public Health (APH), Amsterdam Cardiovascular Science (ACS), Amsterdam Institute for Infection and Immunity (AII), Amsterdam UMC, Vrije Universiteit, Amsterdam, the Netherlands.

E-mail addresses: t.dam@amsterdamumc.nl (T.A. Dam), l.fleuren@amsterdamumc.nl (L.M. Fleuren), l.roggeveen@amsterdamumc.nl (L.F. Roggeveen), m.otten1@amsterdamumc.nl (M. Otten), l.biesheuvel@amsterdamumc.nl (L. Biesheuvel), a.jagesar@amsterdamumc.nl (A.R. Jagesar), r.f.j.kullberg@amsterdamumc.nl (R.F.J. Kullberg), t.hendriks@tue.nl (T. Hendriks), arj.girbes@amsterdamumc.nl (A.R.J. Girbes), m.hoogendoorn@vu.nl (M. Hoogendoorn), p.thoral@amsterdamumc.nl (P.J. Thoral), p.elbers@amsterdamumc.nl (P.W.G. Elbers).

https://doi.org/10.1016/j.ijmedinf.2023.105233
Received 12 July 2023; Received in revised form 15 September 2023; Accepted 21 September 2023
Available online 22 September 2023

1. Introduction

With the advent of artificial intelligence, the secondary use of routinely collected medical data from electronic healthcare records (EHR) has become increasingly popular. Artificial intelligence techniques typically require vast amounts of data as well as external validation. Therefore, it is attractive to combine data from multiple hospitals, either through federated learning or by merging these data into a central database.

However, different hospitals use different implementations of different versions of different electronic health records systems. This typically implies that within these different systems, different names have been assigned for the same medical concepts. In addition, there may be differences in underlying data due to different methods of data collection or different units of measurement [1,2]. These challenges obviously hamper scalable model development and subsequent clinical implementation for decision support.

Medical ontologies such as LOINC, SNOMED, and ATC, and code systems such as UCUM are designed to overcome some of these issues by providing a standardized set of labels and relationships and can help improve data analysis through improving consistency and interoperability [2–6]. However, widespread implementation of a standardized labeling system across electronic health records is typically lacking.

During the COVID-19 pandemic, a national collaboration by Dutch Intensive Care Units led to the first nationwide data sharing initiative of highly granular clinical data from electronic health records. Incoming data from hospitals using one of three dominant EHR systems were combined into a single database with a unified parameter naming scheme for easy and consistent results when querying data and running analyses. This confirmed the absence of implementation of standard ontologies, except for medication records to a limited extent. Therefore, ontology alignment was necessary, ideally using international standard vocabularies, to map parameter names to predefined concepts.

However, no standard vocabulary proved to cover all the data elements that are specifically relevant to intensive care medicine. Therefore, a bespoke vocabulary was defined for this data sharing initiative, aiming to readily capture relevant clinical concepts to which incoming parameters could be manually mapped by a large team of healthcare professionals. The relevance of target concepts was based on intensive care medicine domain expertise as well as previous experience in clinical model development targeted at expeditiously facilitating research projects to answer urgent clinical questions during the pandemic [1]. This approach resulted in a vocabulary containing clinically relevant concepts, which was used to label parameters from participating hospitals sourced from three dominant EHR systems in the Netherlands: Epic (Epic Systems Corporation, Verona, USA), HiX (ChipSoft, Amsterdam, The Netherlands), and MetaVision (iMDsoft, Tel Aviv, Israel) [1].

Currently available methods for mapping hospital data to a target ontology include Regenstrief LOINC Mapping Assistant (RELMA), Usagi, and deep-learning based approaches [7–9]. These methods use text-based techniques to map hospital data to a target ontology. Additionally, Big-data Guided LOINC Mapping (BGLM) uses the underlying data distribution of measurements to map based on similarity in distribution [10]. However, although these are state-of-the-art approaches for ontology alignment of medical data, their application focused on specific domains, such as only laboratory parameters, with disappointing results.

To combine clinical data across domains, which is essential to intensive care medicine, combined with the need for a bespoke vocabulary for which no automated labeling was available, incoming hospital data was manually mapped. As expected, ontology alignment by manual mapping to a standardized list of concepts proved to be very labor-intensive, with an estimated time spent by a team of medical doctors and medical students of nearly 2000 h over 9 months.

Therefore, in order to facilitate data standardization and better facilitate teams to collaborate through data sharing initiatives, we set out to classify incoming hospital data with target labels using a machine learning algorithm. We hypothesized that this augmented intelligence approach, combining natural language processing with comparisons of underlying data distributions, would yield adequate concept mapping suggestions and thus could help reduce mapping workload.

2. Methods

The Medical Ethics Committee at Amsterdam UMC waived the need for patient informed consent and approved of an opt-out procedure for the collection of COVID-19 patient data during the COVID-19 crisis as documented under number 2020.156. This manuscript adheres to the ChAMAI checklist for assessment of medical AI. [11].

As previously reported [1], but repeated here to facilitate detailed understanding of our methods, the vocabulary of target concepts contained various elements for vital signs, ventilator records, hemodynamic monitoring, laboratory records, and medication records with two levels of hierarchy. Two independent clinical experts had been responsible for manual mapping while senior intensivists were consulted to resolve any conflicts or contact the sourcing hospital for clarification. In addition, data distribution plots had been manually reviewed to validate all mappings across hospitals [1].

The final vocabulary set consisted of 1,679 concepts, including 842 medication related items based on ATC and 837 other medical labels. This was used to manually map a total of 76,893 unique parameters that were collected and screened over the course of 6 months, sourced from 30 different hospitals all using one of three dominant EHR systems in the Netherlands. These EHR systems consisted of 6 Epic, 13 HiX, and 11 MetaVision implementations, none of which have implemented a preferential ontology system. As ontology systems were either not implemented or variably implemented, all hospital data was reviewed and mapped manually to ensure consistency in data availability and ontology alignment. Hospital and patient characteristics are reported in the supplementary material.

Of these parameters, 50.49% were mapped to the target concepts, with most records belonging to medication (15.28%) and laboratory values (11.88%). The other 49.51% of parameters were considered irrelevant and labeled “unmapped”. Parameter names were unique for 58,848 parameters (76.53%) across all hospitals.

Using the results of manual mapping, we created machine learning models to predict which concepts parameters should be mapped to. Our models were trained using data from three hospital systems only, covering all three electronic health record systems. These three hospital organizations comprised four large hospitals that were among the first to deliver data and consist of three academic and one general teaching hospital. We chose this approach to mimic the real-world situation in which data from different hospitals becomes available sequentially.

Data was split into a training and test set based on hospital names before any pre-processing steps were performed. Hospital parameter names underwent linguistic normalization through splitting by word based on common delimiters, tokenization, and lemmatization using Natural Language Toolkit (NLTK) WordNetLemmatizer, which groups

can be slow

words into sets of cognitive synonyms based on the lexical database WordNet. These synonyms are subsequently vectorized using term frequency - inverse document frequency (TF-IDF) [12–15]. This resulted in a sparse matrix of 9280 input features, where the occurrence of a word gains more weight if it is used only sparsely in the entire dataset. A stochastic gradient descent classifier was selected from scikit-learn for its efficiency in terms of training and prediction time, limited use of resources, and incremental training, allowing for real-time training and prediction [16]. The best hyper-parameter configuration was selected through a stratified 10-fold cross-validation based exhaustive grid search. The range of hyperparameters and full specification of final hyperparameters are available in Supplementary Table 1.

We predicted the probability of a parameter belonging to each of the concepts and collected the top 10 most probable labels with a probability greater than zero. We calculated performance metrics based on individual labels being included in the top × of the top 10 and grouped the results by relevance, data category, and EHR system. As sensitivity analyses, we removed exact matches on parameter names from the test set. We also performed a leave-one-out analysis where we test on a single hospital to simulate the performance of the model as the number of participating hospitals increased. By iterating over all hospitals, we set out to demonstrate robustness of the model and estimate the performance for new hospitals. As a large number of irrelevant parameters may introduce noise, and irrelevant parameters are usually easily recognized by their names upon manual validation, we retrained the leave-one-out analysis on relevant parameters only. Performance metrics are reported as mean scores over each individual hospital with 95% confidence intervals (95% CI), using bootstrapping with resampling.

In addition, to improve the order of returned labels for numeric parameters, we calculated a similarity statistic based on aggregated data. Available aggregated data included median, IQR, top 3 most frequent values, total number of records, total number of patients, and percentage of all patients in the hospital. The similarity statistic is based on the t-statistic but modified for use with medians and interquartile ranges (formula 1). Using this statistic, instead of hypothesis testing, we quantified the distance between two data distributions and altered the eligibility for predicted labels over various thresholds. As parameters were frequently recorded over 100,000 times, the denominator would become unimportant, resulting in an overrepresentation of differences in medians.

$$u = \frac{|Median_1 - Median_2|}{\sqrt{IQR_1^{\,2} + IQR_2^{\,2}}} \quad (1)$$

As we used medians and quartiles for describing underlying data distributions, we defined the reference group as the outer bounds of quartiles and the weighted mean of medians. The resulting similarity statistic was then used to exclude predictions which were labeled as dissimilar over various thresholds. Predictions for which no similarity could be calculated were not modified.

We also attempted to improve the returned results by deselecting results if the skewness of data was too dissimilar. Using the median, lower quartile and upper quartile, we calculated the skewness of data (formula 2).

$$\nu = \frac{Median - p25}{IQR} \quad (2)$$

Here, a result close to 0 indicates a right-skewed distribution, whereas a result close to 1 indicates a left-skewed distribution. This was used to assess the influence of dropping predictions when the difference in skewness was larger than various thresholds.

3. Results

The model trained on data from 3 hospital organizations, and tested on all other hospitals, yielded an overall precision of 0.733 (95% CI: 0.693–0.766), a recall of 0.751 (95% CI: 0.720–0.778), an F1-score of 0.720 (95% CI: 0.684–0.751), a balanced accuracy of 0.597 (0.534–0.650), and a top-5 accuracy of 0.889 (95% CI: 0.864–0.906) per hospital. When evaluating concepts present in both the training set and at least 5 testing hospitals, the mean balanced accuracy was 0.662 (95% CI: 0.597–0.710) on a total of 708 concepts.

When evaluated on parameters deemed as relevant for clinical research only, precision was 0.673 (95% CI: 0.592–0.742), recall was 0.581 (95% CI: 0.512–0.643), F1-score was 0.597 (95% CI: 0.525–0.660), balanced accuracy was 0.596 (95% CI: 0.533–0.649), and top-5 accuracy was 0.783 (95% CI: 0.727–0.827) per hospital. The overall scores are reported in Supplementary Table 2. For assessing whether the correct concept is in the top 5 of predictions, the cumulative recall was 0.828 for all relevant parameters combined (Fig. 1, Supplementary Table 3).

When excluding parameter names that matched the correct concepts exactly, precision remained 0.598 (95% CI: 0.526–0.664), recall was 0.487 (95% CI: 0.428–0.540), F1-score 0.507 (95% CI: 0.445–0.561), balanced accuracy was 0.472 (95% CI: 0.428–0.508), and top-5 accuracy was 0.724 (95% CI: 0.669–0.764), when evaluated on parameters eventually deemed as relevant (Supplementary Table 4). Performance per concept category and performance per EHR system are individually reported (Tables 1-2).

For the leave-one-hospital-out analysis, when evaluated for parameters deemed as relevant, precision was 0.716 (95% CI: 0.646–0.777), recall was 0.634 (95% CI: 0.572–0.688), F1-score was 0.649 (95% CI: 0.585–0.705), balanced accuracy was 0.640 (95% CI: 0.583–0.705), and top-5 accuracy was 0.879 (95% CI: 0.838–0.905) (Supplementary Table 5). For assessing whether the correct concept is in the top 5 of predictions, recall was 0.954 when evaluated for all incoming parameters combined, 0.905 for relevant parameters and 0.996 for irrelevant parameters. (Supplementary Figure 3). Performance across hospitals and parameter groups are available in the supplementary material (Supplementary Figure 2–3, Supplementary Tables 6-9).

A total of 5,411 parameters were labeled as numeric, with a total of 80,371 predicted concepts. Of these parameters, 73.61% was labeled as relevant for clinical research with most records belonging to laboratory records (26.6%), irrelevant records (26.39%), and respiratory records (17.32%). This subset showed an overall top 1 recall of 0.600, precision of 0.609 and F1-score of 0.562.

The similarity statistic and skewness difference were calculated for 37,458 out of 80,371 predictions. The top 1 prediction was altered for a maximum of 2,805 parameters which decreases as the thresholds for adjustment increase (Supplementary Figure 4). The largest increase in performance as compared to the baseline prediction was at a threshold of 5.4 where baseline recall of affected parameters was 0.190 and the recall of similarity adjusted predictions was 0.419 for a total of 63 numeric parameters (1.16%). The lowest threshold for improvement was 3.6 based on 134 parameters. (Supplementary Figure 4).

For skewness, the largest increase in performance was at a difference in skewness of 0.93 for a total of 76 numeric parameters (1.40%). The lowest skewness difference threshold for improvement was 0.45 based on 283 parameters (5.23%). (Supplementary Figure 5).

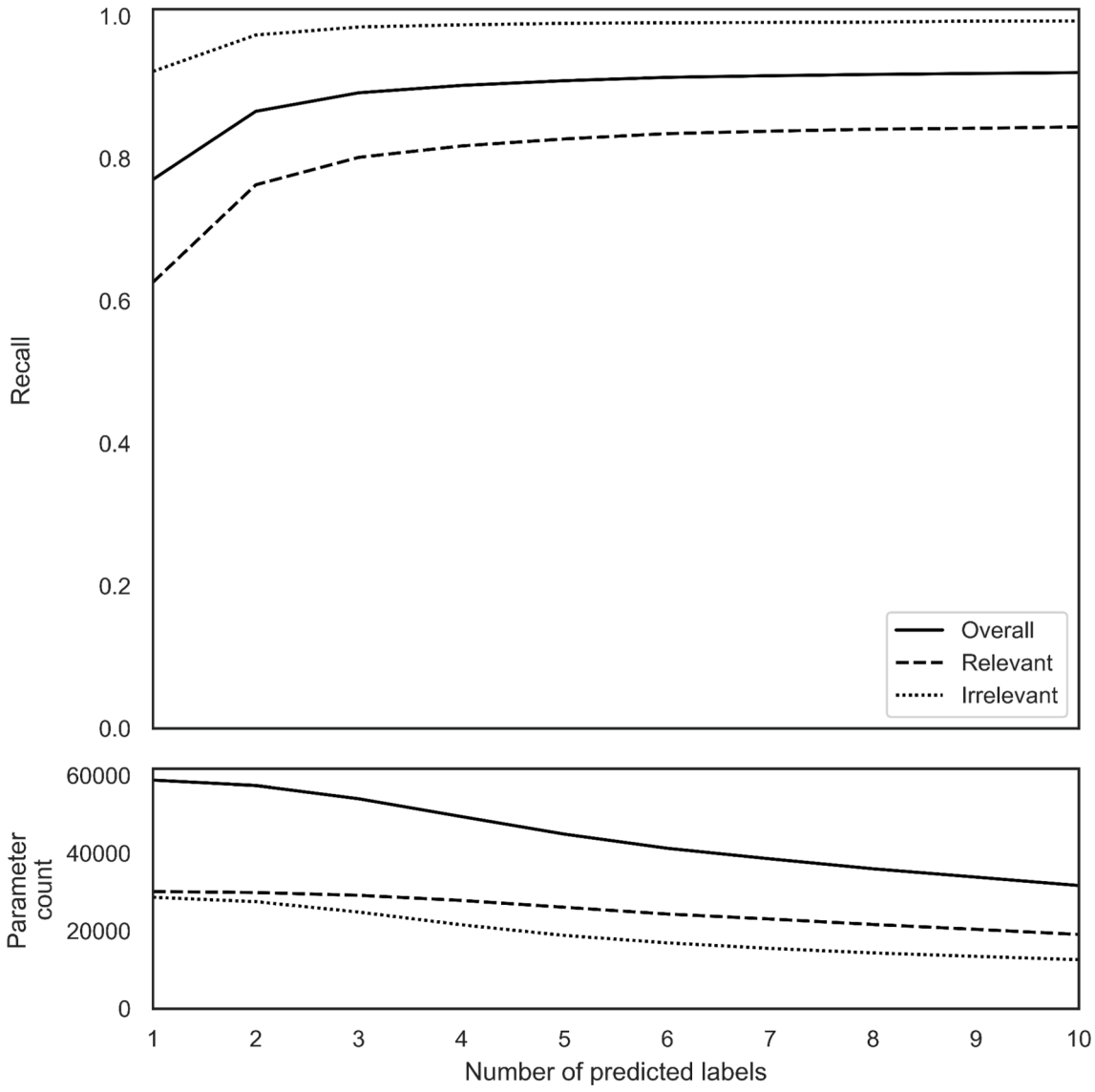


Fig. 1. Label-specific recall for all parameters predicted based on parameter names over increasing number of output labels, split by relevant and irrelevant parameters.

Table 1
Predictive performance, grouped per data category showing the number of underlying concepts, number of underlying parameters, total number of records and based on the label with the highest probability, the precision, recall and F1-score calculated as the weighted average on concept label performance before grouping. Table is sorted by the number of underlying parameters. Prediction model is based on parameter names of three hospital organizations and tested on all data within the testing hospitals. NICE = National Intensive Care Evaluation, SOFA = Sequential Organ Failure Assessment, LDA = Lines/Drains/Airway.

Concept Group	Concept Count	Parameter Count	Record Count	Precision	Recall	F1-score	Balanced Accuracy
unmapped	1	28,666	42,087,263	0.761	0.923	0.834	0.923
medication	748	8,961	1,402,372	0.838	0.854	0.837	0.570
laboratory value	324	6,209	3,415,632	0.577	0.359	0.393	0.248
respiratory	124	5,599	106,580,673	0.781	0.635	0.668	0.270
fluid balance	28	3,260	1,208,411	0.623	0.470	0.519	0.216
hemodynamics	65	2,957	43,958,076	0.856	0.784	0.811	0.472
neurology	14	656	642,602	0.661	0.543	0.582	0.523
infectiology	30	500	2,955,898	0.671	0.608	0.618	0.315
clinical score	30	499	1,616,114	0.722	0.637	0.642	0.504
renal replacement therapy	40	387	9,017,745	0.492	0.351	0.397	0.193
position	4	319	123,165	0.880	0.596	0.710	0.150
NICE data	59	277	54,042	0.246	0.126	0.153	0.094
demographics	11	269	886,330	0.658	0.532	0.558	0.453
LDA	7	102	2,204	0.521	0.627	0.560	0.264
SOFA score	27	82	132,789	0.438	0.500	0.441	0.486
admission information	9	61	16,559	0.060	0.066	0.061	0.081

Table 2
Predictive performance, grouped per EHR system and relevance, based on parameter names, trained on three starting hospital organizations, tested on all data within the testing hospitals. Precision, recall, F1-Score are reported as weighted means. EHR = Electronic Health Record, MV = MetaVision.

EHR	Relevance	Concept Count	Parameter Count	Record Count	Precision	Recall	F1-score	Balanced Accuracy
EPIC	**overall**	1,059	21,755	17,479,756	0.784	0.816	0.786	0.816
EPIC	**relevant**	1,058	6,369	5,797,728	0.616	0.491	0.513	0.491
EPIC	**irrelevant**	1	15,386	11,682,028	0.854	0.951	0.899	0.951
HIX	**overall**	1,094	25,484	15,718,853	0.773	0.752	0.723	0.752
HIX	**relevant**	1,093	19,478	14,206,473	0.837	0.705	0.732	0.705
HIX	**irrelevant**	1	6006	1,512,380	0.564	0.903	0.695	0.903
MV	**overall**	930	11,565	180,901,266	0.705	0.727	0.702	0.727
MV	**relevant**	929	4,291	152,008,411	0.547	0.470	0.473	0.407
MV	**irrelevant**	1	7,274	28,892,855	0.798	0.879	0.837	0.879

4. Discussion

This is the first paper to show that mapping parameters from different EHR systems from different hospitals to concepts using artificial intelligence to augment human intelligence is feasible, with the correct label being absent in the top 5 predictions for only 9.1% of all hospital parameters. Variation in performance both across and within EHR systems was minimal, lending credibility to the robustness of our approach (Tables 1-2, Supplementary Figure 2). In fact, training on data from one to two hospitals per EHR system already performs well, although performance does benefit from more data, with the correct label being absent in the top 5 predictions for only 4.5% of hospital parameters (Supplementary Figures 1, 3).

The marginal increase in performance by the use of data distribution for predicting labels is in line with previous literature [10]. The low performance on numeric data can, at least partially, be explained by variation between hospitals, which has been demonstrated to be significant in bed-side ventilator settings in patients with COVID-19 [17]. Furthermore, inconsistencies in data types of parameters complicate predictions based on data distributions. For example, parameters describing CVVH ultrafiltration measured every 15 min were mapped to the same concept as parameters with hourly measurements and the same type of mismatch occurs for fluid balance parameters which are time dependent but variably measured over time such as urine output. These inconsistencies can be handled by improving the mapping process through a more detailed standardized vocabulary such as LOINC. [3].

This is not the first attempt in automatic ontology alignment of hospital data. Earlier efforts mainly used text-based predictions and were limited to small subsets such as laboratory tests or microbiology records. [7,18–24]. Big data-guided LOINC mapping based on text of laboratory parameters showed a precision of 0.33, which was lower than Regenstrief LOINC Mapping Assistant's (RELMA) precision of 0.43 [7,10]. Our current approach outperforms these methods with a precision of 0.729 for relevant parameters and 0.744 on all parameters (Supplementary Table 2). However, this was lower than the precision of 0.87 in Russian laboratory terms which underwent careful preparation and correction of spelling mistakes and abbreviation errors. [25]. More advanced methods, such as deep-learning automated terminology mapping in OMOP-CDM, yielded slightly higher recall scores for the top-1 prediction than the base performance of Usagi (0.490 versus 0.376), which uses Levenshtein distance to calculate term similarity. However, both methods were again outperformed by our current approach (0.626 for relevant parameters, Fig. 1, Supplementary Table 3). [8,9].

Despite earlier works in the field of intelligent and automated mapping of radiology and laboratory records to international standards, hospital systems in the Netherlands still have not implemented a hospital-wide standardized labeling system. The current approach, however, shows that labeling can be domain-agnostic, providing a method for easily transforming the entirety of the EHR system to a standardized naming scheme. This should improve interoperability by facilitating external model validation and model implementation and is expected to contribute to machine learning models reaching higher technology levels of readiness [26,27].

When stratifying to illustrate robustness over the various groups of interest, the balanced accuracy needs to be interpreted carefully. As the balanced accuracy requires false positives and true negatives, stratifying to groups of interest will influence these counts. In the case of hospitals, as different hospitals use different levels of granularity, not all concept labels were used for each hospital. When stratifying for concept categories or hospitals, false positive labels given to records belonging to other groups will not be available for calculation of the balanced accuracy.

Our approach suffers from a number of limitations. First, our current classification is based on a locally defined set of target labels. However, this set of target labels is under revision to link to LOINC [3]. Performance is expected to remain similar as predicted entities do not change. This will be assessed during the mapping process for the current ICUdata collaboration of Dutch Intensive Care units sharing data on all intensive care patients [28,29]. Second, although the accuracy is generally high, in order to ensure data integrity, manual validation is still required. Third, classification is trained on Dutch hospital systems, where abbreviations may differ compared to English based systems. Fourth, access to source data was limited to summarized statistics before labeling and parameter context was not available, limiting the use of advanced techniques on underlying data distributions, metrics for co-occurrence based on time and context within the EHR system, or identifiers linking to commonalities in devices or standards. However, even without the use of these direct links, a high performance is still achieved based on parameter names. Finally, we have shown a proof-of-concept through the use of a linear classifier using stochastic gradient descent. As parameter names overlap partially for checklists rather than actual measurements, more complex multiclass classification models may improve performance. [30].

However, our approach elevates the present-day state-of-the-art in concept mapping through increased precision and recall, through acceptance of a diverse set of data sources instead of a singular data source, and our approach can be applied without any manual preprocessing of input data. With the current approach, mapping data to a common ontology will require less resources and adoption of a common ontology across hospitals will be more feasible.

5. Conclusions

Augmented intelligence is a promising method to improve concept mapping of parameter names from raw electronic health record data exports. We propose a robust method for mapping data across various domains, facilitating the integration of diverse data sources. However, recall is not perfect, and therefore manual validation of mapping remains essential. Adjustment based on data distribution might be possible, but only impacts a minimal proportion of relevant parameters.

Summary Table

What was already known on the topic:

1. Hospital data is stored using highly variable parameter names.
2. Data standardization is required to facilitate research across hospitals.
3. Concept mapping is challenging and labor-intensive.

What this study added to our knowledge:

1. Augmented intelligence models can improve concept mapping.
2. Use of data distribution has limited impact on model performance.
3. Adequate mapping is feasible without developing domain-specific algorithms.

Authors contributions

Tariq Dam contributed to conceptualization, data collection, processing, analysis and drafted the manuscript. Lucas Fleuren contributed to conceptualization, data collection, processing and analysis. Luca Roggeveen, Martijn Otten, Laurens Biesheuvel and Ameet Jagesar contributed to analysis and code review. Robbert Lalisang, Bob Kullberg, Tom Hendriks contributed to data collection and processing. All authors critically reviewed the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Dutch ICU Data Sharing Against COVID-19 Collaborators:

From collaborating hospitals having shared data:

Diederik Gommers, PhD, Department of Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands.

Olaf L. Cremer, PhD, Intensive Care, UMC Utrecht, Utrecht, The Netherlands.

Rob J. Bosman, ICU, OLVG, Amsterdam, The Netherlands.

Sander Rigter, MD, Department of Anesthesiology and Intensive Care, St. Antonius Hospital, Nieuwegein, The Netherlands.

Evert-Jan Wils, MD, PhD, Department of Intensive Care, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands.

Tim Frenzel, MD, PhD, Department of Intensive Care Medicine, Radboud University Medical Center, Nijmegen, The Netherlands.

Dave A. Dongelmans, MD, PhD, Department of Intensive Care Medicine, Amsterdam UMC, Amsterdam, The Netherlands.

Remko de Jong, MD, Intensive Care, Bovenij Ziekenhuis, Amsterdam, The Netherlands.

Marco A.A. Peters, MD, Intensive Care, Canisius Wilhelmina Ziekenhuis, Nijmegen, The Netherlands.

Marlijn J.A Kamps, MD, Intensive Care, Catharina Ziekenhuis Eindhoven, Eindhoven, The Netherlands.

Dharmanand Ramnarain, MD, Department of Intensive Care, ETZ Tilburg, Tilburg, The Netherlands.

Ralph Nowitzky, MD, Intensive Care, HagaZiekenhuis, Den Haag, The Netherlands.

Fleur G.C.A. Nooteboom, MD, Intensive Care, Laurentius Ziekenhuis, Roermond, The Netherlands.

Wouter de Ruijter, MD, PhD, Department of Intensive Care Medicine, Northwest Clinics, Alkmaar, The Netherlands.

Louise C. Urlings-Strop, MD, PhD, Intensive Care, Reinier de Graaf Gasthuis, Delft, The Netherlands.

Ellen G.M. Smit, MD, Intensive Care, Spaarne Gasthuis, Haarlem en Hoofddorp, The Netherlands.

D. Jannet Mehagnoul-Schipper, MD, PhD, Intensive Care, VieCuri Medisch Centrum, Venlo, The Netherlands.

Tom Dormans, MD, PhD, Intensive care, Zuyderland MC, Heerlen, The Netherlands.

Cornelis P.C. de Jager, MD, PhD, Department of Intensive Care, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands.

Stefaan H.A. Hendriks, MD, Intensive Care, Albert Schweitzerziekenhuis, Dordrecht, The Netherlands.

Sefanja Achterberg, MD, PhD, ICU, Haaglanden Medisch Centrum, Den Haag, The Netherlands.

Evelien Oostdijk, MD, PhD, ICU, Maasstad Ziekenhuis Rotterdam, Rotterdam, The Netherlands.

Auke C. Reidinga, MD, ICU, SEH, BWC, Martiniziekenhuis, Groningen, The Netherlands.

Barbara Festen-Spanjer, MD, Intensive Care, Ziekenhuis Gelderse Vallei, Ede, The Netherlands.

Gert B. Brunnekreef, MD, Department of Intensive Care, Ziekenhuisgroep Twente, Almelo, The Netherlands.

Alexander D. Cornet, MD, PhD, FRCP, Department of Intensive Care, Medisch Spectrum Twente, Enschede, The Netherlands.

Walter van den Tempel, MD, Department of Intensive Care, Ikazia Ziekenhuis Rotterdam, Rotterdam, The Netherlands.

Age D. Boelens, MD, Anesthesiology, Antonius Ziekenhuis Sneek, Sneek, The Netherlands.

Peter Koetsier, MD, Intensive Care, Medisch Centrum Leeuwarden, Leeuwarden, The Netherlands.

Judith Lens, MD, ICU, IJsselland Ziekenhuis, Capelle aan den IJssel, The Netherlands.

Harald J. Faber, MD, ICU, WZA, Assen, The Netherlands.

A. Karakus, MD, Department of Intensive Care, Diakonessenhuis Hospital, Utrecht, The Netherlands.

Robert Entjes, MD, Department of Intensive Care, Adrz, Goes, The Netherlands.

Paul de Jong, MD, Department of Anesthesia and Intensive Care, Slingeland Ziekenhuis, Doetinchem, The Netherlands.

Thijs C.D. Rettig, MD, PhD, Department of Anesthesiology, Intensive Care and Pain Medicine, Amphia Ziekenhuis, Breda, The Netherlands.

Sesmu Arbous, MD, PhD, Intensivist, LUMC, Leiden, The Netherlands.

Julia Koeter, MD, Intensive Care, Canisius Wilhelmina Ziekenhuis, Nijmegen, The Netherlands.

Roger van Rietschote, Business Intelligence, Haaglanden MC, Den Haag,The Netherlands.

M.C. Reuland, MD, Department of Intensive Care Medicine, Amsterdam UMC, Universiteit van Amsterdam, Amsterdam, The Netherlands.

Laura van Manen, MD, Department of Intensive Care, BovenIJ Ziekenhuis, Amsterdam, The Netherlands.

Leon Montenij, MD, PhD, Department of Anesthesiology, Pain Management and Intensive Care, Catharina Ziekenhuis Eindhoven, Eindhoven, The Netherlands.

Jasper van Bommel, MD, PhD, Department of Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands.

Roy van den Berg, Department of Intensive Care, ETZ Tilburg, Tilburg, The Netherlands.

Ellen van Geest, Department of ICMT, Haga Ziekenhuis, Den Haag, The Netherlands.

Anisa Hana, MD, PhD, Intensive Care, Laurentius Ziekenhuis, Roermond, The Netherlands.

B. van den Bogaard, MD, PhD, ICU, OLVG, Amsterdam, The Netherlands.

Prof. Peter Pickkers, Department of Intensive Care Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands.

Pim van der Heiden, MD, PhD, Intensive Care, Reinier de Graaf Gasthuis, Delft, The Netherlands.

Claudia (C.W.) van Gemeren, MD, Intensive Care, Spaarne Gasthuis, Haarlem en Hoofddorp, The Netherlands.

Arend Jan Meinders, MD, Department of Internal Medicine and Intensive Care, St Antonius Hospital, Nieuwegein, The Netherlands.

Martha de Bruin, MD, Department of Intensive Care, Franciscus Gasthuis & Vlietland, Rotterdam, The Netherlands.

Emma Rademaker, MD, MSc, Department of Intensive Care, UMC Utrecht, Utrecht, The Netherlands.

Frits H.M. van Osch, PhD, Department of Clinical Epidemiology, VieCuri Medisch Centrum, Venlo, The Netherlands.

Martijn D. de Kruif, MD, PhD, Department of Pulmonology, Zuyderland MC, Heerlen, The Netherlands.

Nicolas Schroten, MD, Intensive Care, Albert Schweitzerziekenhuis, Dordrecht, The Netherlands.

Klaas Sierk Arnold, MD, Anesthesiology, Antonius Ziekenhuis Sneek, Sneek, The Netherlands.

J.W. Fijen, MD, PhD, Department of Intensive Care, Diakonessenhuis Hospital, Utrecht, The Netherland.

Jacomar J.M. van Koesveld, MD, ICU, IJsselland Ziekenhuis, Capelle aan den IJssel, The Netherlands.

Koen S. Simons, MD, PhD, Department of Intensive Care, Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands.

Joost Labout, MD, PhD, ICU, Maasstad Ziekenhuis Rotterdam, The Netherlands.

Bart van de Gaauw, MD, Martini ziekenhuis, Groningen, The Netherlands.

Michael Kuiper, Intensive Care, Medisch Centrum Leeuwarden, Leeuwarden, The Netherlands.

Albertus Beishuizen, MD, PhD, Department of Intensive Care, Medisch Spectrum Twente, Enschede, The Netherlands.

Dennis Geutjes, Department of Information Technology, Slingeland Ziekenhuis, Doetinchem, The Netherlands.

Johan Lutisan, MD, ICU, WZA, Assen, The Netherlands.

Bart P. Grady, MD, PhD, Department of Intensive Care, Ziekenhuisgroep Twente, Almelo, The Netherlands.

Remko van den Akker, Intensive Care, Adrz, Goes, The Netherlands.

Tom A. Rijpstra, MD, Department of Anesthesiology, Intensive Care and Pain Medicine, Amphia Ziekenhuis, Breda, The Netherlands.

Roos Renckens, MD, PhD, Department of Internal Medicine, Northwest Clinics, Alkmaar, the Netherlands,

From collaborating hospitals having signed the data sharing agreement:

Daniël Pretorius, MD, Department of Intensive Care Medicine, Hospital St Jansdal, Harderwijk, The Netherlands.

Menno Beukema, MD, Department of Intensive Care, Streekziekenhuis Koningin Beatrix, Winterswijk, The Netherlands.

Bram Simons, MD, Intensive Care, Bravis Ziekenhuis, Bergen op Zoom en Roosendaal, The Netherlands.

A.A. Rijkeboer, MD, ICU, Flevoziekenhuis, Almere, The Netherlands.

Marcel Aries, MD, PhD, Department of Intensive Care, MUMC+, School of Mental Health and Neurosciences (MHENS), University Maastricht, Maastricht, The Netherlands.

Niels C. Gritters van den Oever, MD, Intensive Care, Treant Zorggroep, Emmen, The Netherlands.

Martijn van Tellingen, MD, EDIC, Department of Intensive Care Medicine, afdeling Intensive Care, ziekenhuis Tjongerschans, Heerenveen, The Netherlands.

Annemieke Dijkstra, MD, Department of Intensive Care Medicine, Het Van Weel-Bethesda Ziekenhuis, Dirksland, The Netherlands.

Rutger van Raalte, Department of Intensive Care, Tergooi hospital, Hilversum, The Netherlands.

From the Center for Critical Care Computational Intelligence (C4I):

Fuda van Diggelen, MSc, Quantitative Data Analytics Group, Department of Computer Science, Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands.

Ali el Hassouni, PhD, Quantitative Data Analytics Group, Department of Computer Science, Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands.

David Romero Guzman, PhD, Quantitative Data Analytics Group, Department of Computer Science, Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands.

Sandjai Bhulai, PhD, Analytics and Optimization Group, Department of Mathematics, Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands.

Dagmar M. Ouweneel, PhD, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Ronald Driessen, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Jan Peppink, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Harm-Jan de Grooth, MD, PhD, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

G.J. Zijlstra, MD, PhD, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

A.J. van Tienhoven, MD, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Evelien van der Heiden, MD, Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Jan Jaap Spijkstra, MD, PhD, Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Hans van der Spoel, MD, Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Angelique M.E. de Man, MD, PhD, Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Thomas Klausch, PhD, Department of Clinical Epidemiology, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

Heder J. de Vries, MD, Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.

From Pacmed:

Sebastiaan J.J. Vonk, MSc, Pacmed, Amsterdam, The Netherlands.

Willem E. Herter, BSc, Pacmed, Amsterdam, The Netherlands.

Michele Tonutti, MRes, Pacmed, Amsterdam, The Netherlands.

Daan P. de Bruin, MSc, Pacmed, Amsterdam, The Netherlands.

Mattia Fornasa, PhD, Pacmed, Amsterdam, The Netherlands.

Tomas Machado, Pacmed, Amsterdam, The Netherlands.

Michael de Neree tot Babberich, Pacmed, Amsterdam, The Netherlands.

Olivier Thijssens, MSc, Pacmed, Amsterdam, The Netherlands.

Lot Wagemakers, Pacmed, Amsterdam, The Netherlands.

Hilde G.A. van der Pol, Pacmed, Amsterdam, The Netherlands.

Julie Berend, Pacmed, Amsterdam, The Netherlands.

Virginia Ceni Silva, Pacmed, Amsterdam, The Netherlands.

Taco Houwert, MSc, Pacmed, Amsterdam, The Netherlands.

Hidde Hovenkamp, MSc, Pacmed, Amsterdam, The Netherlands.

Roberto Noorduijn Londono, MSc, Pacmed, Amsterdam, The Netherlands.

Davide Quintarelli, MSc, Pacmed, Amsterdam, The Netherlands.

Martijn G. Scholtemeijer, MD, Pacmed, Amsterdam, The Netherlands.

Aletta A. de Beer, MSc, Pacmed, Amsterdam, The Netherlands.
Giovanni Cinà, PhD, Pacmed, Amsterdam, The Netherlands.
Adam Izdebski, Pacmed, Amsterdam, The Netherlands.
From RCCnet:
Leo Heunks, MD, PhD, Department of Intensive Care Medicine, Amsterdam Medical Data Science, Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands.
Nicole Juffermans, MD, PhD, ICU, OLVG, Amsterdam, The Netherlands.
Arjen J.C. Slooter, MD, PhD, Department of Intensive Care Medicine, UMC Utrecht, Utrecht University, Utrecht, the Netherlands.
From other collaborating partners:
Martijn Beudel, MD, PhD, Department of Neurology, Amsterdam UMC, Universiteit van Amsterdam, Amsterdam, The Netherlands.

Tariq Dam contributed to conceptualization, data collection, processing, analysis and drafted the manuscript. Lucas Fleuren contributed to conceptualization, data collection, processing and analysis. Luca Roggeveen, Martijn Otten, Laurens Biesheuvel and Ameet Jagesar contributed to analysis and code review. Robbert Lalisang, Bob Kullberg, Tom Hendriks contributed to data collection and processing. All authors critically reviewed the manuscript.

Ethics approval and consent to participate.

The Medical Ethics Committee at Amsterdam UMC waived the need for patient informed consent and approved of an opt-out procedure for the collection of COVID-19 patient data during the COVID-19 crisis as documented under number 2020.156.

Funding.

Partially funded by the Netherlands Organization for Health Research and Development under project number 10430012010003.

Data and code availability.

All participating hospitals have access to the Dutch ICU Data Warehouse. External researchers can get access in collaboration with any of the participating hospitals. Contact details can be found on amsterdammedicaldatascience.nl. The code used for analysis is publicly available at github.com/tariqdam/AutoMap

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105233.

References

[1] L.M. Fleuren, T.A. Dam, M. Tonutti, et al., The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients [Internet]. Crit Care Lond Engl 2021; 25[cited 2021 Oct 18] Available from: https://pubmed.ncbi.nlm.nih.gov/34425864/.
[2] E.G. Phimister, M.A. Haendel, C.G. Chute, P.N. Robinson, Classification, Ontology, and Precision Medicine, The New England Journal of Medicine 379 (15) (2018) 1452–1462.
[3] A.W. Forrey, C.J. McDonald, G. DeMoor, et al., Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results, Clin Chem 42 (1996) 81–90.
[4] Overview of SNOMED CT [Internet]. [cited 2022 Nov 16] Available from: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html.
[5] WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs, 2022. Oslo, Norway 2021 [Internet]. Available from: https://www.whocc.no/atc_ddd_methodology/purpose_of_the_atc_ddd_system/.
[6] G. Schadow, C.J. McDonald, The Unified Code for Units of Measure [Internet]. Unified Code Units Meas 2017; [cited 2022 Nov 16] Available from: https://ucum.org/ucum.
[7] About RELMA [Internet]. *LOINC* [cited 2023 Mar 21] Available from: https://loinc.org/relma/.
[8] Usagi [Internet]. [cited 2023 Mar 2] Available from: http://ohdsi.github.io/Usagi/.
[9] B. Kang, J. Yoon, H.Y. Kim, et al., Deep-learning-based automated terminology mapping in OMOP-CDM, J Am Med Inform Assoc JAMIA 28 (2021) 1489–1496.
[10] K. Liu, M. Witteveen-Lane, B.S. Glicksberg, et al., BGLM: big data-guided LOINC mapping with multi-language support, JAMIA Open 5 (2022) ooac099.
[11] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies, International Journal of Medical Informatics 153 (2021).
[12] Steven Bird, Edward Loper, et al., Natural Language Processing with Python, O'Reilly Media Inc., 2009.
[13] TF–IDF [Internet]. In: Sammut C, Webb GI, editor(s). Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010. p. 986–987.[cited 2022 Nov 16] Available from: https://doi.org/10.1007/978-0-387-30164-8_832.
[14] C. Fellbaum, WordNet: An Electronic Lexical Database. [Internet]. Bradford Books; 1998. Available from: https://mitpress.mit.edu/9780262561167/.
[15] J. Euzenat, P. Shvaiko, Ontology Matching [Internet]. Berlin, Heidelberg: Springer; 2013. [cited 2023 Jun 7] Available from: https://link.springer.com/10.1007/978-3-642-38721-0.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[17] T.A. Dam, H.-J. de Grooth, T. Klausch, et al., Some Patients Are More Equal Than Others: Variation in Ventilator Settings for Coronavirus Disease 2019 Acute Respiratory Distress Syndrome, Crit Care Explor 3 (2021).
[18] L.M. Lau, K. Johnson, K. Monson, et al., A method for the automated mapping of laboratory results to LOINC, Proc AMIA Symp (2000) 472–476.
[19] D.J. Vreeman, C.J. McDonald, Automated Mapping of Local Radiology Terms to LOINC, American Medical Informatics Association Annual Symposium Proceedings 2005 (2005) 769–773.
[20] M. Fidahussein, D.J. Vreeman, A corpus-based approach for automated LOINC mapping, J Am Med Inform Assoc JAMIA 21 (1) (2014) 64–72.
[21] H. Kim, R. El-Kareh, A. Goel, FNU Vineet, W.W. Chapman, An Approach to Improve LOINC Mapping through Augmentation of Local Test Names, Journal of Biomedical Informatics 45 (4) (2012) 651–657.
[22] S.K. Parr, M.S. Shotwell, A.D. Jeffery, et al., Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database, J Am Med Inform Assoc JAMIA 25 (2018) 1292–1300.
[23] J. Kelly, C. Wang, J. Zhang, et al., Automated Mapping of Real-world Oncology Laboratory Data to LOINC, American Medical Informatics Association Annual Symposium Proceedings 2021 (2021) 611.
[24] C.-Y. Yeh, S.-J. Peng, H.C. Yang, et al., Logical Observation Identifiers Names and Codes (LOINC®) Applied to Microbiology: A National Laboratory Mapping Experience in Taiwan, Diagn Basel Switz 11 (2021) 1564.
[25] G. Kopanitsa, Application of a Regenstrief RELMA vol 6.6 to Map Russian Laboratory Terms to LOINC, Methods of Information in Medicine 55 (02) (2016) 177–181.
[26] L.M. Fleuren, P. Thoral, D. Shillan, A. Ercole, P.W.G. Elbers, Machine learning in intensive care medicine: ready for take-off? Intensive Care Medicine 46 (7) (2020) 1486–1488.
[27] D. van de Sande, M.E. van Genderen, J. Huiskens, D. Gommers, J. van Bommel, Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit, Intensive Care Medicine 47 (7) (2021) 750–760.
[28] P. Elbers, P. Thoral, T. Dam, et al., Sharing is caring: How covid-19 led to large-scale collaboration for icudata.nl. *Neth*, Journal of Critical Care 29 (2021) 85–86.
[29] icudata.nl - the Dutch ICU Data Warehouse [Internet]. [cited 2022 Nov 16] Available from: https://www.icudata.nl/.
[30] 1.12. Multiclass and multioutput algorithms [Internet]. Scikit-Learn [cited 2022 Nov 16] Available from: https://scikit-learn/stable/modules/multiclass.html.