EHR-QC: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes

Yashpal Ramakrishnaiah[a], Nenad Macesic[a,c], Anton Y. Peleg[a,c,*] and Sonika Tyagi[a,b,*]

[a]Department of Infectious Diseases, The Alfred Hospital and Central Clinical School, Monash University, Melbourne 3000, VIC, Australia
[b]School of Computing Technologies, RMIT University, Melbourne 3000, VIC, Australia
[c]Centre to Impact AMR, Monash University, Melbourne 3000, VIC, Australia

ARTICLE INFO

Keywords:
Digital Health
Electronic Health Records
EHR
Clinical Outcome Prediction
Machine Learning

ABSTRACT

The adoption of electronic health records (EHRs) has created opportunities to analyze historical data for predicting clinical outcomes and improving patient care. However, non-standardized data representations and anomalies pose major challenges to the use of EHRs in digital health research. To address these challenges, we have developed EHR-QC, a tool comprising two modules: the data standardization module and the preprocessing module. The data standardization module migrates source EHR data to a standard format using advanced concept mapping techniques, surpassing expert curation in benchmarking analysis. The preprocessing module includes several functions designed specifically to handle healthcare data subtleties. We provide automated detection of data anomalies and solutions to handle those anomalies. We believe that the development and adoption of tools like EHR-QC is critical for advancing digital health. Our ultimate goal is to accelerate clinical research by enabling rapid experimentation with data-driven observational research to generate robust, generalisable biomedical knowledge.

Highlights

- EHR-QC accepts EHR data from a relational database or as a flat file and provide an easy-to-use, customized, and comprehensive solution for data handling activities.
- It offers a modular standardization pipeline that can convert any EHR data to a standardized data model i.e. OMOP-CDM.
- It includes an innovative algorithmic solution for clinical concept mapping that surpasses the current expert curation process.
- We have demonstrated that the imputation performance depends on the nature and missing proportion, hence as part of EHR-QC we included a method that searches for the best imputation method for the given data.
- It also contains an end-to-end solution to handle other anomalies such as outliers, errors, and other inconsistencies in the EHR data.

List of Figures

*Corresponding author
anton.peleg@monash.edu (A.Y. Peleg); sonika.tyagi@rmit.edu.au (S. Tyagi)
ORCID(s): 0000-0002-2213-8348 (Y. Ramakrishnaiah); 0000-0002-7905-628X (N. Macesic); 0000-0002-2296-2126 (A.Y. Peleg); 0000-0003-0181-6258 (S. Tyagi)

List of Tables

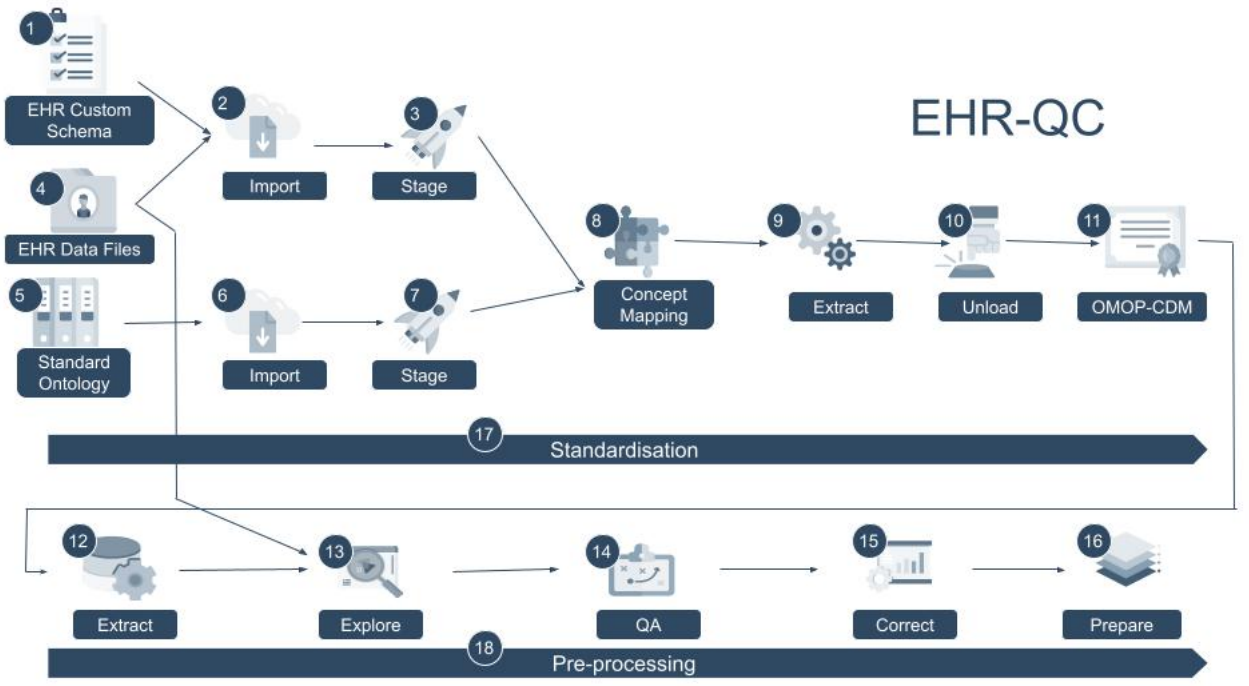


Figure 1: EHR-QC architecture diagram showing the standardisation and the preprocessing modules. The EHR data can be converted to OMOP-CDM standard using the standardisation module. This module imports and stages the data in its native format along with the reference vocabularies in an intermediate schema upon which the data is transformed into the standard entities. An important step in the process is to map the non-standard terms in the source data to a concept from the standard ontologies for which the pipeline provides an automated solution. This pipeline unloads the standardised data to the OMOP CDM schema as the final step. The EHR-QC also contains the preprocessing pipeline to assist with exploratory data analysis and quality assurance (QA) of the EHR data.

1. Introduction

Electronic Health Records (EHR) have been widely adopted and contain an incredible wealth of digital health information including demographics, observations, investigations, diagnoses, treatments, procedures, and clinical notes. This has allowed EHR data to be used for many purposes, such as in public health surveillance [1, 2], disease modelling [3], predictive analytics [4, 5, 6], assessment of medical treatments and procedures [7], decision making [8] and policy development [9], and data-driven research [10, 11, 12]. These applications are possible because of the increased adoption of EHRs coupled with the emergence of data-driven machine-learning techniques that facilitate the ability to leverage large amounts of data to uncover hidden knowledge. However, there are significant limitations in use of EHR due to non-standardisation and inherent biases in the data [13, 14, 15, 16, 17].

One of the major challenges in conducting research using EHR data is the presence of anomalies, such as missing data, outliers, errors, and inconsistencies [18]. EHR data comprises various data types collected from different systems, some of which are obtained directly from monitoring devices, while others are entered manually [19]. Additionally, as the data is collected primarily for administrative purposes, it may not undergo the same level of rigorous vetting as manually collected research data leading to poor research outcomes when used unprocessed [20]. Even seemingly insignificant errors can have severe consequences, as evidenced by a study in which children whose weights were wrongly recorded resulted in drug overdose in one in three cases [21]. To handle these anomalies, several domain-specific frameworks [14, 20] and tools such as Achilles [22], DataQualityDashboard (DQD) [14], ARES, MIRACUM [23], mosaicQA [24], and Mind the gap [25] have been developed. However, many of these solutions are limited to a specific source format or scope, and do not offer a means to address identified anomalies. Therefore, it is necessary to create more effective quality control frameworks.

Furthermore, the lack of standardization of EHR data poses a significant challenge, especially with respect to data formatting, clinical terminology, analytical methods, and procedures. In our study, we prioritize two key aspects of standardization: data format and clinical terminology. Standardizing data format allows for seamless integration and analysis across different sources, while harmonizing clinical terminology facilitates accurate interpretation and comparison of findings [26, 27]. Conversely, non-standard data representation can negatively impact the adaptability of tools and methods to various sources leading to poor generalization of results, duplicated efforts, and laborious downstream processing [28, 29].

The EHR is typically stored in institution-specific databases, each with its own data representation format, also known as a "schema". Standardizing the database schema involves maintaining a consistent data representation that are interoperable. To establish a standardized representation of EHR data, open-source common data models (CDMs) have been developed. The Observational Medical Outcomes Partnership-Common Data Models (OMOP-CDM), created by the OHDSI consortium, is one such example. Efforts are underway to convert custom schemas to the OMOP-CDM standard while preserving its content [30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. This conversion allows for consistent data analysis and integration across different healthcare systems, facilitating collaborative research and improving interoperability [40] [41, 42].

However, the scope of these efforts is limited since they are tailored to a specific EHR format and cannot be repurposed. To address this issue, a few generic utilities have been developed such as the user-interface-based "White Rabbit", and "Rabbit In A Hat" and purely code-based “Extract, transform, load (ETL) framework” [43] to facilitate the OMOP conversion. The user-interface-centric design of these tools makes them incompatible with version control systems, resulting in a need for manual intervention, scalability challenges, replication difficulties, and a higher risk of errors. In contrast, the code-based approach is less efficient, and its performance may not be sufficient for time-sensitive applications, particularly when dealing with large datasets.

The second challenge to EHR data standardisation is the use of free-text to document and store medical concept information such as clinical findings, procedures, and outcomes. To address this, clinical terminology standardization involves mapping clinical concepts to a standard vocabulary through a process known as concept mapping. In 2003, five controlled terminologies were proposed to represent different types of concepts, such as SNOMED-CT for clinical terms [44], LOINC for laboratory test orders and results [45, 46], RxNorm for clinical drugs [47], NDF RT for pharmacologic properties of medications, and UMDNS for medical devices [48](refer to supplementary table S1 for more details). These terminologies typically use hierarchical organizations and ontologies to define relationships between concepts. For example, an ontology may define the "is-a-type-of" relationship between "myocardial infarction" and "heart disease." Repositories like UMLS [49] and Athena [50] provide access to these ontologies. However, the current process of mapping clinical concepts to standard ontologies is time-consuming and requires expert knowledge, even with tools like Usagi [32, 51, 52, 53]. Some concepts are particularly difficult to map, such as those in the drug exposure category, with only 38% being mapped in one study [32]. As a result, fully automating this process remains a significant challenge. The current best-performing methods involve an initial automated mapping followed by manual curation by experts. However, given the large number of medical concepts in an EHR dataset, manual mapping at scale is impractical in many cases. Overall, improving the quality and standardization of EHR systems is crucial to enhancing the scalability, reproducibility, and reliability of EHR-based research [54]. This would ultimately result in better healthcare quality, reduced costs, and wider adoption of EHR systems [55]. To tackle the above mentioned challenges, we have developed the EHR-QC toolkit. This fully-automated pipeline is specifically designed for the standardization of EHR data encoding, fully automated concept mapping, and comprehensive quality assurance of healthcare data. The EHR-QC toolkit has the potential to become an integral part of digital health workflows that rely on EHR data to perform observational studies.

2. Methods

The EHR-QC is composed of two main modules namely "Standardization Pipeline" (Figure 1.17) and the "Preprocessing Pipeline" (Figure 1.18) consisting of various utility Python code functions (Figure 1.1 - 1.16) for handling EHR. This toolkit is a command-line Python utility with a straightforward setup and user interface. A complete step-by-step guide to running the pipeline has been provided (`https://ehr-qc-tutorials.readthedocs.io/`). The following sections describe the technical details of different modules of the pipeline and provide case studies to demonstrate its utility.

2.1. Data sources

Two different data sources are used at different stages for developing and benchmarking EHR-QC modules. During the development of the standardization pipeline, we used the Medical Information Mart for Intensive Care (MIMIC) IV data, a de-identified EHR dataset collected from critical care settings in a US hospital [56]. Both the original MIMIC schema and its OMOP-CDM conversion were used to validate preprocessing functionalities. We also used a benchmark dataset obtained from a recent paper [57] on the conversion of UK Biobank EHR to OMOP-CDM. This dataset included concepts from three categories - "Operations," "Non-Cancer Illnesses," and "Cancers" - obtained from various EHR sources, along with their mappings to a standard ontology. The mappings were curated by a team of experts using Usagi tool in a semi-automated manner. The availability of expert-curated mappings was the primary reason for selecting this dataset as a benchmark, as it helps to ensure the accuracy and consistency of the concept mapping process.

2.2. Custom configuration setup

The pipeline is completely flexible to allow inputs both as an existing database schema and as flat text files in ".CSV" format containing any type and range of attributes. A collection of the module functions can be invoked as a single pipeline, such as the "Standardization Pipeline" (Figure 1.17) and the "Preprocessing Pipeline" (Figure 1.18) enabling complete automation of the end-to-end EHR data processing activity. Appropriate initial configurations are provided through a configuration file. The configuration file allows users to manage the customisations such as the database connection parameters, intermediate schema names for lookup, source, extract and, CDM tables, standard vocabulary file paths, paths of EHR source files, column mappings for each of these files, and boundary values for various attributes for performing data quality checks. These configuration options make the pipeline flexible by adapting to any variation in the source data and also to run in a fully automated manner. Detailed custom use cases are provided in our online documentation of the pipeline.

2.3. Migrating the EHR data to the OMOP-CDM schema

The OMOP-CDM migration module provides utility functions to facilitate the process of converting any EHR representation to the OMOP-CDM schema. The database templates for the migration scripts obtained from an earlier migration effort [58] are embedded within the Python codebase, which dynamically builds queries based on the configurations, forming a layer of abstraction for users. This module provides automatic end-to-end migration functionality of any EHR to the OMOP-CDM, including standard vocabulary import and concept mapping.

The first step in the migration process is to import the standard ontologies (Figure 1.5, Figure 1.6) and the EHR (Figure 1.2) in their raw format into a database. The EHR data can be sourced either from a custom schema (Figure 1.1) or as a structured tabular flat file (Figure 1.4) typically formatted as CSV files. In this step, appropriate column mappings are to be provided if the data structure varies from the expected convention as shown in Figure S1. In the next step, the imported information is dumped into staging tables (Figure 1.2, Figure 1.6), from which the standard entities are extracted by the process known as ETL without affecting the raw data in the import tables. Depending on the source schema, extracting the OMOP-CDM entities might involve filtering information or merging attributes from staging tables. Mapping non-standard concepts in the source EHR to a standard ontology term known as concept mapping (Figure 1.8) is the most crucial and time-consuming step of the process. This step involves standardising non-standardised concepts in the EHR by either automatically mapping them with the desired standard ontologies or importing a pre-built custom mapping file. To facilitate concept mapping, we have developed a novel method to automatically perform this process, the details of which are discussed in the next section 2.4. Next, during the extract step (Figure 1.9), the EHR data is cleaned if needed, mapped to concepts available in the vocabulary tables, and OMOP-CDM entities are extracted. In the final step, the extracted entities are unloaded (Figure 1.10) to the final OMOP-CDM database (Figure 1.11). In this module, all the intermediate tables are automatically created and stored for any further analysis and audit. Further, this enables individual stages to be run independently, also resuming from where it is left off when run in pipeline mode.

2.4. Mapping clinical concepts to controlled vocabulary

Concept mapping typically involves three scenarios. The first scenario is when the source data already adheres to the desired standard. In this case, data can be directly moved to the target schema after performing basic code integrity checks. The second scenario is when the source data is standardized using a different standard than the target ontology. For well-established and compatible standards, a pre-existing mapping can be used to obtain the corresponding desired standard ontology terms. However, when no readily available mapping exists, *de novo* mapping between the source

EHR standard and the desired ontology standard must be performed. The third scenario is when the source contains concepts that are not mapped to any ontology and might be collected as free-flow text. In this case, concepts from the source EHR need to be mapped individually onto a standard ontology term. This is summarised in Supplementary Table S2

There are several techniques available for performing concept mapping by measuring the similarity of a search term to standard ontology terms. One such technique is approximate string matching, also known as fuzzy matching. Fuzzy matching provides a similarity score based on the Levenshtein distance [59] between two strings, which quantifies the character-level differences between them. This method works best when the source text is taken from a controlled vocabulary that has only minor variations from standard concepts. To increase flexibility when dealing with diverse vocabularies, fuzzy algorithms can be extended through reverse indexing, word tokenization, and conditional mapping of tokens. A very popular tool, Usagi (`https://github.com/OHDSI/Usagi`) developed by OHDSI consortium, uses an extended reverse indexing based technique internally to provide the matching terms. Next, semantic matching algorithms seek to find the closest meaning match between two phrases, instead of just comparing their textual composition. To accomplish this, word embeddings are generated, which are multi-dimensional distributed vectors representing each phrase. In an n-dimensional space, the embeddings of similar phrases are closer to each other while the embeddings of opposing phrases are more distant. Therefore, a possible match can be identified by selecting the standard concept whose embedding is closest to the embedding of the search phrase. The medical concept annotation tool Medcat used this technique to detect clinical concepts in texts [60]. Semantic matching takes into consideration the semantics of concepts, allowing for the mapping of a more generic and diverse terminology. However, it falls short of human-level performance, making standalone automatic mapping algorithms highly error-prone and not a viable alternative for semi-automatic expert curation. Basically, no single algorithm is the most effective in all scenarios, as their effectiveness depends on the nature of the data to be mapped, as depicted in Figure S2.

Standalone concept mapping techniques are not as effective as expert curation. This makes them unsuitable for complete automation. Therefore, we implemented Majority Voting, a composite approach that only retains mappings supported by more than one standalone algorithm. Although this approach improves mapping accuracy, it also reduces mapping coverage. To address this issue, we developed another composite algorithm called "Majority Voting Plus." In the first step, this algorithm identifies all mappings supported by two or more algorithms, like Majority Voting. For the unmapped concepts, we use Medcat, Usagi, and Fuzzy in that order of preference to obtain the first available mapping. With this approach, we resolved the low coverage issue while retaining superior performance. We have included Majority Voting Plus as part of the standardisation module in EHR-QC, which provides the only fully automated solution for EHR standardisation to the best of our knowledge. The mappings can also be saved as a CSV file for manual review later.

2.5. Data preprocessing to perform exploratory analysis and the quality assurance

2.5.1. Exploratory data analysis and reporting anomalies

The data preprocessing module is equipped with various functions that perform monotonous data preprocessing activities like extraction, exploration, quality assurance (QA), correction, and preparation of EHR data for downstream analysis tasks. The *extract* function can be invoked to generate flat files by specifying connection details to the source repository stored as a relational database such as in SQL or postgres. Subsequent functions can be executed independently, decoupled from the data source, since they accept flat files as inputs. This module's objective is to standardise the EHR data preprocessing process by providing a convenient library.

The extract module (Figure 1.12) fetches the demographic, vitals, and lab measurement information from the OMOP-CDM schema by default, additionally, it can also be configured to read the data in the MIMIC IV format and saves it in a csv file. Next, the *exploration* module (Figure 1.13) is used to generate reports aimed at providing a comprehensive overview of the healthcare data. The reports contain information about the attributes' type, count, range, distribution, and summary statistics, along with information on anomalies such as missing values.

Further, the *QA* module (Figure 1.14) can be used to generate visualisations and statistics highlighting common anomalies such as missing data, outliers, errors, and other systematic inconsistencies. Additionally, this module not only identifies anomalies but also quantifies each category and offers remediation recommendations. For instance, the report displays the count and percentage of missing data and outliers. It also detects the presence of multiple data standards or distributions which can indicate data contamination, by obtaining the data modality. Lastly, to check the plausibility of systematic inconsistencies the distributions of the attributes are plotted against the predetermined boundary conditions and are visualised.

2.5.2. Handing anomalies: missingness

The correction module (Figure 1.15 - *ehrqc.qc.Impute*) handles the tasks of dealing with the missing data. It performs a comparative analysis of various imputation algorithms such as mean imputation, median imputation, K-nearest neighbours (KNN) imputation, MissForest [61], Expectation Maximisation [62], and Multiple Imputation [63] are performed on the dataset. The imputation algorithms are applied to a random set of missing values with a specific percentage of missingness simulated artificially, and the root mean square error (RMSE) values are calculated to determine the best imputation strategy for the given dataset. This is repeated for all the algorithms and the RMSE in each case is compared to determine which algorithm works the best for the given data and the given proportion of missingness. The best-performing algorithm is applied automatically to impute missing values with the least RMSE score.

2.5.3. Handling anomalies: outliers

The correction module (Figure 1.15 - *ehrqc.qc.Anomalies*) utilizes Item Response Theory (IRT) [64], an ensemble of unsupervised outlier detection algorithms to detect the outliers. This algorithm returns an ensemble score for every data point which is a combination of the outlier scores obtained from multiple unsupervised algorithms. This technique avoids the use of hard-set boundaries found in traditional outlier detection methods, which can lead to biased analysis and the removal of genuine data. This technique combines results from multiple unsupervised outlier detection algorithms to assign an overall anomaly score to each data point. Data points exceeding a threshold score are considered extreme values and excluded from further analysis. Lastly, this module (Figure 1.16) includes two functions for standardising and re-scaling the data which is essential for many machine learning tasks. These utility functions enable the user to generate a quality-assured data matrix that can be used to perform predictive modelling.

3. Results and discussion

Evaluation of MIMIC-IV EHR data migration to OMOP-CDM reveals improved data quality and utility

To validate the migration process, we utilized our standardization pipeline to transfer MIMIC-IV EHR data to the OMOP-CDM schema. The process resulted in significant improvements in quality and utility of the EHR. For instance, the pipeline efficiently excluded any unusable patient entries that lacked subsequent entries. Moreover, a new data table was created during the migration process to store death information extracted from the admissions table, to increase the accessibility. The flow of data from the source schema to the OMOP-CDM through intermediate tables is illustrated in Figure 2. As part of the migration workflow, we have developed an automated concept mapping technique which will be explicated in the following section. This provides an opportunity to resolve discrepancies that may arise from the use of multiple units of measurement and harmonise redundant concepts. In the entire migration process, intermediate tables function as audit tables that ensure complete transparency (S3). It is possible to monitor the data that has been successfully migrated to the destination tables, as well as the data that has not been migrated due to various reasons such as inadequate quality or mapping failure. We successfully migrated 337,942 individuals, 2,435,481 visit occurrences, 468,919,408 measurements, and 9,331 death records from MIMIC-IV EHR to the OMOP-CDM structure in the process.

Majority Voting Plus outperforms expert-curated mappings with comprehensive coverage and high alignment

We benchmarked concept mapping performance of our pipeline using a published dataset [57]. Our results indicate that the Majority Voting approach yields a better alignment with curated concepts and minimises non-overlapping mappings when compared to standalone techniques (see Figure S3). However, since the Majority Voting gives a mapping only if there is a consensus amongst two or more standalone algorithms, it has a poor overall coverage as a substantial portion of the concepts remained unmapped due to the lack of consensus (see Figure 3A). The Majority Voting Plus technique on the other hand boosts coverage to a level similar to that of standalone algorithms (Figure 3A).

Further, we have evaluated the performance of our algorithm by comparing the mappings with the expert-curated concepts. The alignment between the two is presented in Figure (Figure 3A) where the proportions of mappings that match, do not match, or are not mapped to the curated values are displayed. The matching percentage gives the proportion of concepts in agreement with the curated concepts, providing a measure of the quality of the mapping. It is important to note that a non-matching percentage does not necessarily imply incorrectness, as both the curated and

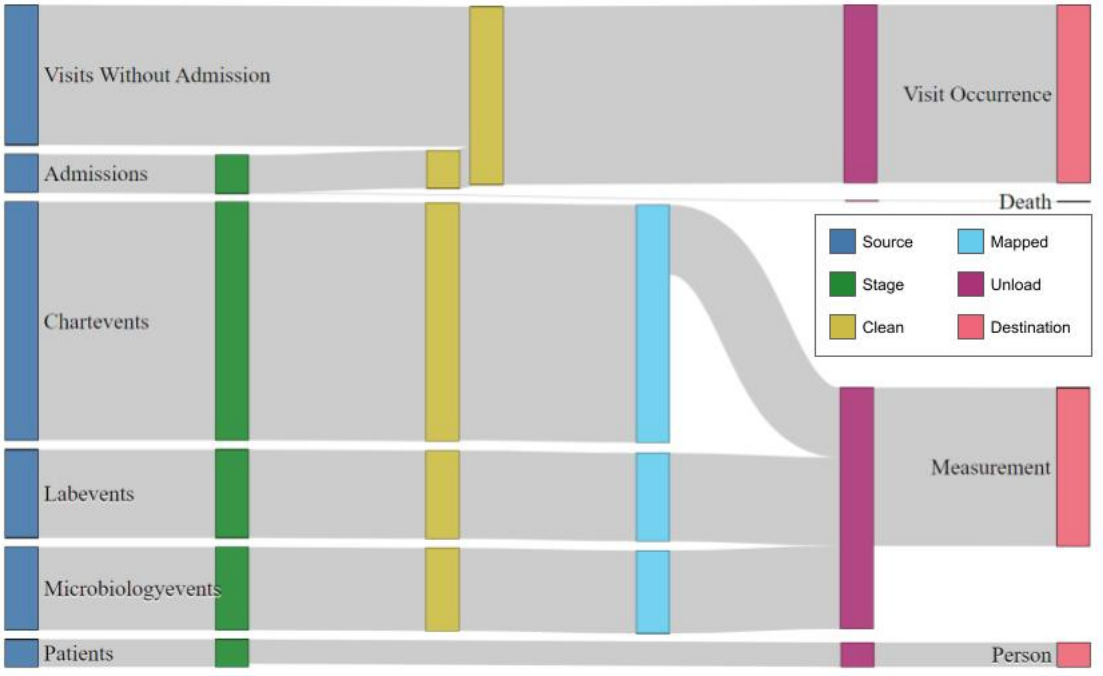


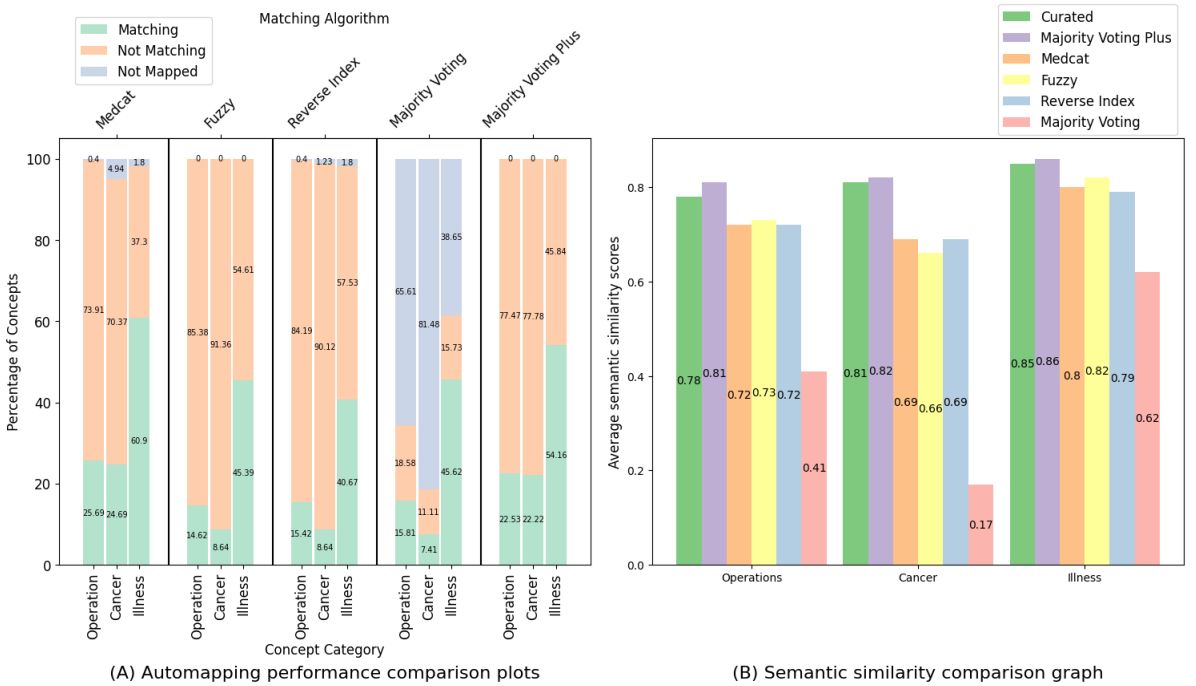
Figure 2: The figure illustrates the data flow diagram of four different entities. In all cases, the data from the source schema is imported into a staging table which is transformed into standard entities after performing cleaning and mapping to a desired standard. The standard entities are dumped in the unloading table which will be finally pushed to the destination OMOP-CDM schema.

the mapped concepts may be correct in some cases, as shown in Supplementary Table S4. Our analysis demonstrates that concepts within the Illness category that are well-standardized exhibit a better alignment with curated mappings than those in other categories, as evidenced by the higher matching percentage of the algorithms. Conversely, the cancer concept type, which lacks a well-established standard vocabulary, performs poorly, particularly with text content matching algorithms like Fuzzy and Reverse Index. However, Medcat, a semantic matching technique, performs well in this category. Further optimization of Medcat's performance can be achieved by fine-tuning it on more cancer-related text, as approximately 5% of the values remain unmapped. Our analysis also indicates that the Majority Voting Plus algorithm provides a mapping for every queried concept while maintaining a mapping percentage only slightly lower than that of Medcat. In summary, this figure shows that Majority Voting Plus and Medcat are the best-performing algorithms in terms of coverage and fidelity of the mappings.

Continuing our assessment of concept mapping, we utilised Semantic Similarity Score as a final criterion to measure the proximity of the mappings to the intended meaning of the concepts. Figure 3B compares the Semantic Similarity Scores between the query concepts and the mappings generated by the algorithms for three concept categories. The graphs show that expert-curated mappings were more similar to the query concepts than mappings derived from standalone techniques. However, our study found that the Majority Voting Plus approach consistently outperformed expert-curated mappings for all concept categories. This approach has comprehensive coverage and strong alignment with curated concepts, making it suitable for automatic concept mapping without compromising quality. This allowed us to fully automate the standardisation pipeline as part of the EHR-QC utility.

Data exploration and quality reports provides an overview of the data and detect anomalies

We preprocessed EHR data in the standard OMOP-CDM format. Our first step was to generate data quality reports that contained exploration and anomaly graphs. Exploration graphs provided a comprehensive overview of the data, while anomaly graphs showed common anomalies such as missingness, outliers, inconsistencies, and systematic



(A) Automapping performance comparison plots (B) Semantic similarity comparison graph

Figure 3: (A) An illustration of the percentage of coverage and mapping performance for Fuzzy, Reverse Index, Medcat, Majority Voting, and Majority Voting Plus techniques under three concept categories namely, Operation, Cancer, and Illness. The source texts that are not mapped are included under the Not Mapped category whereas the mappings that are in agreement with the expert-curated concepts are counted as Matching while the rest are included under Not Matching. (B) The average semantic similarity score which in this case is a population means taken for the measure of closeness in meaning to the source text. It is plotted for the expert-curated concepts along with the mappings from Fuzzy, Reverse Index, Medcat, Majority Voting, and Majority Voting Plus techniques. The medical concepts includes are Operations, Cancer, and Illness concept categories.

errors. The supplementary figure (Figure S4) presents the anomaly graphs generated by EHR-QC on the left, and the corresponding corrected data on the right. Our reports also included summary statistics of the data, such as the type and number of attributes, missingness and outliers, errors, and the proportion of data within user-specified value ranges. Overall, these reports are useful for gaining an overview of the EHR data and identifying anomalies.

Missing data imputation simulates multiple imputation strategies and applies the best one for the given dataset

In addition to providing a general overview of the data and the anomalies, the EHR-QC preprocessing module includes utility functions to rectify the identified anomalies. Figure 4 presents the performance of missing data comparison and imputation utilities using various imputation techniques. In Plot 4A, the reconstruction r-squared score is plotted against different missing proportions ranging from 0 to 50 for various missing value imputation techniques. According to our analysis, the Expectation Maximisation algorithm performed the poorest among all the algorithms tested on the given data, regardless of the proportion of missing data. On the other hand, when the proportion of missing data ranged from 10-25%, MissForest showed the best performance, while for the 25-50% range, KNN outperformed the other algorithms. Interestingly, for missingness beyond 40%, Mean Imputation displayed the best performance. These findings demonstrate that the optimal algorithm for imputing missing data depends on the proportion of missing data in the dataset. Therefore, our results can guide the selection of the most appropriate algorithm based on the amount of missing data present in a given dataset.

The plots 4B and 4C display the data with and without any missing values, respectively. Whereas, plots 4D, 4E, and 4F display the correlation between the principal components of the original data and the imputed data for simulated missingness of 5%, 20%, and 35%, respectively. These plots illustrate the performance degradation of the missing data

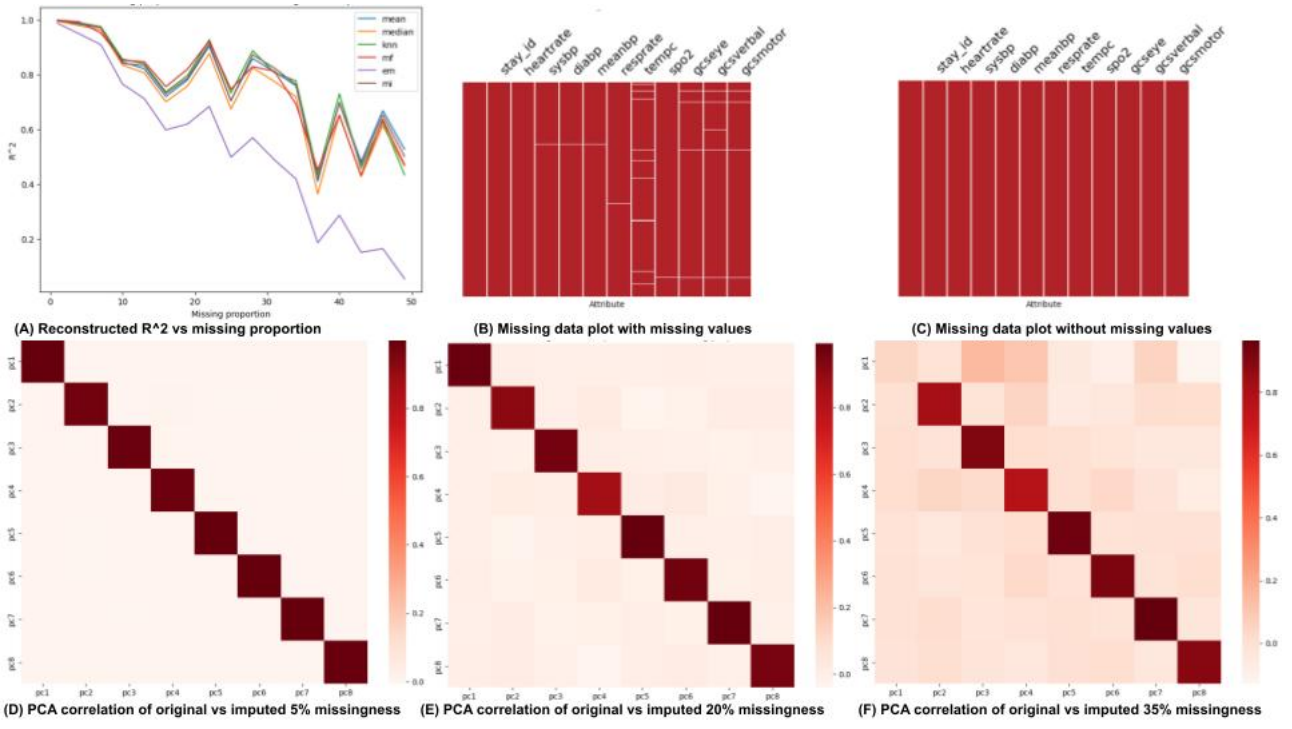


Figure 4: A) The reconstruction r squared values for some of the missing value imputation algorithms for a range of missing proportions. B) Missing data plots with missing values in the data. C) Missing data plots without missing values in the data. D) Correlation between the principal components of the original data and the imputed data from 5% missingness. E) Correlation between the principal components of the original data and the imputed data from 20% missingness. F) Correlation between the principal components of the original data and the imputed data from 35% missingness.

imputation with the increase in missing data proportion highlighting the need for choosing an appropriate imputation technique suitable for the data provided.

Outlier detection utility helps in the detection and treatment of outliers in an adaptive manner using an ensemble of unsupervised algorithms

We demonstrated how EHR-QC facilitates the identification and removal of outliers. Plots 5A, 5B and 5C in Figure 5 demonstrate EHR-QC's ability to adaptively detect outliers. To demonstrate this, a simplified dataset with two attributes, "systolic blood pressure" and "heart rate," was plotted in three scenarios. Plot 5A shows the unprocessed data, where extreme values are observed for both attributes. The ensemble score reaches as high as 25, indicating the presence of eccentric data points, represented by darker-colored points on the graph, which are extreme outliers. Plot 5B, obtained from the same dataset after applying a conventional univariate rule-based method to remove outliers, demonstrates the limitations of using inflexible hard cutoff values for classifying outliers. The boundaries imposed by this method result in the truncation of natural data clusters. Although this technique removed extreme outliers, a few data points with ensemble scores up to 10 remained. Plot 5C, obtained by applying an unsupervised method called IRT to identify outliers, demonstrates that the algorithm effectively removed outliers while retaining the entire cluster with its natural boundaries. The highest ensemble score for any data point in this plot did not exceed the value 6. Plots 5D, 5E, and 5F, demonstrating how EHR-QC plots aid in addressing inconsistencies in the data. Plot 5D shows the density plot of a single attribute, temperature, obtained from the raw data, indicating the presence of extreme values. After removing the outliers from this attribute, other anomalies now become apparent as shown in Plot 5E which in this case is the existence of multiple units of measurement. This plot uses vertical lines indicating normal ranges to aid in identifying such inconsistencies. Ideally, the majority of values should fall within the normal range, as shown in plot 5F, which was obtained after unifying the measurement standard.

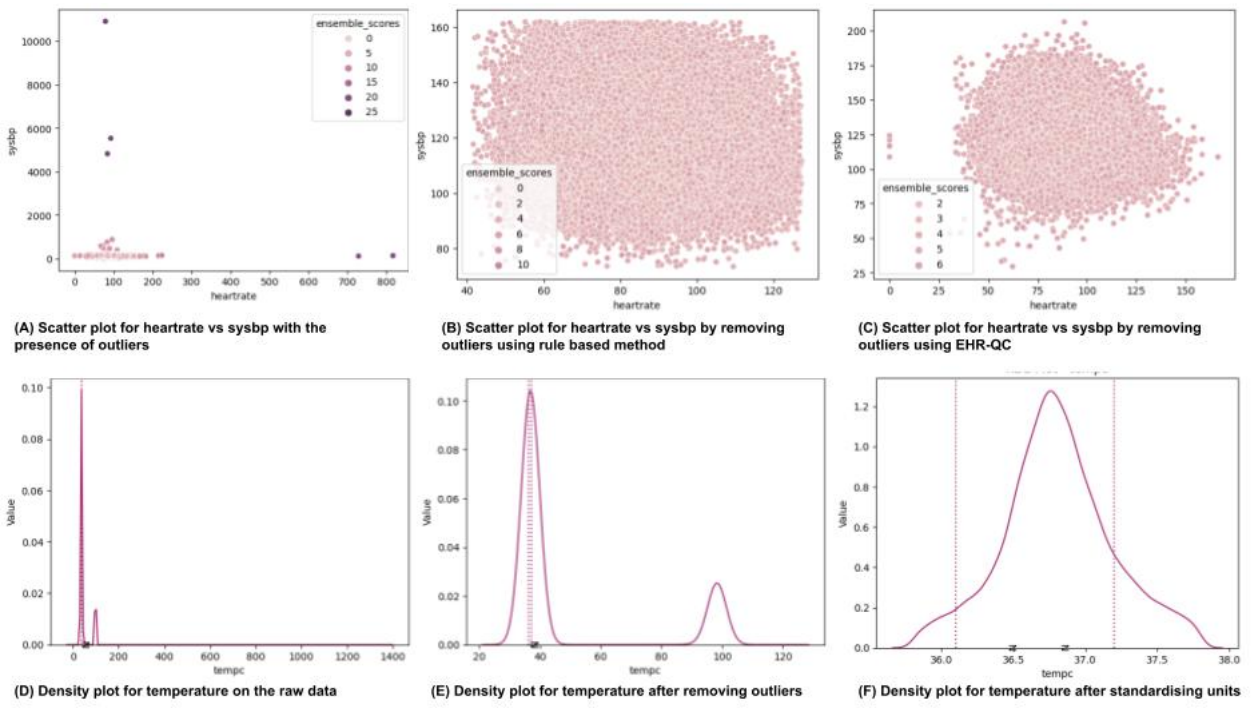


Figure 5: A) Scatter plot displaying ensemble outlier scores for systolic blood pressure vs heart rate plot for the raw dataset. B) Scatter plot displaying ensemble outlier scores for systolic blood pressure vs heart rate plot by removing outliers using a rule-based method. C) Scatter plot displaying ensemble outlier scores for systolic blood pressure vs heart rate plot by removing outliers using IRT. D) Density plot for the temperature attribute generated using the raw data. E) Density plot for the temperature attribute generated after removing the outliers. F) Density plot for the temperature attribute generated after removing the outliers and unifying the measurement standards.

4. Conclusion

Machine learning in digital health relies on large-scale healthcare data but is often limited to single-site data, hindering generalizability. Our work addresses this by providing a feasible EHR data harmonization workflow, enabling reproducible research outcomes through model validations on multi-site data.

Here, we introduced EHR-QC, a modular quality control pipeline that enables the conversion of EHR data to the standardized OMOP-CDM format. We also presented an innovative algorithmic solution for clinical concept mapping that surpasses the current expert curation process. This automation of data standardization represents a significant advancement, promoting the adoption of EHR standards and facilitating the development of more generalisable data-driven models. As a result, researchers can expect more efficient, reproducible, and robust research outcomes.

The EHR-QC pipeline also includes preprocessing functionalities for efficient exploration, quality assurance, and data preparation for downstream machine learning applications. Overall, EHR-QC offers a comprehensive and user-friendly solution for handling healthcare data, ensuring reproducible and robust research outcomes.

5. Acknowledgements

We acknowledge contributions from two research interns, Esha Singh and Geeta Kole for their initial analysis of MIMIC to OMOP conversion workflow. We also thank Tyrone Chen for proofreading the manuscript, and Jerico Revote from Monash eResearch Centre for his invaluable support.

AP, NM, ST acknowledge Medical Research Future Fund funding for the SuperbugAI flagship project. YR received Monash Graduate Scholarship for his PhD.

The authors also extend their sincere appreciation to the open research community responsible for making the following resources available which were instrumental in facilitating the execution of this research; IRT [64], MIMIC IV [56], MIMIC IV to OMOP CDM Conversion (https://github.com/OHDSI/MIMIC), UK Biobank transformation to OMOP-CDM [57], Athena (https://athena.ohdsi.org/), SNOMED (https://www.snomed.org/), Usagi (https://github.com/OHDSI/Usagi), Medcat [60], and HuggingFace [65].

6. Code availability

The EHR-QC source code is now accessible to researchers for investigative purposes through the following Git repository: https://gitlab.com/superbugai/ehrqc. Comprehensive documentation for the utility can also be found at https://ehr-qc-tutorials.readthedocs.io.

CRediT authorship contribution statement

Yashpal Ramakrishnaiah: Formal analysis, Investigation, Visualisation, Writing – review and editing, Software Development. **Nenad Macesic:** Investigation, Funding Acquisition, Supervision, Writing – review and editing. **Anton Y. Peleg:** Investigation, Funding Acquisition, Resources, Supervision, Writing – review and editing. **Sonika Tyagi:** Conceptualisation, Formal analysis, Investigation, Funding Acquisition, Resources, Supervision, Writing – review and editing.

References

[1] Hargobind S Khurana, Robert H Groves, Jr, Michael P Simons, Mary Martin, Brenda Stoffer, Sherri Kou, Richard Gerkin, Eric Reiman, and Sairam Parthasarathy. Real-Time automated sampling of electronic medical records predicts hospital mortality. *Am. J. Med.*, 129(7):688–698.e2, July 2016.

[2] Chad Anderson, Mala Kaul, and Dana Edberg. Increasing affordance potency through process improvement: case study of a healthcare system, 2019.

[3] Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D. Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, Wendy Chapman, and Rina Dutta. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *J. Biomed. Inform.*, 88:11–19, December 2018.

[4] Catherine Tong, Emma Rocheteau, Petar Veličković, Nicholas Lane, and Pietro Liò. Predicting patient outcomes with graph representation learning, 2022.

[5] Shahid Ali Choudhry, Jing Li, Darcy Davis, Cole Erdmann, Rishi Sikka, and Bharat Sutariya. A Public-Private partnership develops and externally validates a 30-day hospital readmission risk prediction model. *OJPHI*, 5(2), June 2013.

[6] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):1–10, May 2018.

[7] Jonathan Austrian, Felicia Mendoza, Adam Szerencsy, Lucille Fenelon, Leora I Horwitz, Simon Jones, Masha Kuznetsova, and Devin M Mann. Applying A/B testing to clinical decision support: Rapid randomized controlled trials. *J. Med. Internet Res.*, 23(4):e16651, April 2021.

[8] Michael J Rothman, Steven I Rothman, and Joseph Beals, 4th. Development and validation of a continuous measure of patient condition using the electronic medical record. *J. Biomed. Inform.*, 46(5):837–848, October 2013.

[9] Sebastian Salas-Vega, Adria Haimann, and Elias Mossialos. Big data and health care: Challenges and opportunities for coordinated policy development in the EU. *Health Syst Reform*, 1(4):285–300, May 2015.

[10] M P H Ellen Kim MD, Samuel M Rubinstein, Mphil Kevin T. Nead MD, Andrzej P Wojcieszynski, M S E Peter E. Gabriel MD, and M S Jeremy L. Warner MD. The evolving use of electronic health records (EHR) for research. *Semin. Radiat. Oncol.*, 29(4):354–361, October 2019.

[11] Alan Tomines, Heather Readhead, Adam Readhead, and Steven Teutsch. Applications of electronic health information in public health: uses, opportunities & barriers. *EGEMS (Wash DC)*, 1(2):1019, October 2013.

[12] Ravi B Parikh, Meetali Kakad, and David W Bates. Integrating predictive analytics into High-Value care: The dawn of precision delivery. *JAMA*, 315(7):651–652, February 2016.

[13] Behrouz Ehsani-Moghaddam, Ken Martin, and John A Queenan. Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. *Health Information Management Journal*, 50(1-2):88–92, December 2019.

[14] Michael G Kahn, Tiffany J Callahan, Juliana Barnard, Alan E Bauck, Jeff Brown, Bruce N Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G Johnson, Siaw-Teng Liaw, Marianne Hamilton-Lopez, Daniella Meeker, Toan C Ong, Patrick Ryan, Ning Shang, Nicole G Weiskopf, Chunhua Weng, Meredith N Zozus, and Lisa Schilling. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*, 4(1):1244, September 2016.

[15] Lily A Cook, Jonathan Sachs, and Nicole G Weiskopf. The quality of social determinants data in the electronic health record: a systematic review. *Journal of the American Medical Informatics Association*, 29(1):187–196, October 2021.

[16] Priscilla Y. A. Attafuah, Patience Aseweh Abor, Aaron Asibi Abuosi, Edward Nketiah-Amponsah, and Immaculate Sabelile Tenza. Satisfied or not satisfied? electronic health records system implementation in ghana: Health leaders' perspective. *BMC Medical Informatics and Decision Making*, 22(1), September 2022.

[17] Mark S. Iscoe, Robert M. McLean, and Edward R. Melnick. Restoring meaningful content to the medical record: Standardizing measurement could improve EHR utility while decreasing burden. *Mayo Clinic Proceedings*, 97(11):1971–1974, November 2022.

[18] Frank Fox, Vishal R Aggarwal, Helen Whelton, and Owen Johnson. A data quality framework for process mining of electronic health record data, 2018.

[19] Kitty S Chan, Jinnet B Fowles, and Jonathan P Weiner. Review: Electronic health records and the reliability and validity of quality measures: A review of the literature, 2010.

[20] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J. Am. Med. Inform. Assoc.*, 20(1):144–151, January 2013.

[21] Kristin M Hirata, Ann H Kang, Gina V Ramirez, Chieko Kimata, and Loren G Yamamoto. Pediatric weight errors and resultant medication dosing errors in the emergency department. *Pediatr. Emerg. Care*, 35(9):637–642, September 2019.

[22] Vojtech Huser, Xiaochun Li, Zuoyi Zhang, Sungjae Jung, Rae Woong Park, Juan Banda, Hanieh Razzaghi, Ajit Londhe, and Karthik Natarajan. Extending achilles heel data quality tool with new rules informed by Multi-Site data quality comparison. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1488–1489. IOS Press, 2019.

[23] Lorenz A Kapsner, Marvin O Kampf, Susanne A Seuchter, Gaetan Kamdje-Wabo, Tobias Gradinger, Thomas Ganslandt, Sebastian Mate, Julian Gruendner, Detlef Kraska, and Hans-Ulrich Prokosch. Moving towards an EHR data quality framework: The MIRACUM approach. *Stud. Health Technol. Inform.*, 267:247–253, September 2019.

[24] Martin Bialke, Henriette Rau, Thea Schwaneberg, Rene Walk, Thomas Bahls, and Wolfgang Hoffmann. mosaicQA - a general approach to facilitate basic data quality assurance for epidemiological research. *Methods Inf. Med.*, 56(7):e67–e73, May 2017.

[25] Keri N Althoff, Cherise Wong, Brenna Hogan, Fidel Desir, Bin You, Elizabeth Humes, Jinbing Zhang, Yuezhou Jing, Sharada Modur, Jennifer S Lee, Aimee Freeman, Mari Kitahata, Stephen Van Rompaey, W Christopher Mathews, Michael A Horberg, Michael J Silverberg, Angel M Mayor, Kate Salters, Richard D Moore, Stephen J Gange, and North American AIDS Cohort Collaboration on Research and Design. Mind the gap: observation windows to define periods of event ascertainment as a quality control method for longitudinal electronic health record data. *Ann. Epidemiol.*, 33:54–63, May 2019.

[26] Hyeoun Ae Park and Nick Hardiker. Clinical terminologies: A solution for semantic interoperability. *Journal of Korean Society of Medical Informatics*, 15(1):1–11, March 2009.

[27] S Trent Rosenbloom, Randolph A Miller, Kevin B Johnson, Peter L Elkin, and Steven H Brown. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. *J. Am. Med. Inform. Assoc.*, 13(3):277–288, February 2006.

[28] Snezana Savoska, Blagoj Ristevski, and Vladimir Trajkovik. Personal health record Data-Driven integration of heterogeneous data, 2023.

[29] Shahid Munir Shah and Rizwan Ahmed Khan. Secondary use of electronic health record: Opportunities and challenges, 2020.

[30] Rupa Makadia and Patrick B Ryan. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. *EGEMS (Wash DC)*, 2(1):1110, November 2014.

[31] Seng Chan You, Seongwon Lee, Soo-Yeon Cho, Hojun Park, Sungjae Jung, Jaehyeong Cho, Dukyong Yoon, and Rae Woong Park. Conversion of national health insurance Service-National sample cohort (NHIS-NSC) database into observational medical outcomes Partnership-Common data model (OMOP-CDM). *Stud. Health Technol. Inform.*, 245:467–470, 2017.

[32] Nicolas Paris, Antoine Lamer, and Adrien Parrot. Transformation and evaluation of the MIMIC database in the OMOP common data model: Development and usability study. *JMIR Med Inform*, 9(12):e30970, December 2021.

[33] Juan Espinoza, Abu Sikder, Armine Lulejian, and Barry Levine. Development of an OpenMRS-OMOP ETL tool to support informatics research and collaboration in LMICS. *Available at SSRN 4075625*, April 2022.

[34] Andrea Haberson, Christoph Rinner, Alexander Schöberl, and Walter Gall. Feasibility of mapping austrian health claims data to the OMOP common data model, 2019.

[35] Daniel M Lima, Jose F Rodrigues-Jr, Agma J M Traina, Fabio A Pires, and Marco A Gutierrez. Transforming two decades of ePR data to OMOP CDM for clinical research. *Stud. Health Technol. Inform.*, 264:233–237, August 2019.

[36] Dukyong Yoon, Eun Kyoung Ahn, Man Young Park, Soo Yeon Cho, Patrick Ryan, Martijn J Schuemie, Dahye Shin, Hojun Park, and Rae Woong Park. Conversion and data quality assessment of electronic health record data at a korean tertiary teaching hospital to a common data model for distributed network research. *Healthc. Inform. Res.*, 22(1):54–58, January 2016.

[37] Yue Yu, Nansu Zong, Andrew Wen, Sijia Liu, Daniel J Stone, David Knaack, Alanna M Chamberlain, Emily Pfaff, Davera Gabriel, Christopher G Chute, Nilay Shah, and Guoqian Jiang. Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. *J. Biomed. Inform.*, 127:104002, March 2022.

[38] Nicolas Paris and Adrien Parrot. MIMIC in the OMOP common data model. August 2020.

[39] Michael Kallfelz, Anna Tsvetkova, Tom Pollard, Manlik Kwong, Gigi Lipori, Vojtech Huser, Jeffrey Osborn, Sicheng Hao, and Andrew Williams. MIMIC-IV demo data in the OMOP common data model, June 2021.

[40] Hui Xing Tan, Desmond Chun Hwee Teo, Dongyun Lee, Chungsoo Kim, Jing Wei Neo, Cynthia Sung, Haroun Chahed, Pei San Ang, Doreen Su Yin Tan, Rae Woong Park, and Sreemanee Raaj Dorajoo. Applying the OMOP common data model to facilitate Benefit-Risk assessments of medicinal products using Real-World data from singapore and south korea. *Healthc. Inform. Res.*, 28(2):112–122, April 2022.

[41] Yuan Peng, Elisa Henke, Ines Reinecke, Michéle Zoch, Martin Sedlmayr, and Franziska Bathelt. An ETL-process design for data harmonization to participate in international research with german real-world data based on FHIR and OMOP CDM. *Int. J. Med. Inform.*, 169:104925, January 2023.

[42] Najia Ahmadi, Yuan Peng, Markus Wolfien, Michéle Zoch, and Martin Sedlmayr. OMOP CDM can facilitate Data-Driven studies for cancer prediction: A systematic review. *Int. J. Mol. Sci.*, 23(19), October 2022.

[43] Juan C Quiroz, Tim Chard, Zhisheng Sa, Angus Ritchie, Louisa Jorm, and Blanca Gallego. Extract, transform, load framework for the conversion of health databases to OMOP. *PLoS One*, 17(4):e0266911, April 2022.
[44] Kevin Donnelly. SNOMED-CT: The advanced terminology and coding system for ehealth. *Stud. Health Technol. Inform.*, 121:279–290, 2006.
[45] S M Huff, R A Rocha, C J McDonald, G J E De Moor, T Fiers, W D Bidgood, A W Forrey, W G Francis, W R Tracy, D Leavelle, F Stalling, B Griffin, P Maloney, D Leland, L Charles, K Hutchins, and J Baenziger. Development of the logical observation identifier names and codes (LOINC) vocabulary, 1998.
[46] A W Forrey, C J McDonald, G DeMoor, S M Huff, D Leavelle, D Leland, T Fiers, L Charles, B Griffin, F Stalling, A Tullis, K Hutchins, and J Baenziger. 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(1):81–90, January 1996.
[47] S Liu, Wei Ma, R Moore, V Ganesan, and S Nelson. RxNorm: prescription for electronic drug information exchange. *IT Prof.*, 7(5):17–23, September 2005.
[48] J A Gaev. The universal medical device nomenclature system. *Stud. Health Technol. Inform.*, 28:127–130, 1996.
[49] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70, January 2004.
[50] ÖzcanFatma. ATHENA. *Proceedings VLDB Endowment*, August 2016.
[51] Joao Rafael Almeida and Jose Luis Oliveira. Multi-language concept normalisation of clinical cohorts, 2020.
[52] Hao Liu, Simona Carini, Zhehuan Chen, Spencer Phillips Hey, Ida Sim, and Chunhua Weng. Ontology-based categorization of clinical studies by their conditions. *J. Biomed. Inform.*, 135:104235, November 2022.
[53] Elzo Pereira Pinto Junior, Priscilla Normando, Renzo Flores-Ortiz, Muhammad Usman Afzal, Muhammad Asaad Jamil, Sergio Fernandez Bertolin, Vinícius de Araújo Oliveira, Valentina Martufi, Fernanda de Sousa, Amir Bashir, Edward Burn, Maria Yury Ichihara, Maurício L Barreto, Talita Duarte Salles, Daniel Prieto-Alhambra, Haroon Hafeez, and Sara Khalid. Integrating real-world data from brazil and pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the global south. *J. Am. Med. Inform. Assoc.*, 30(4):643–655, March 2023.
[54] Obinwa Ozonze, Philip J Scott, and Adrian A Hopgood. Automating electronic health record data quality assessment. *J. Med. Syst.*, 47(1):23, February 2023.
[55] Christie Divine Akwaowo, Humphrey Muki Sabi, Nnette Ekpenyong, Chimaobi M Isiguzo, Nene Francis Andem, Omosivie Maduka, Emem Dan, Edidiong Umoh, Victory Ekpin, and Faith-Michael Uzoka. Adoption of electronic medical records in developing countries-a multi-state study of the nigerian healthcare system. *Front Digit Health*, 4:1017231, November 2022.
[56] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2023.
[57] Vaclav Papez, Maxim Moinat, Erica A Voss, Sofia Bazakou, Anne Van Winzum, Alessia Peviani, Stefan Payralbe, Michael Kallfelz, Folkert W Asselbergs, Daniel Prieto-Alhambra, Richard J B Dobson, and Spiros Denaxas. Transforming and evaluating the UK biobank to the OMOP common data model for COVID-19 research and beyond. *J. Am. Med. Inform. Assoc.*, 30(1):103–111, October 2022.
[58] Michael Kallfelz, Anna Tsvetkova, Tom Pollard, Manlik Kwong, Gigi Lipori, Vojtech Huser, Jeffrey Osborn, Sicheng Hao, and Andrew Williams. Mimic-iv demo data in the omop common data model, 2021.
[59] Joseph B Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237, 1983.
[60] Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. MedGPT: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*, July 2021.
[61] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.
[62] T.K. Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996.
[63] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.
[64] Sevvandi Kandanaarachchi. Unsupervised anomaly detection ensembles using item response theory. *Information Sciences*, 587:142–163, 2022.
[65] Shashank Mohan Jain. *Hugging Face*, pages 51–67. Apress, Berkeley, CA, 2022.