Journal of Clinical and Translational Science

www.cambridge.org/cts

EHRchitect: An open-source software tool for medical event sequences data extraction from Electronic Health Records†

Kostiantyn Botnar[1], Justin T. Nguyen[1], Madison G. Farnsworth[2], George Golovko[1] and Kamil Khanipov[1]

[1]Department of Pharmacology and Toxicology, University of Texas Medical Branch at Galveston, Galveston, TX, USA and [2]Department of Human Pathophysiology and Translational Medicine, University of Texas Medical Branch at Galveston, Galveston, TX, USA

Research Article

Cite this article: Botnar K, Nguyen JT, Farnsworth MG, Golovko G, and Khanipov K. EHRchitect: An open-source software tool for medical event sequences data extraction from Electronic Health Records. *Journal of Clinical and Translational Science* **9**: e79, 1–9. doi: 10.1017/cts.2025.55

Received: 14 March 2024
Revised: 10 March 2025
Accepted: 18 March 2025

Keywords:
Electronic Health Records; Python; research study configuration; database; data quality; medical event; event sequence; open-source software; data cleaning; data selection; electronic medical records

Corresponding author:
K. Botnar; Email: kobotnar@utmb.edu

†This article has been updated since original publication. A notice detailing the change has been published at doi: 10.1017/cts.2025.10183.

Abstract

Background: Electronic Health Records (EHR) analysis is pivotal in advancing medical research. Numerous real-world EHR data providers offer data access through exported datasets. While enabling profound research possibilities, exported EHR data requires quality control and restructuring for meaningful analysis. Challenges arise in medical events (e.g., diagnoses or procedures) sequence analysis, which provides critical insights into conditions, treatments, and outcomes progression. Identifying causal relationships, patterns, and trends requires a more complex approach to data mining and preparation. *Methods:* This paper introduces EHRchitect – an application written in Python that addresses the quality control challenges by automating dataset transformation, facilitating the creation of a clean, formatted, and optimized MySQL database (DB), and sequential data extraction according to the user's configuration. *Results:* The tool creates a clean, formatted, and optimized DB, enabling medical event sequence data extraction according to users' study configuration. Event sequences encompass patients' medical events in specified orders and time intervals. The extracted data are presented as distributed Parquet files, incorporating events, event transitions, patient metadata, and events metadata. The concurrent approach allows effortless scaling for multi-processor systems. *Conclusion:* EHRchitect streamlines the processing of large EHR datasets for research purposes. It facilitates extracting sequential event-based data, offering a highly flexible framework for configuring event and timeline parameters. The tool delivers temporal characteristics, patient demographics, and event metadata to support comprehensive analysis. The developed tool significantly reduces the time required for dataset acquisition and preparation by automating data quality control and simplifying event extraction.

© The Author(s), 2025. Published by Cambridge University Press on behalf of Association for Clinical and Translational Science. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial licence (https://creativecommons.org/licenses/by-nc/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use.

Introduction

While Electronic Health Records (EHR) systems have proven invaluable in improving patient care coordination and administrative efficiency, their primary focus has traditionally been on clinical and financial aspects of healthcare delivery [1]. Consequently, EHR systems may not always be optimally configured for large-scale research, resulting in limitations when accommodating diverse and complex requirements [2]. To overcome these limitations, numerous sophisticated tools and methodologies have been developed for health records data analysis. Many of them rely on statistical methods and machine learning (ML) models [3]. ML has been applied for many different tasks, including data collection, management, and analysis in both public health and clinical research [4], summarizing and visualizing data [5], quality of ambulatory care [6], and surgery complications [7]. Effective utilization of analytical tools necessitates an accurately curated data source, which may be hard to achieve for EHR data, considering its incomplete nature. This leads to the need for extensive EHR data preprocessing efforts before applying analytical methodologies and computational models, entailing a substantial investment of time. Furthermore, considering the constant increase in EHR data, it is imperative to ensure that the preprocessing procedures are replicable to uphold the integrity of research outcomes [8,9]. Given the frequent utilization of ML and artificial intelligence (AI) approaches within the healthcare domain, there is a demand for systematic and reproducible approaches to EHR data preprocessing.

When assessing data selection tasks for specific research purposes, it is essential to note that researchers rarely focus on a single medical event, such as a diagnosis or procedure. More commonly, patient cohorts that follow a sequence of events are given consideration. For example, patients who have experienced a disease, followed by various types of treatment, and observed outcomes over a defined period. Integrating this sequence of events with time

constraints and additional inclusion or exclusion criteria makes the data selection process particularly challenging.

Commercial EHR data repositories like TriNetX, Cosmos Epic, and IBM MarketScan offer curated medical data and accompanying analytical tools. Furthermore, most platforms primarily deliver basic descriptive statistics for essential cohorts. For instance, Cosmos Epic offers a suite of tools for investigators, such as the Slicer Dicer tool for cohort selection, but lacks suitable options for event sequence selection [10]. TriNetX has a query builder that is quite flexible and allows temporal relationships between events to be built. However, there are no integrated ML tools to run the modeling on the resulting cohort. Few EHR systems provide an export of the selected data as raw tables, requiring an additional data cleaning and preparation process (e.g., TriNetX). Thus, despite the availability of curated data, there is a lack of flexibility regarding advanced data selection tools and state-of-the-art analytical tools application.

Along with commercial solutions, a wide range of open-source software tools complements their commercial counterparts, designed to undertake a spectrum of data preprocessing tasks. Notable examples include growthcleanr [11] and ActiveClean [12], which excel in data cleaning and formatting, albeit without the capacity to filter data in line with specific study configurations, often resulting in the delivery of all initial records regardless of the target events. This limitation may pose challenges in handling extensive datasets containing millions of patient records. Conversely, FIDDLE [13], METRE [14], MIMIC-IV-Data-Pipeline [15], and MIMIC-Extract [16] focus on a limited set of medical parameters, generating simplified tables conducive to ML model usage. While these tools permit the accumulation of time series data for selected medical parameters, they cannot accommodate diverse event sequences, such as post-interventional medication treatment followed by an outcome occurrence. Other tools, like EHR-QC [17] and mosaicQA [18], specialize in data quality control and offer comprehensive services, including data standardization, preprocessing, and quality reporting. Although these software tools effectively fulfill their designated roles, they do not provide the flexibility to configure medical event sequences or selectively extract pertinent records, thereby failing to reduce the dataset complexity for downstream analysis.

This paper describes EHRchitect, the software tool engineered to streamline the automation of patient data preprocessing and selection according to the medical event sequences. It achieves this by creating a structured MySQL database optimized for data retrieval and selection, along with program modules dedicated to event sequence processing as configured by the user. Data selection configuration is facilitated by a JSON (JavaScript Object Notation) file [19] adhering to a specific format. EHRchitect output comprises files encapsulating patient records for each event, event transitions completed with time-related attributes, and patient and event metadata. It allows a significant reduction of the initial data set, effortless access to patient data corresponding to any event of interest, and subsequently engaging in statistical analyses or constructing ML models.

Methods

EHRchitect was developed using Python programing language (version 3.9) utilizing such libraries as pandas (version 2.1.4), SQLAlchemy (version 2.0.28), and sshtunnel (version 0.4.0). Our tool can be used across Windows (version 10 or 11 Home or Professional Edition) or Linux (tested on Rocky Linux version 8.10) operating systems on a local computer or remote server. It can connect to a remote or local server compatible with MySQL (version 8.0 or higher). Running the application necessitates a Python environment with libraries mentioned in the requirements.txt file stored in the project. Due to the multiprocessing approach, the most efficiency can be achieved by running EHRchitect on a computer with at least eight cores processor and a local MySQL server with version 8.0 or higher. Computational resource consumption (e.g. CPU time or RAM volume) depends on the amount of data and selection criteria.

Data source

EHRchitect works with remote and local datasets as long as they adhere to the specified structure. Our program uses a raw dataset to create a new database on an existing MySQL server with credentials set up in the program configuration file. Raw data transformation requires the data to be a set of CSV (comma-separated values) files with a compatible data structure [20] archived in a ZIP file (Figure 1). In the case of remote data set storage, EHRchitect facilitates the downloading of the necessary archive via a provided HTTP URL. Before uploading data to a MySQL database, it undergoes deduplication, date format standardization, and data transformation. Furthermore, the program imports the General Equivalence Mapping table [21] to translate codes from ICD-10 (International Classification of Diseases, 10th Revision) [22] to ICD-9 (International Classification of Diseases, 9th Revision) [23]. Following the completion of the database setup, EHRchitect generates a JSON file containing the host IP address and credentials, enabling seamless data access during the subsequent program usage.

EHRchitect data structure accommodates core data objects such as patient, encounter, diagnosis, procedure, medication, laboratory result, and vital signs. Patient and encounter records are determined by unique identifiers, which are mandatory and not nullable across all tables except those with medical code metadata. The patient table contains demographic fields like sex, race, ethnicity, marital status, and birth and death dates in a text format. All date fields across the DB have the format "YYYYMMDD." Encounter information is encompassed in the Encounter table and characterized by encounter type, start and end dates, and the patient's ID linked to the encounter. Tables Diagnoses, Medication, LabResult, VitalSign, and Procedure contain patient records with the information according to the table name and have common columns "code," "date," and "code_system." Laboratory results and vital signs are additionally characterized by the numerical or text value, or both, and these values' measurement unit. Each medication record may optionally have the route, brand, and strength.

Study configuration

Utilizing a clean and formatted DB, EHRchitect allows data to be selected according to user configuration. The study configuration file determines the arrangement of medical events within a chronological sequence. This file adheres to a predefined hierarchical structure and needs to be stored in JSON format. Using this configuration file, the program selects data according to all determined inclusion, exclusion, and temporal requirements. It delivers resulting records with metadata and temporal characteristics in the form of distributed Parquet file [24] (Figure 2.).

The configuration file contains a study description defined at the root by the parameters title, output directory path, and the

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press

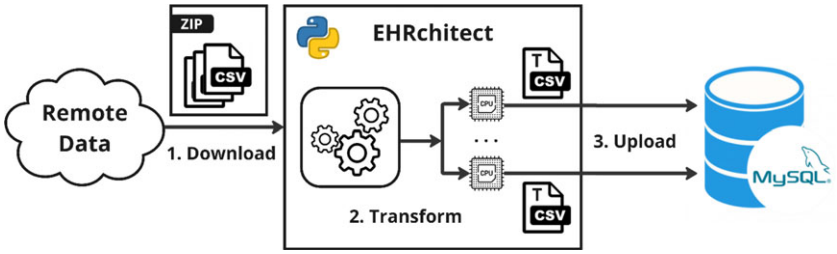


Figure 1. EHRchitect database preparation pipeline. Comma-separated values (CSV) files with raw data packed in a ZIP archive are downloaded using the URL the user provides. Each CSV file is transformed to the EHRchitect database format, along with data cleaning and transformation. The program creates a new MySQL database using MySQL server credentials provided by the user, and uploads transformed data with the following optimization.

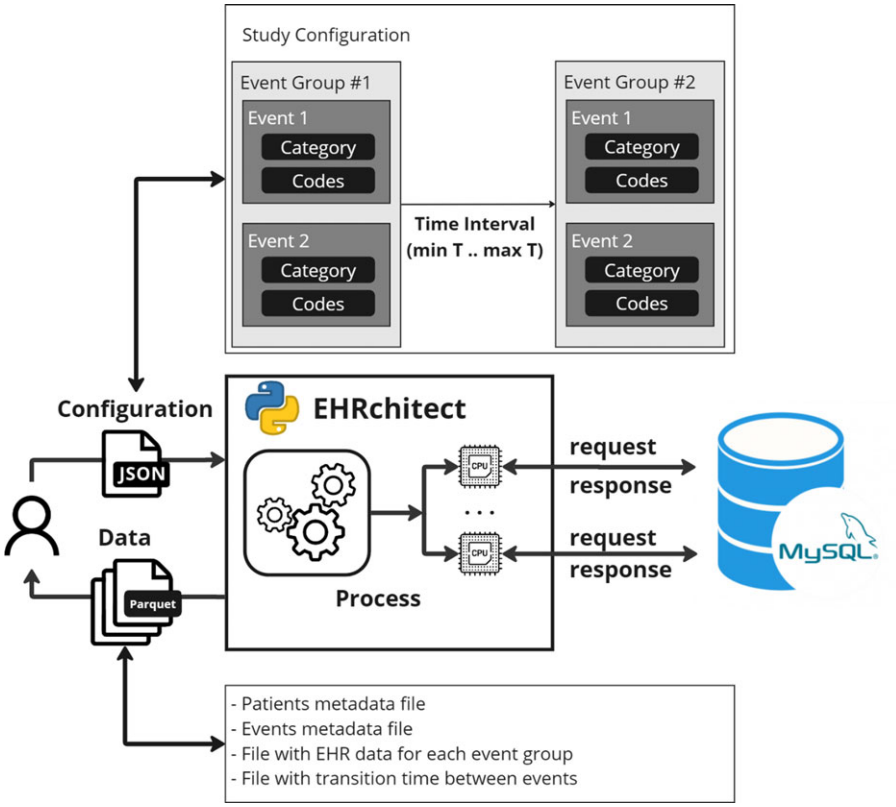


Figure 2. EHRchitect data extraction pipeline. The User describes a study in a JavaScript Object Notation file and passes it to the program. EHRchitect selects data according to all determined inclusion and exclusion criteria and time restrictions and delivers the resulting records with metadata and temporal characteristics in distributed Parquet files. Configuration file describes a study as a sequence of events with specified time constraints. Each event is determined by a list of codes (e.g., ICD-10, RxNorm) and a category (e.g., diagnosis, medication.

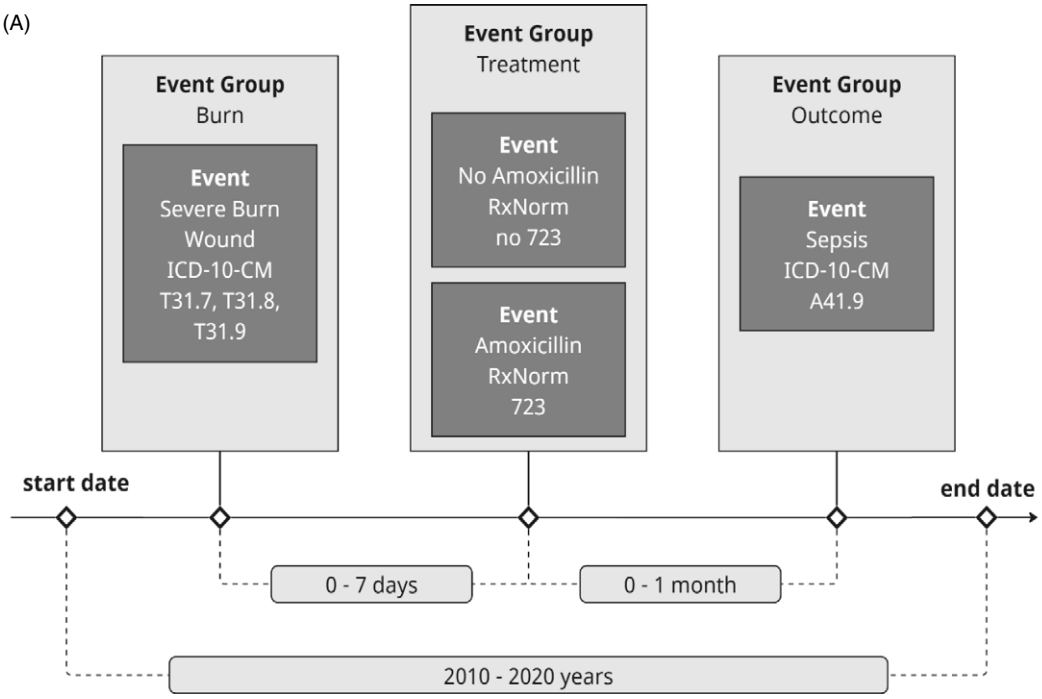
study timeframe. Each study must contain at least one event group. Every event group must include a numerical identifier for defining their chronological order, a list of events, and, optionally, a time limitation for these events. Each event must fall within a category, such as diagnosis, procedure, laboratory result, medication, or vital sign, and optionally specify a list of codes. Additionally, we have introduced a special event, denoted as patient death, which falls under the "patient" category and is assigned the code "DEATH."

Each configuration object has mandatory parameters. Every study must include at least one event group. Each event group must have a chronological order number and at least one event. Event groups will be searched in ascending order of their numbers. Every event within the group must be assigned a category. Event codes are optional. The search will encompass all records within the designated category and time interval if no codes were defined. Additional optional parameters provide the flexibility to define the study's overall timeframe, specify criteria for subsequent event searches, establish event exclusion and inclusion conditions, and explore intervals where events are absent.

For instance, if a study objective is to identify the impact of amoxicillin treatment within the first week after the severe burn wound (SBW) on the sepsis appearance within the following month during 2010–2020 years, we must create a configuration involving three event groups: first for the SBW, second one for the amoxicillin treatment, and the third one for the sepsis outcome (Figure 3). In this configuration, the SBW as the first event group is assigned group number of one and does not have any time restrictions. SBW events may be characterized by the diagnosis

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press

(A)



(B)

```
{ "name": "Amoxicillin impact on sepsis after SBW", period:{ "min_d": "01-01-2010", "max_d": "31-12-2020"},
  "event_groups": [
   { "event_group": 1, "name": "Burn",
     "events": [{"name": "Severe burn wound","id": "sbw","category": "diagnosis", "codes": ["T31.7","T31.8","T31.9"]} ]
},
   { "event_group": 2, "name": "Treatment",
     "period": {"min_t": 0,"max_t": 7,"unit": "day"},
     "events": [{"name": "Amoxicillin treatment","id": "amtr","category": "medication","codes": ["723"]},
          {"name": "No Amoxicillin treatment","id": "noamtr","category": "medication","codes": ["723"],"negation":
true}]},
   { "event_group": 3, "name": "Outcome",
     "period": {"min_t": 0,"max_t": 1,"unit": "month"},
     "events": [{"name": "Sepsis","id": "seps","category": "diagnosis","codes": ["A41.9"]}]}
  ]}
```

Figure 3. Study configuration example. A – an example of a study schema. The study explores an amoxicillin treatment impact on sepsis outcomes among patients with severe burn wounds (SBW). SBW is defined through a set of ICD-10 codes ("T31.7," "T31.8," "T31.9"). Amoxicillin is defined through the RxNorm code "723." Sepsis is defined through the ICD-10 code ("A41.9"). Study temporal parameters: Amoxicillin should be prescribed within seven days after the SBW. The outcome should appear within one month after the treatment or after the SBW in the not-treated cohort. Records from the 2010-2020 years only are considered. B – the study configuration file.

ICD-10 codes "T31.7," "T31.8," or "T31.9." The treatment group is assigned group number two and one-week time intervals, indicating that its events will be searched within one week after the SBW event. The treatment group will have two events: amoxicillin treatment and its absence. Amoxicillin may be defined by RxNorm (Prescription Normalized Naming System) code "723." Treatment absence is defined similarly as presence but with an additional Boolean parameter "negation" set to True. The last event group with the number of three contains one event with sepsis definition through the diagnosis ICD-10 code "A41.9" and the time interval of one month.

Patient records selection

In the initial step, the algorithm identifies all patients who meet the criteria of the first event group. Subsequently, these patients are divided into multiple subgroups that are processed concurrently. Subsequent data selection is carried out in individual processes for each group of patients. Simultaneously, each interaction with the SQL server within each process occurs in a distinct thread.

The study configuration encompasses various temporal parameters that influence the search for data records. At the highest level, the study time frame describes the restriction of the search period for all events in the study. Within each event group, the "period" parameter sets the search time boundaries for events within that group. If the period is not explicitly defined, events within the group are searched throughout all time following the events of the preceding event group. Additionally, each event can include an optional "period" parameter that overrides the search time interval for the event specifically. The number of events is controlled by the "match_mode" parameter that can take on two values: "all_matches," the default, which incorporates all records

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press

satisfying the configuration into the outcome, and "first_match," which includes only the chronologically first record for each patient.

Each event has a range of optional parameters, offering flexibility in record selection according to the researcher's requirements. These parameters enable a search with all possible subcodes of the event codes, a patient cohort selection that did not experience specific events within a designated time interval, and the establishment of complex exclusion and inclusion criteria for the event.

EHRchitect supports multiple code systems corresponding to various medical events, specifically ICD-10, ICD-9, LOINC (Logical Observation Identifiers Names and Codes) [25], RxNorm, CPT (Current Procedural Terminology) [26], HCPCS (Healthcare Common Procedure Coding System) [27], and SNOMED (Systematized Nomenclature of Medicine) [28]. Considering the transition from ICD-9 to ICD-10 code systems in October 2015, the tool automates the mapping of ICD-10 codes to their ICD-9 counterparts using the General Equivalence Mapping table version from 2018[21] while requesting the data to provide a comprehensive dataset. All result data records for events defined with ICD-10 codes include ICD-9 records as well but with ICD-10 labels.

Output description

The output of EHRchitect is a collection of Parquet files. These files contain event metadata, patient metadata, event records, and event transition records. The metadata is consolidated within a single file, while the event and transition records are stored across multiple distributed Parquet files.

The patient metadata file contains a unique patient identifier, date of birth, date of death, sex, race, and ethnicity. The event metadata file contains the event identifier, event name, code, category, code description, and event group name. Each event file includes patient identifiers, event identifiers, event codes, and the corresponding dates discovered for these patients. Transition files include patient identifiers, source events, and destination events, along with the time elapsed in days between these two events.

Results

EHRchitect reduces the researcher's time and effort on EHR data preparation and selection. A dedicated DB optimized for a search by specific parameters provides a stable data source with guaranteed results reproducibility. Flexible study configuration allows complex event sequence logic with different levels of temporal restrictions and inclusion and exclusion criteria. Changing study parameters or data selection conditions requires only configuration changes without rewriting SQL requests or programing code. The result files delivered by the program contain comprehensive details regarding events and patient metadata, event order, and temporal attributes, enhancing the depth of data analysis.

Our program is an open-source program that is available on GitHub [20]. It has been applied across various research domains, including organ transplantation [29], burn wound management, and cancer treatment [30]. Most recently, EHRrcitecht was used in a study to evaluate patient outcomes depending on specific pulmonary embolism treatment [31]. The study's setup is reflected in the configuration file shown in Figure 4. This study categorized events of interest into three event groups: pulmonary embolism (PE), intervention, and outcome.

Pulmonary embolism was identified using the ICD-10 diagnosis code "I26." The intervention comprised three events: thrombectomy or catheter-directed thrombolysis (any of CPT codes "37187," "37188," "37212," "37213," "37214"), pulmonary artery embolectomy (CPT code "1006322"), and thrombolytics (any of RxNorm codes "259280," "40028," "76895," "8410"). The outcomes group included six events: recurrent PE (ICD-10 code "I26"), stroke (ICD-10 code "I63.9"), patient death, other embolisms (ICD-10 code "I74"), gastrointestinal bleeding (any of ICD-10 codes "K92.0," "K92.1," "K92.2"), and intracerebral hemorrhage (ICD-10 code "I61" or "I62"). Patient death was classified as a specific event, determined by a reserved keyword "DEATH," and verified using the patient's date of death. The study was conducted with specific temporal conditions: only patients who underwent any intervention events within three days of the PE diagnosis were included. Outcomes were tracked for up to one year after the intervention, except for recurrent PE, defined as any PE event occurring at least ten days after the intervention. We selected the first occurrence of relevant codes, including subcodes for each event in the sequence. Additional exclusion criteria were applied to enhance study quality and precision. Patients with any record of the designated outcomes before the initial PE event were excluded. Additionally, those treated with thrombolytics within 30 days prior to PE diagnosis were also excluded. Each intervention event was further refined by applying the outcome exclusion criteria. Exclusion criteria for events were specified using the "exclude" parameter (Figure 5). The full configuration file is available in the supplementary material.

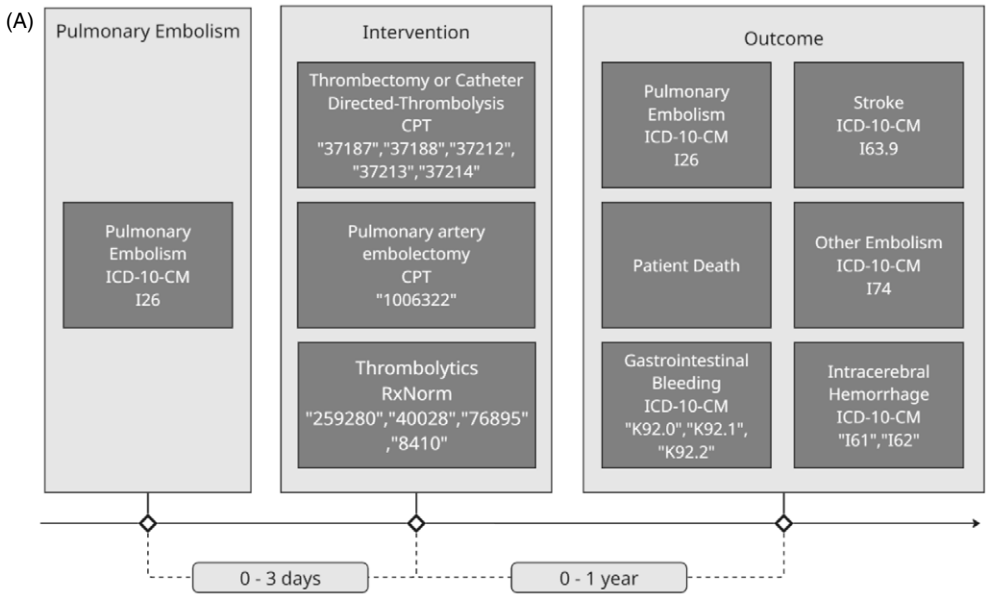
The output of the study configuration processing by EHRchitect is a set of tables containing patient records corresponding to each event group, transitions between events, and metadata for both patients and events. For this specific case, the resulting tables, which include metadata, PE events, interventions, and transitions between them, are shown in Figure 6.

The initial dataset included 2,224 patients with 5,817,500 records of prescribed drugs, 1,511,170 records of diagnosis, and 1,267,792 records of different procedures. Running EHRchitect on this data and described configuration utilizing a computer with 63 CPUs and 516 gigabytes of RAM, we received the result dataset within 23 minutes. The result included 460 patients who underwent PE, 339 of which had a following intervention, and 268 patients within the intervention group, who experienced any of the predefined outcomes. Combining all results in one table, we received a table with 615 patient records, including demographic parameters, event descriptions, and time intervals between events. This data enabled statistical analysis at multiple levels, including codes, events, and event groups. With demographic information available, we could compare different intervention cohorts statistically, calculate odds ratios, and assess treatment effects based on the defined outcomes.

Discussion

Integrating ML and AI in healthcare is becoming increasingly prevalent. A key requirement for many approaches is access to a well-curated and preprocessed dataset, particularly those targeting specific patient cohorts. While tools such as growthcleanr [11] and ActiveClean [12] offer data cleaning and quality control functionalities, they are often limited to single-table data. This may work well for a small dataset packed in a single file, but

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press



(B)

```
{"name": "Pulmonary Embolism Study",
 "levels": [{"level": 0, "name": "Pulmonary Embolism",
             "events": [{"name": "Pulmonary Embolism","id": "pemb",
                         "category": "diagnosis", "include_subcodes": true,"codes": ["I26"]}]},
            {"level": 1, "name": "Intervention","match_mode": "first_match",
             "period":{"min_t": 0,"max_t": 3,"unit": "day"},
             "events": [{"name": "Thrombectomy OR Catheter Directed-Thrombolysis","id": "thromect",
                         "category": "procedure","codes":["37187","37188","37212","37213","37214"]},
                         {"name": "Pulmonary artery embolectomy", "id": "embolectomy",
                         "category": "procedure","codes":["1006322"]},
                         {"name": "Thrombolytics", "id": "thrombolytics",
                         "category": "medication","codes":["259280","40028","76895","8410"]}]},
            {"level": 2,"name": "Outcome","match_mode": "first_match",
             "period":{"min_t": 0,"max_t": 1,"unit": "year"},
             "events": [{"name": "Pulmonary Embolism","id": "out_pemb",
                         "category": "diagnosis","include_subcodes": true,"codes": ["I26"],
                         "period":{"min_t": 10,"max_t": 365,"unit": "day"}},
                         {"name": "Stroke","id": "out_stroke",
                         "category": "diagnosis","include_subcodes": true,"codes": ["I63.9"]},
                         {"name": "Patient death","id": "out_death",
                         "category": "patient","codes": ["DEATH"]},
                         {"name": "Other Embolism","id": "out_oemb",
                         "category": "diagnosis","include_subcodes": true,"codes": ["I74"]},
                         {"name": "GI Bleed","id": "out_gibleed",
                         "category": "diagnosis","include_subcodes": true,"codes": ["K92.0","K92.1","K92.2"]},
                         {"name": "ICH","id": "out_ich",
                         "category": "diagnosis","include_subcodes": true,"codes": ["I61","I62"]}]}]}
```

Figure 4. The pulmonary embolism treatment research. A. Schematic research configuration. B. Example of the EHRchitect configuration file for the research.

it might become difficult to apply these tools to a dataset distributed throughout many files, encapsulating millions of records of different types (diagnoses, medications, vital signs, etc.). In contrast, EHRchitect adopts a more comprehensive approach by incorporating various data types and centralizing information in a MySQL DB to serve as a unified research data source.

An alternative EHR data mining and analysis approach involves using large clinical data platforms such as TriNetX, Cosmos, and IBM Watson. These platforms offer access to millions of patients' data and built-in analysis tools. However, the primary challenges of this approach are the high costs and the limitations of the available analytical tools. While exporting data from some of these platforms could potentially overcome the analytical limitations, researchers are left with raw datasets that require verification and formatting. EHRchitect was explicitly developed to address this issue, providing functionality to clean and preprocess datasets for further analysis. Although our software does not offer a direct connection to the existing EHR platforms, it has a built-in capability to operate with the datasets exported from TriNetX or with any other datasets compatible with our program data structure. Due to the EHRchitect's flexible and modular architecture and open source code, it is relatively easy for anyone to adapt it to new data sources. Minor adaptations like field names or formats can be handled with minimal effort, while more complex differences may require custom transformation logic.

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press

```
{"name": "Pulmonary Embolism Study",
"levels": [
        {"level": 0, "name": "Pulmonary Embolism",
        "events": [
                {"name": "Pulmonary Embolism",      "id": "pemb",
                "category": "diagnosis", "include_subcodes": true,
                "codes": ["I26"],
                "exclude": {
                        "events": [
                                {"name": "Previous outcome cases", "id": "pe_excl_out",
                                "category": "diagnosis", "include_subcode": true,
                                "codes": ["I26", "I63.9", "I74", "K92.0", "K92.1", "K92.2", "I61", "I62"]},
                                {"name": "Previous Thrombolytics", "id": "pe_excl_thrombolytics",
                                "category": "medication", "codes": ["259280", "40028", "76895", "8410"],
                                "period": {"min_t": -30, "max_t": -1, "unit": "day"}}
                        ]}
        ]},
...
}
```

Figure 5. Description of exclusion criteria in the configuration file. All exclusion criteria are described as events under the "exclude" object in the parent event they should be allied. If the period is absent, as in the "Previous outcome cases" event, the exclusion is applied to the entire period before the parent event.

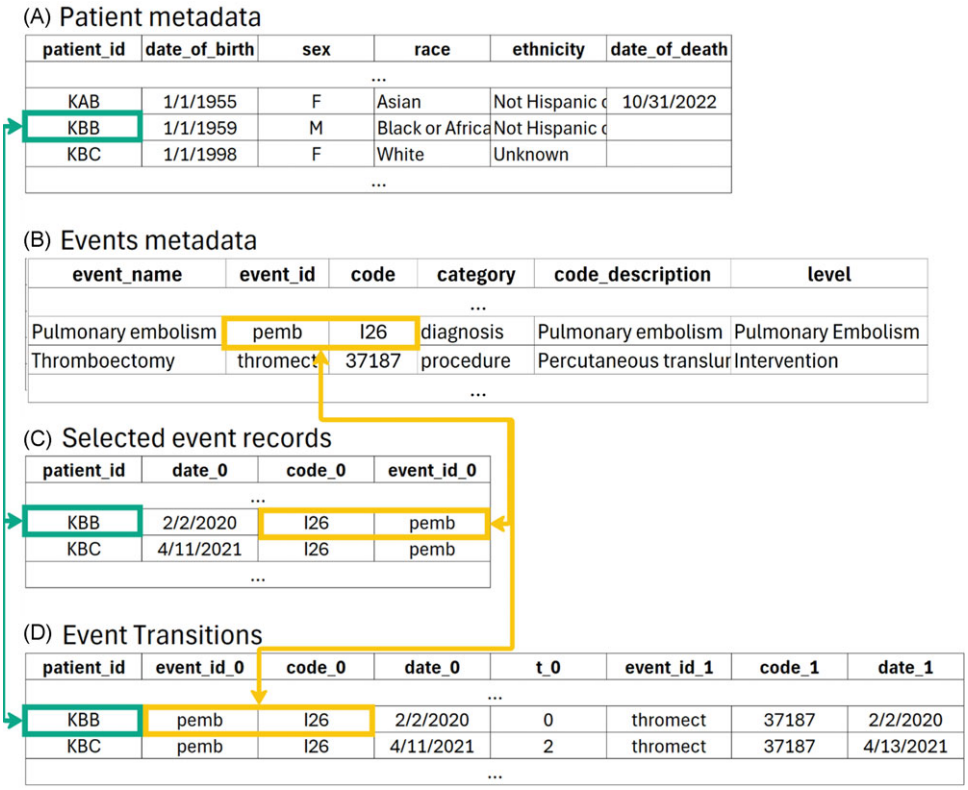


Figure 6. Result tables. A. The patients metadata table contains the demographic parameters of all patients across the study. B. The events metadata table describes the study events. C. Each event group includes patient records selected according to its description. D. The transition table shows patient records of the consequent events that satisfied time conditions. Columns with the siffix "_0" report the start event. Columns with the suffix "_1" report the finish event. Column "t_0" contains a number of days between the start and finish events. All tables are linked by the "patient_id" parameter. Event records are identified by the "event_id" and "code" fields.

A key advantage of EHRchitect is its ability to extract event data in a defined chronological order using a JSON file that encapsulates the entire study configuration. This file accommodates any number of events, each with unique criteria such as event category, chronological order, event negation, etc. Users can impose multiple time constraints across different levels of the study utilizing the parameter "period" with a different measurement scale: days, months, and years. The hierarchical organization of study configuration – ranging from specific codes at the most detailed level to event groups at a broader level – facilitates analysis at different scales. However, using a JSON format for study configuration may pose a learning curve, particularly for individuals unfamiliar with the format. Understanding this, we are working on a web interface to streamline the study configuration process and make it more accessible. This interface will enable users to configure studies intuitively, enhancing the overall user experience and accessibility.

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press

The data selection process following the study configuration outputs results as a set of Parquet files. We chose the Parquet format due to its higher efficiency than traditional column-oriented formats like CSV and TSV and its suitability for distributed data processing [32]. Although Parquet files are not supported by standard text editors or tools like MS Excel, they are compatible with a wide range of cloud data storage platforms, including Azure, Snowflake, AWS, and software like Tableau and Power BI. Additionally, Parquet files can be easily processed using common libraries of leading programing languages in data science, such as pandas and polars in Python and arrow in R, along with frameworks like Hadoop and Spark, making them an ideal choice for large-scale data analysis [33].

Although utilizing a MySQL DB offers significant advantages, including optimized data selection, rapid searches by medical codes and dates, and research reproducibility, it also presents certain challenges. Most significant among these is the complexity of MySQL server setup and maintenance. These tasks can be a significant barrier for researchers without technical expertise or adequate support, potentially leading to disengagement from our program. Furthermore, using MySQL may be excessive for smaller datasets with less than a thousand patients, as more straightforward solutions like CSV files could be more efficient. To address these issues, we plan to develop an alternative option that utilizes flat files instead of a relational database.

Conclusion

EHRchitect is a software tool that automates a routine data preparation process and medical event sequence data extraction. It streamlines the transformation process, automating the creation of a well-organized and optimized MySQL database, thus simplifying data extraction for medical event sequences based on user-defined study configurations. These event sequences are carefully curated, providing insights into patients who have experienced medical events specified by the user order and time intervals. The study's extracted data is efficiently presented as distributed Parquet files, encompassing a comprehensive dataset of events, event transitions, patient metadata, and event metadata. EHRchitect offers scalability, making it suitable for multi-processor systems by enabling concurrent data transformation and selection.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/cts.2025.55.

Acknowledgement. The authors acknowledge the Sealy Center for Structural Biology and Molecular Biophysics at the University of Texas Medical Branch at Galveston for providing research resources. The authors also wish to acknowledge Ka-Yiu Wong, the research scientist in the Sealy Center for Structural Biology & Molecular Biophysics at the University of Texas Medical Branch, for his valuable technical support in setting up and maintaining the server and database systems essential for this project.

Author contributions. The authors confirm contribution to the paper as follows: conception and design: Botnar K, Golovko G, Khanipov K; method development: Botnar K; manuscript preparation: Botnar K, Golovko G, Khanipov K, Farnsworth M, Nguyen JT; testing and approval: Botnar K, Golovko G, Khanipov K, Farnsworth M, Nguyen JT; responsibility for the manuscript as a whole: Khanipov K. All authors reviewed the results and approved the final version of the manuscript.

Funding statement. This study was conducted with the support of the Institute for Translational Sciences at the University of Texas Medical Branch, which is partially funded by a Clinical and Translational Science Award (UL1TR001439) from the National Center for Advancing Translational Sciences at the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Competing interests. The authors declare none.

References

1. **Holmes JH, Beinlich J, Boland MR, *et al.*** Why Is the electronic health record so challenging for research and clinical care? *Methods Inf Med.* 2021;**60**(1-02):32–48.
2. **Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL.** The evolving use of electronic health records (EHR) for research. *Semin Radiat Oncol.* 2019;**29**(4):354–361.
3. **Nadella GS, Satish S, Meduri K, Meduri SS.** A systematic literature review of advancements, challenges and future directions of AI and ML in healthcare. 2023;**5**(3):16.
4. **Marks M, Lal S, Brindle H, *et al.*** Electronic data management in public health and humanitarian crises. Upgrades, scalability and impact of ODK. Published online 2020.
5. **Ma KL.** Machine learning to boost the next generation of visualization technology. *Ieee Comput Graph.* 2007;**27**(5):6–9.
6. **Linder JA, Ma J, Bates DW, Middleton B, Stafford RS.** Electronic health record use and the quality of ambulatory care in the United States. *Arch Intern Med.* 2007;**167**(13):1400–1405.
7. **Weller GB, Lovely J, Larson DW, Earnshaw BA, Huebner M.** Leveraging electronic health records for predictive modeling of post-surgical complications. *Stat Methods Med Res.* 2017;**27**(11):3271–3285.
8. **Begley CG, Ioannidis JP.** Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res.* 2015;**116**(1):116–126.
9. **Denaxas S, Direk K, Gonzalez-Izquierdo A, *et al.*** Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *BioData Min.* 2017;**10**(1):31.
10. **Saini V, Jaber T, Como JD, Lejeune K, Bhanot N.** 623. Exploring slicer dicer, an extraction tool in EPIC, for clinical and epidemiological analysis. *Open Forum Infect Dis.* 2021;**8**(Supplement_1):S414–S415.
11. **Lin PD, Rifas-Shiman SL, Aris IM, *et al.*** Cleaning of anthropometric data from PCORnet electronic health records using automated algorithms. *JAMIA Open.* 2022;**5**(4):ooac089.
12. **Krishnan S, Wang J, Wu E, Franklin MJ, Goldberg K.** Activeclean: Interactive data cleaning while learning convex loss models. arXiv preprint arXiv: 160103797. Published online. 2016.
13. **Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J.** Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc.* 2020;**27**(12):1921–1934.
14. **Liao W, Voldman J.** A multidatabase exTRaction pipEline (METRE) for facile cross validation in critical care research. *J Biomed Inform.* 2023;**141**:104356.
15. **Gupta M, Gallamoza B, Cutrona N, Dhakal P, Poulain R, Beheshti R.** An extensive data processing pipeline for MIMIC-IV. *Proc Mach Learn Res.* 2022;**193**:311–325.
16. **Wang S, McDermott MBA, Chauhan G, Ghassemi M, Hughes MC, Naumann T.** MIMIC-extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. Proceedings of the ACM Conference on Health, Inference, and Learning. Published online 2020:222-235. https://doi.org/10.1145/3368555.3384469
17. **Ramakrishnaiah Y, Macesic N, Webb GI, Peleg AY, Tyagi S.** EHR-QC: a streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes. *J Biomed Inform.* 2023;**147**:104509.
18. **Bialke M, Rau H, Schwaneberg T, Walk R, Bahls T, mosaicQA HW.** A general approach to facilitate basic data quality assurance for epidemiological research. *Methods Inf Med.* 2017;**56**(7):e67–e73.
19. **International E.** The JSON data interchange Syntax . 2017. Accessed January 27, 2025. https://ecma-international.org/wp-content/uploads/ECMA-404_2nd_edition_december_2017.pdf
20. **Botnar K.** EHRchitect. October 3, 2024. Accessed January 27, 2025. https://github.com/kostabotnar/EHRchitect

21. **Prevention C of DC and General Equivalence Mapping Files**. 2018. Accessed January 27, 2025. https://ftp.cdc.gov/pub/health_statistics/nchs/publications/ICD10CM/2018/.
22. **World Health Organization**. ICD-10: International statistical classification of diseases and related health problems: Tenth revision. 2nd ed. World Health Organization; 2004. Accessed January 27, 2025. https://iris.who.int/handle/10665/42980
23. **World Health Organization**. International classification of diseases : [9th] ninth revision, basic tabulation list with alphabetic index. World Health Organization; 1978. Accessed January 27, 2025. https://iris.who.int/handle/10665/39473
24. **Apache**. Apache parquet documentation. 2022. Accessed January 27, 2025. https://parquet.apache.org/docs/
25. **Regenstrief Institute I**. LOINC. 2023. Accessed January 27, 2025. https://loinc.org/
26. **Nelson SJ, Zeng K F, Kilbourne J F, Powell T F, Moore R.** Normalized names for clinical drugs: RxNorm at 6 years. (1527-974X (Electronic)). 2011.
27. **Services C for M& M**. HCPCS - General information. *Medicare*. 2012. Accessed January 27, 2025. http://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html
28. **SNOMED International**. 2024. Accessed January 27, 2025. https://www.snomed.org/.
29. **Dongur L, Samman Y, Golovko G**, ***et al.*** Cancer incidence following bariatric surgery in renal transplant recipients: a retrospective multi-center analysis. *Surg Obes Relat Dis*. Published online 2024;**20**(12):1198–1205.
30. **Zou J, Nelson N, Botnar K, Khanipov K, Klimberg VS.** Clinical trends in granulomatous mastitis incidence, prevalence, and treatment: a retrospective study highlighting ethnic differences in care. *J Surg Res*. 2024;**302**: 732–738.
31. **Tsukagoshi JWB.** *Effect of BMI on the Outcomes of Patients with Pulmonary Embolism Requiring Endovascular Interventions*. In: Vascular Annual Meeting.2024.
32. **Gohil A, Shroff A, Garg A, Kumar S.** A compendious research on big data file formats. In: *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). Published online*, 2022:905–913.
33. **Plase D, Niedrite L, Taranovs R.** Accelerating data queries on Hadoop framework by using compact data formats. In: *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2016:1–7.

https://doi.org/10.1017/cts.2025.55 Published online by Cambridge University Press