Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230241

Machine Learning for Medical Data Integration

Armin MÜLLER [a,1], Lara-Sophie CHRISTMANN [b], Severin KOHLER [b], Roland EILS [b] and Fabian PRASSER [a]

[a] *Center of Health Data Science, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany*
[b] *Digital Health Center, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany*

ORCiD ID: Armin Müller https://orcid.org/0000-0003-3566-8687, Lara-Sophie Christmann https://orcid.org/0000-0001-6453-8633, Severin Kohler https://orcid.org/0000-0002-7718-6187, Roland Eils https://orcid.org/0000-0002-0034-4036, Fabian Prasser https://orcid.org/0000-0003-3172-3095

Abstract. Making health data available for secondary use enables innovative data-driven medical research. Since modern machine learning (ML) methods and precision medicine require extensive amounts of data covering most of the standard and edge cases, it is essential to initially acquire large datasets. This can typically only be achieved by integrating different datasets from various sources and sharing data across sites. To obtain a unified dataset from heterogeneous sources, standard representations and Common Data Models (CDM) are needed. The process of mapping data into these standardized representations is usually very tedious and requires many manual configuration and refinement steps. A potential way to reduce these efforts is to use ML methods not only for data analysis, but also for the integration of health data on the syntactic, structural, and semantic level. However, research on ML-based medical data integration is still in its infancy. In this article, we describe the current state of the literature and present selected methods that appear to have a particularly high potential to improve medical data integration. Moreover, we discuss open issues and possible future research directions.

Keywords. medical data integration, common data models, machine learning

1. Introduction

Insights generated through data-driven medical research methods based on the secondary use of health data have the potential to make future medicine more predictive, preventive and personalized [1]. This can improve cost efficiency and foster adoption to demographic change. To make health data from a diverse set of sources available for research and enable real-world evidence generation, they need to be integrated into common data models that facilitate comparability. Well-known models in the health field in-

[1]Corresponding Author: Armin Müller, Center of Health Data Science, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany; E-mail: armin.mueller@charite.de

clude the OMOP Common Data Model [2], HL7 Fast Healthcare Interoperability Resources [3] and openEHR [4]. When data was not collected in accordance to such models at source – which is common to date – it needs to be harmonized and transformed.

Due to the large number of autonomous information systems within typical health IT infrastructures, data is usually heterogeneous along three axes: (1) syntax (e.g. regarding the meaning of symbols), (2) structure (e.g. regarding the organization of properties of health-related data entities) and (3) semantics (e.g. regarding terminologies as well as codes and their meaning).

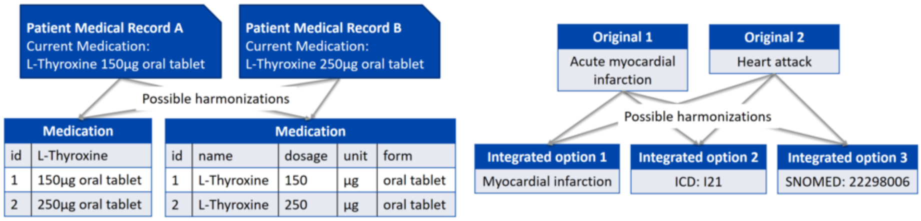

Figure 1. Example for structural heterogeneity (left) and semantic heterogeneity (right).

An example for syntactical heterogeneity are the different possible encodings of the character "μ" in the drug prescription "L-Thyroxine 150 μg oral tablet", which can be encoded as "\u03BC" in Unicode, as "230" in ASCII, and as "μ" in HTML. Figure 1 illustrates a simple example of structural and semantic heterogeneity. On the left, the same information about the administration of a drug is structured in different ways. The right side demonstrates that multiple terms can refer to the same concept.

To date, harmonization is mostly achieved by using manually specified rules and algorithms in so-called Extract-Transform-Load (ETL) processes, supported by tools such as Pentaho Data Integration [5] or Talend Open Studio [6]. While innovative approaches, e.g. based on declarative specifications of target representations [7], can help to reduce some efforts, this process is usually still very time and resource consuming [8].

ML and Artificial Intelligence (AI) technologies are one of the core tools of data-driven medical research. The general idea is that instead of providing computers with rules to follow, they extract knowledge and discover rules themselves from training data provided [9]. For these algorithms and models to produce reasonable results, they rely heavily on large datasets, clearly demonstrating the need for medical data integration. With their predictive and generative capabilities, ML methods can also potentially be a powerful tool for data integration itself. In recent years, machine learning has already been applied very successfully to various knowledge extraction and standardization tasks. One important field is natural language processing, where ML has been very successful at understanding the structure of clinical documents [10] and extracting medical concepts from clinical free-text reports [11].

2. Objective

The aim of this paper is to investigate the potential of ML for medical data integration tasks and to provide a concise overview of the current state of the field. We focus on methods suited for integrating structured health data into standardized data models. More specifically, this paper presents examples of methods that have been suggested for ML-based harmonization of data on the syntactic, structural and semantic level. We further discuss their potential value for medical data integration tasks and highlight limitations as well as open research questions.

3. Method

Early papers from the data integration community have suggested that ML could be used to automate or support many of the tasks needed to harmonize and integrate structured data. Starting from a seminal paper by Dong and Rekatsinas, which, to our knowledge, is the first – and one of the only ones so far – to address the potential of ML for data integration tasks [12], we reviewed the state of the literature and present selected highlights.

We screened all of the papers citing the work by Dong and Rekatsinas [12] ($n = 117$ in December 2022). We excluded nine non-English papers ($n = 108$) and then selected all papers whose titles indicate a possible focus on data integration ($n = 78$). From this, we selected all papers with abstracts suggesting that they address structured tabular data or health data. We also excluded reviews as they only covered narrow aspects of the general topic relevant to our work. This resulted in $n = 22$ papers of which 14 addressed the topic of entity resolution (sometimes also called entity matching), which is an important aspect of semantic integration that seems to have received quite some scientific attention.

In the following section, we present in more detail selected papers from the body of literature identified in the described search process as well as selected papers discovered during a preparatory exploration of the field.

4. Results

4.1. Syntactic Heterogeneity

While it is already difficult to automatically detect the character set with which files are encoded [17], determining the composition and types of data items is even more challenging. One aspect that is particularly relevant for structured data integration, focuses on extracting the orientation and sub-components of tables, which is an important first step in any transformation process. Recently, Habibi et al. proposed a deep learning method for classifying table orientation, achieving an F_1-score of 76% [18]. Other relevant systems include MIT's Sherlock, which can detect data types (e.g. dates) in structured data with an F_1-score of 89% using deep neural networks [19].

4.2. Structural Heterogeneity

Just recently, Sahay et al. studied self-organizing maps combined with a priori knowledge about the target structure to conquer structural heterogeneity, achieving an F_1-score of 71% [20]. Anderson inspected column embeddings as input into a bidirectional LSTM model to label columns and tilt tables [13]. Both approaches are applicable to target data models with a specific pre-defined structure, such as the OMOP CDM. Toutanova et al. used neural nets to extract facts in the form of (subject, predicate, object)-triples, outperforming prior work in precision [21]. This is suited for mapping to generic models based on fact-tables, such as Informatics for Integrating Biology & the Bedside (i2b2) [22].

4.3. Semantic Heterogeneity

To tackle semantic heterogeneity, Kate used support vector machines to map terms from clinical narratives to SNOMED CT codes [23], achieving an F1-score of 88%. Wang et al. proposed using contrastive representation learning to facilitate multiple data integration tasks including entity and column matching [16]. Parr et al. focused on lab values and LOINC codes using logistic regression and a random forest multiclass classifier [24], achieving an F1-score of up to 62%, while Mirzaei et al. compared a logistic regression classifier, a random forest classifier and a fully connected neural network classifier to standardize variables within and across datasets [14]. Recently, Zhang et al. have approached the problem of detecting the semantic type of data items using a deep neural network for single column predictions. The results can then be forwarded into a multi-layer structured prediction model that outputs the final classification per column [25].

5. Discussion

In this paper we presented a concise overview of selected ML-based methods supporting core steps of medical data integration. In theory, a combination of such individual methods could be used to develop end-to-end ML-based data integration processes. In practice, however, several challenges have to be overcome to make this vision a reality.

First, most of the current solutions provide a performance that is not sufficient to reliably support data integration processes without a lot of manual intervention (cf. the F_1-scores presented in Section 4). Hence, further research on improved methods and human-in-the-loop approaches is needed. For example, Graph Neural Networks seem promising to improve the accuracy of structural data integration steps [15]. Since health data is inherently different from many other data domains - e.g., due to it being longitudinal and sometimes of low quality - the applicability of methods developed for non-health data remains to be evaluated.

Second, in addition to novel methods, also more comprehensive training and test sets are needed. Given the amount of work that has already went into medical data integration on a global scale, we are confident that large enough sets of matching original and harmonized data could be created. Nonetheless, they would also need to be curated and sharing them will likely pose privacy risks. An additional option would be to integrate knowledge about standardized data representations, e.g. from openEHR, FHIR or the OMOP CDM, into the ML models.

Finally, approaches are needed to introduce ML-based methods incrementally and in synergy with more traditional processes, e.g. as atomic operators in common data integration platforms. Considering the vast amount of heterogeneity in health data, we believe that this would be a significant step for advancing medical research.

In future work, it would be interesting to benchmark combinations of ML-based integration approaches along multiple axes and to compare their performance in terms of their integration accuracy and scalability. Furthermore, the work presented in this paper only provides a first overview of the topic. In the future, we aim to build upon this work by performing a more in-depth structured review on ML-based medical data integration.

References

[1] Hood L, Balling R, Auffray C. Revolutionizing Medicine in the 21st Century Through Systems Approaches. Biotechnol J. 2012 Aug;7(8):992-1001. doi: 10.1002/biot.201100306. Epub 2012 Jul 20.

[2] Voss EA, Makadia R, Matcho A, et al. Feasibility and Utility of Applications of the Common Data Model to Multiple, Disparate Observational Health Databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. doi: 10.1093/jamia/ocu023. Epub 2015 Feb 10. PMID: 25670757; PMCID: PMC4457111.

[3] Bender D, Sartipi K. HL7 FHIR: An Agile and RESTful Approach to Healthcare Information Exchange. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013 Jun; pp. 326-331, doi: 10.1109/CBMS.2013.6627810.
[4] Kalra D et al. The openEHR Foundation. Stud Health Tech Inform. 2005;115:153-73. PMID: 16160223.
[5] Casters M, Bouman R, van Dongen J. Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration. Indianapolis, IN: Wiley, Inc; 2010. 720p. ISBN: 9780470635179.
[6] Talend Open Studio: Open-source ETL and Free Data Integration. Talend. Available at: https://www.talend.com/products/talend-open-studio/ (Accessed: December 12, 2022).
[7] Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn K, Prasser F. Enabling Agile Clinical and Translational Data Warehousing: Platform Development and Evaluation. JMIR Med Inform 2020;8(7):e15918. doi: 10.2196/15918.
[8] Kimball R, Ross M. The Data Warehouse Toolkit: The definitive guide to dimensional modeling. Indianapolis, IN: John Wiley & Sons, Inc.; 2013. 608p. ISBN: 9781118530801.
[9] Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. 2016 Sep 29;375(13):1216-9. doi: 10.1056/NEJMp1606181. PMID: 27682033.
[10] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, Component Evaluation and Applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13. doi: 10.1136/jamia.2009.001560. PMID: 20819853.
[11] Kholghi M, Sitbon L, Zuccon G, Nguyen A. Active Learning: A Step Towards Automating Medical Concept Extraction. J Am Med Inform Assoc. 2016 Mar;23(2):289-96. doi: 10.1093/jamia/ocv069.
[12] Dong XL, Rekatsinas T. Data Integration and Machine Learning: A Natural Synergy. Proc. 2018 Int. Conf. Management of Data. 2018 May. pp. 1645–1650. doi: 10.1145/3183713.3197387.
[13] Anderson MR. It's Data All the Way Down: Exploring the Relationship Between Machine Learning and Data Management [dissertation]. [Michigan]: University of Michigan; 2019. 124p.
[14] Mirzaei A, Aslani P, Schneider CR. Healthcare Data Integration Using Machine Learning: A Case Study Evaluation with Health Information-seeking Behavior Databases. Res Social Adm Pharm. 2022 Dec;18(12):4144-4149. doi: 10.1016/j.sapharm.2022.08.001. Epub 2022 Aug 8. PMID: 35965198.
[15] Krivosheev E, Atzeni M, Mirylenka K, Scotton P, Casati F. Siamese Graph Neural Networks for Data Integration. eprint arXiv. 2020 Jan. doi: 10.48550/arXiv.2001.06543.
[16] Wang R, Li Y, Wang J. Sudowoodo: Contrastive Self-supervised Learning for Multi-purpose Data Integration and Preparation. arXiv preprint. 2022 Oct. doi: 10.48550/arXiv.2207.04122
[17] Kikui GI. Identifying, the Coding System and Language, of On-line Documents on the Internet. Proc. 16th Conf. Comp. Linguistics. 1996 Aug. pp. 652–657. doi: 10.3115/993268.993282.
[18] Habibi M, Starlinger J, Leser U. DeepTable: A Permutation Invariant Neural Network for Table Orientation Classification. Data Mining and Knowledge Discovery. 2020 Sep;34(6):1963-83. doi: 10.1007/s10618-020-00711-x.
[19] Hulsebos M, Hu K, Bakker M, et al. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019 Jul. pp. 1500-1508. doi: 10.1145/3292500.3330993.
[20] Sahay T, Mehta A, Jadon S. Schema Matching Using Machine Learning. 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN). 2020 Apr. doi: 10.1109/SPIN48934.2020.9071272.
[21] Toutanova K, Chen D, Pantel P, Poon H, Choudhury P, Gamon M. Representing Text for Joint Embedding of Text and Knowledge Bases. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015 Sep. doi: 10.18653/v1/D15-1174.
[22] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. doi: 10.1136/jamia.2009.000893. PMID: 20190053.
[23] Kate RJ. Towards Converting Clinical Phrases into SNOMED CT Expressions. Biomed Inform Insights. 2013 Jun 24;6(Suppl 1):29-37. doi: 10.4137/BII.S11645. PMID: 23847425; PMCID: PMC3702194.
[24] 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 1;25(10):1292-1300. doi: 10.1093/jamia/ocy110. PMID: 30137378.
[25] Zhang D, Hulsebos M, Suhara Y, Demiralp Ç, Li J, Tan W-C. Sato: Contextual Semantic Type Detection in Tables. Proceedings of the VLDB Endowment. 2020 Jun;13(12):1835–1848. doi: 10.14778/3407790.3407793.