A Query Engine for Self-controlled Case Series, with an application to COVID-19 EHR data

Xiaojin Li, PhD[1,3,⋆], Yan Huang, PhD[1,3], Licong Cui, PhD[2,3], Guo-Qiang Zhang, PhD[1,2,3]
[1]**McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX**
[2]**School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX**
[3]**Texas Institute for Restorative Neurotechnologies, The University of Texas Health Science Center at Houston, Houston, TX**
⋆ Corresponding author

Abstract

Self-controlled case series (SCCS) is a statistical method in epidemiological study design that uses individuals as their own controls, with comparisons made within the same individuals at different time points of observation. SCCS has been applied in settings where it is difficult to identify comparison or control groups. To provide computational support for SCCS, we introduce a query engine called Self-Controlled Case Query (SCCQ) and use it to extract cohorts of self-controlled case series from a large-scale COVID-19 Electronic Health Records (EHR) dataset. Visual summary of the queried population through the R-Shiny visualization framework offers SCCQ's query result dashboard to the researcher. SCCQ allows the export of query-generated raw data files with a portable format that researchers can extend to create more intricate and robust visualization capabilities without needing a high-level of technical or statistical background. Our validation and evaluation experiments uncovered COVID-19 outcomes to be consistent with existing research findings. With SCCQ, cohort exploration, data extraction, and information visualization can be provided for structured EHR data to lower the barrier for clinical and epidemiological research.

1 Introduction

Electronic Health Records (EHR) is intended as a comprehensive, up-to-date record of a patient's healthcare information in a digital format. These records can be shared across different healthcare settings, and they typically include a range of data types, including demographics, medication, diagnoses, vital signs, past medical history, immunization status, radiology reports, and laboratory and test results. EHR has been broadly adopted in the United States in the last two decades for improving the quality of care, enhancing patient privacy and security, improving efficiency and productivity, and reducing healthcare costs[1–3]. The wide adoption of EHR resulted in a rapid growth of clinical data available electronically[4]. This growth offers a unique opportunity for the secondary analysis of clinical data in research beyond its original purpose of documenting care, and creative analysis of EHR data can generate insight for improving care and informing healthcare policy making[5–10].

However, EHR data can be large, complex, biased, and maybe stored in disparate databases or repositories[11]. Healthcare organizations face increasing challenges in making EHR data suitable for secondary analysis[12]. Indeed, even where data is effectively incorporated and made accessible, extracting analysis-ready data (i.e., there is no additional preprocessing is needed when using the data for statistical analysis or for generating AI models) can be time-consuming and difficult for researchers without a background in informatics or data science. As EHR systems become more widespread, there is a pressing data-science need for effective support in making sense of data. Interactive visualization is a way to improve the understanding of complex data, thereby benefiting the generation of insights from EHR data[11].

EHR data provides an information source for epidemiological studies such as cohort and case-control studies which attempt to ascertain the strength of the association between specific exposures and health outcomes and have been widely applied in clinical research[13]. Challenges in applying standard epidemiological study designs to EHR data include aspects such as: 1) difficult to identify suitable comparison/control groups, e.g., when investigating adverse effects to vaccines and medications[14]; 2) difficult to provide effect estimates and adjust confounders such as body mass index and comorbidities[15], understanding that existing studies applied different strategies to reduce confounding factors such as multivariate analysis[16] and propensity score matching[17,18]; and 3) high-level computational burden in

statistical modeling involving tens of millions of individuals observed over multiple years[19].

Self-controlled case series (SCCS) uses intra-person comparisons in a population of individuals with both the outcome and exposure of interest[20]. Only exposed individuals who have experienced an event are included. This strategy is advantageous in that virtually all time-invariant confounding factors are eliminated. It offers an alternative to more established methods of cohort or case-control studies for identifying associations between time-varying exposures and outcomes[21]. SCCS was initially designed to investigate associations between vaccination and acute potential adverse events[22,23]. Subsequently, the method was applied in several other areas, including pharmacoepidemiology and epidemiology such as in studying autism[20]. To use the SCCS method for EHR data, it is necessary to perform and combine different types of temporal queries, such as absolute temporal query and relative temporal query[24,25]. However, EHR data exploration/query systems involve substantial technical obstacles for their end-users without proper interface design. Some systems need their users to search by using command-line-based queries, such as SQL or its extensions, which are difficult to learn, and it is conceptually complex to describe temporal queries[11]. Thus, despite the fact that temporal information is essential for clinical research, there is a general lack of systematic and dedicated methodology support for temporal queries that cover backend data model design, query language, and user interfaces[24].

We introduce Self-Controlled Case Query (SCCQ) engine to fill these gaps in cohort identification and cohort information visualization involving the SCCS method for EHR data. SCCQ extracts clinical cohorts with any SCCS queries and provides visualization from a large-scale COVID-19 EHR dataset so that as to make it easier for researchers to make sense of data without requiring a high-level informatics or statistical background. The SCCS query builder interface provides researchers with a straightforward way to perform initial cohort exploration. Users can benefit from cohort information visualization by leveraging their ability to recognize patterns and connections which can make the discovery of deeper details easier[26–28]. In the fields of healthcare, medical informatics, and imaging informatics, visualization has been well studied but is relatively new to clinical research informatics[29]. It has been demonstrated that visualization can reduce task completion times, but its impact on finding and retrieving individual patient records is uncertain and highly dependent on the visualization interface[29]. We create SCCQ to assist researchers in constructing a query and retrieving a population. Once a population is retrieved, the interface can support its visualization and assist researchers in understanding population characteristics. SCCQ seeks to create a paradigm with the ability to combine interactive visualization and on-the-fly hypothesis exploration, thus increasing the productivity of researchers by lowering the barriers for sense-making of data.

2 Background

2.1 Self-controlled case series method

Based on a retrospective cohort model, the self-controlled case-series method is applied to a defined observation period, conditionally on the number of events experienced by each individual over the observation period[30]. Within the observational period, the time is classified as at-risk or as control time in relation to exposures that are considered to be fixed, and it is based on within-person comparisons in individuals with both the outcome and the exposure of interest[14,15]. Incidence rate ratios are derived, comparing the rate of events during exposed periods of time with the rate during all other observed time periods. The key advantage is that controls for all fixed confounders, measured or otherwise, and allows for age-variation in the baseline incidence of events. Thus, it removes the potential confounding effect of characteristics that vary between individuals, such as frailty and risk factors for vascular disease[31,32]. Different kinds of SCCS methods have been proposed in the last few years[14,33]. However, most of these methods are difficult to apply, which limits their usage by researchers whose primary expertise is not statistics. Several software packages (R, STATA, and SAS)[34,35] have been developed to allow researchers to utilize these techniques more easily.

2.2 R-Shiny visualization framework

R-Shiny[36] is an open package in R, which provides a web application framework to create interactive visualizations of data. It provides extensive pre-built and customizable output widgets for displaying plots, tables, and printed output of R objects, which save time in the construction, automation, and distribution of data visualizations and statistical analyses. R-Shiny has been widely used to develop interactive dashboards and web applications with EHR data[37–39].

3 Methods

We design the system architecture (as shown in Figure 1) to be comprised of 4 core architectural elements: 1) a web-based interface written in React[40] (Figure 1.A), which includes SCCS query builder, cohort information visualization and data exporting. The query builder is a powerful and intuitive interface that has been designed and developed to enable researchers to quickly create the SCCS criteria and perform exploratory queries. 2) A query engine for searching EHR data implemented with Ruby on Rails[41] (Figure 1.B). Such a query engine translates user input-based queries into executable database query languages, which consists of three parts, a query translation module, a query execution module, and a data reorganization module. 3) A MongoDB database for storing EHR data based on Event-level Inverted Index (ELII)[24], which provides good performance for temporal queries (Figure 1.C), and 4) a small application database with MongoDB for tracking queries, storing query results for visualization, and logging activity. The purpose of this small and independent database is to separate the system application data and EHR data so that they will not affect each other when upgrading the system or updating the data.

3.1 Data and backend database

As an application of SCCQ, we use the OPTUM® de-identified COVID-19 Electronic Health Record dataset, which is drawn from dozens of healthcare providers in the United States, including more than 700 hospitals and 7,000 clinics. It includes EHR data for 7 million unique individuals who have documented clinical care with a documented diagnosis of COVID-19 or acute respiratory illness after 02/01/2020 and/or documented COVID-19 testing regardless of their results. The data incorporates a wide swath of raw clinical data, including new, unmapped COVID-specific clinical data points from both inpatient and ambulatory electronic medical records, which include patient-level information: demographics, diagnoses, procedures, lab tests, care settings, medications prescribed or administered, and mortality. These data are certified as de-identified by independent statistical experts in accordance with Health Insurance Portability and Accountability Act (HIPAA) statistical de-identification rules and are managed in accordance with the OPTUM® customer data use agreement.

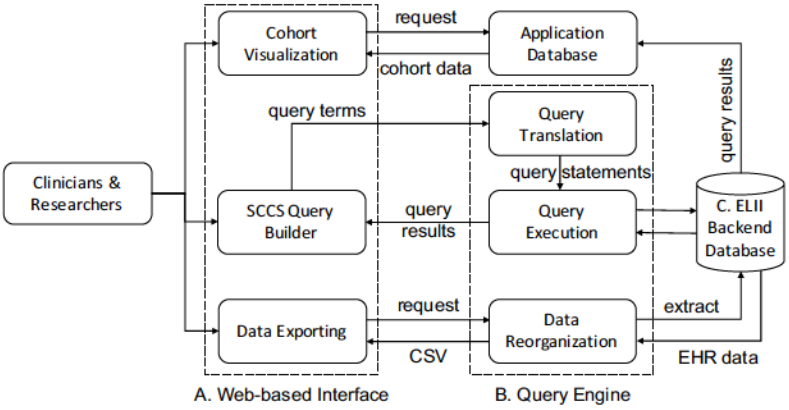

Figure 1: The system architecture of SCCQ.

We build the ELII with the OPTUM® COVID-19 EHR data. ELII consists of 4 components[24]: 1) conventional inverted index, which contains the inverted indices of time-invariant variables, 2) timeline inverted index, which consists of inverted indices of time-dependent variables, 3) patient timeline, which includes all clinical events and related information for each patient, and 4) a global lookup table, which contains forward indices of all variables and associated inverted indices. With such a design, the performance of temporal queries can be significantly improved (26-88 times improvement)[24]. Leveraging the advantages of ELII in temporal queries, we design and implement a query engine for SCCS query and further data exporting.

We choose MongoDB as the backend database since it has the following advantages: 1) good query performance with the large-scale dataset, 2) flexible data models so that we can build our own models for the inverted index to improve the query performance, especially for temporal queries, 3) highly and easily scalable, and we can easily scale-up by adding commodity servers, and 4) easy for developers and leading to faster development time and fewer bugs.

3.2 SCCS query engine

The SCCS query for cohort exploration of a specific clinical study (e.g., the risk of stroke after diagnosis of COVID-19) is illustrated in Figure 2. In Figure 2(a), there are two events, named index event (exposure) and outcome event

(outcome), and an SCCS query uses the date of index event as the index date and finds patients who had outcome event within a given observation window before and after the index date. The index date is different for each patient, and the period before the index date is considered the control period, while the period after the index date is treated as the exposure period.

As users add query terms, such as index event, time period of index event (e.g., start date and end date), outcome event, and the observation window, the query is then generated. The query builder interface creates an array of objects (key-value pairs) in JavaScript Object Notation (JSON) format that represent the current user interface state, the query definition, and additional metadata describing the query. The query translation module automatically translates the user's query into actual MongoDB statements, and then the query execution module sends the translated MongoDB statements to the corresponding data sources to execute the query. Leveraging the advantages of ELII design, SCCQ can return the list of unique patient identifiers instead of returning the patient counts. Such a list offers our interface extra features for simple demographic visualization and row-level cohort data access with a single click.

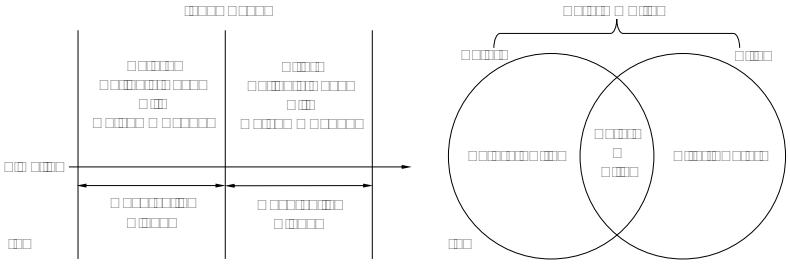

Figure 2: The SCCS query for cohort exploration.

After the SCCS query, the patients are categorized into two groups: 1) *Before* group, which contains all patients who had outcome event before the index event, and 2) *After* group, which includes all patients who had outcome event after the index event. Figure 2(b) shows the Venn diagram to visually represent the logic relationship between two groups: 1) *Before+After*, which contains all patients who had outcome event before or after the index event, 2) *Before-After*, which includes all patients who only had outcome event before the index event, 3) *Before&After*, which consists of all patients who had outcome event before and after the index event, and 4) *After-Before*, which contains all patients who only had outcome event after the index event. Thus, we have a total of 6 different patient groups, each of which can be exported and visualized independently.

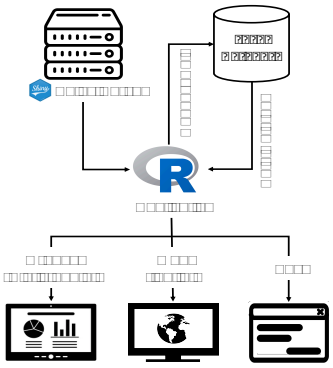

Figure 3: The workflow of R-Shiny application.

3.3 Cohort information visualization

The results of SCCS queries are saved to the local database and can be exported to CSV files. The CSV files provide the detailed data necessary for analysis and the query results saved in the local database will be the data source for the visualizations. With dedicated R packages, including plotly, ggplot, and leaflet, we develop several R-Shiny applications (as shown in Figure 3), which contain different modalities for interactive visualization of cohort information, such as demographic distributions, diagnosis, and medication history. Such visualizations allow the researcher to get an overview of the information of extracted cohort in the first place, so that query details can be adjusted and refined in the further step of the study. Another benefit is that the researcher may identify new patterns and conduct research accordingly. For highly motivated researchers, new visualizations can be added by leveraging the same data in the local database or by using the CSV files in an analysis software package of their choosing.

4 Experiments and Results

4.1 Data repository

Over 30 billion records (with dataset version of 20210916) were processed from more than 2.5 TB of raw text files and stored in our MongoDB database, including demographics, diagnoses, procedures, lab tests, care settings, and

medications. Table 1 presents a summary of the demographic of the patients in the database. There were about 7 million patients in our database, of which 56.29% were female, 43.60% were male, and 0.11% had unknown gender. The age was normally distributed, with the majority (81.65%) of patients between 20 and 80 years of age, 12.26% of patients were younger than 20 years old, and 6.07% of patients were older than 80 years old. The race distribution showed that 72.89% of patients were Caucasian, 10.93% were African American, 2.18% were Asian, and 14% were other/unknown. The ethnicity distribution demonstrated that 77.25% of patients were not Hispanic, 6.86% were Hispanic, 15.89% were unknown. The region distribution presented that 45.35% of patients were from Midwest, 23.80% were from the Northeast, 16.22% were from the South, 10.22% were from the West, and 4.41% were from other regions.

4.2 SCCS query builder interface

Table 1: The summary of OPTUM® COVID-19 EHR data.

Category		Number of Patients	Percentage (%)
Gender	**Female**	3,919,943	56.29
	Male	3,036,319	43.60
	Unknown	7,512	0.11
Age	**less than 20 years old**	853,768	12.26
	20-80 years old	5,685,793	81.65
	more than 80 years old	423,143	6.07
Race	**Caucasian**	5,075,755	72.89
	African American	760,830	10.93
	Asian	151,570	2.18
	Other/Unknown	975,619	14.00
Ethnicity	**Not Hispanic**	5,379,257	77.25
	Hispanic	477,760	6.86
	Unknown	1,106,757	15.89
Region	**Midwest**	3,158,067	45.35
	Northeast	1,657,653	23.80
	South	1,129,698	16.22
	West	711,323	10.22
	Other/Unknown	307,033	4.41
COVID-19 Confirmed		990,842	14.23
COVID-19 Vaccinated	**Pfizer (only)**	793,729	11.40
	Moderna (only)	504,944	7.25
	Janssen	57,856	0.83
	Moderna and Pfizer	5,361	0.08
	AstraZeneca	714	0.01

SCCQ allows the expression of SCCS criteria as variables/query terms using SCCS query builder interface, which can be executed as a one-time search to estimate the eligible patient cohort. Figure 4 shows the SCCS query builder interface with three areas annotated. In the outcome event area (Figure 4.A), researchers can define the outcome diagnosis event with the 10th revision of the International Classification of Diseases (ICD-10) code, for example, R430 was used to define the outcome event as anosmia. Index event area (Figure 4.B) provided three pre-defined events, in the corresponding drop-down menu, for COVID-19 related studies: 1) COVID-19 PCR-positive, returning patients with positive COVID-19 PCR test results, 2) COVID-19 vaccination, returning patients who have received at least one dose of COVID-19 vaccine (regardless of vaccine brand), and 3) influenza vaccination, returning patients who have been vaccinated against influenza. For COVID-19 PCR-positivity and COVID-19 vaccination, we used the first PCR-positive/vaccination date as the index date, while for influenza, the most recent date was considered as the index date. As shown in Figure 4.B, it supported multiple index events, each of which was treated as a separate study for SCCS analysis (i.e., there is no relationship between the index events). The purpose of this feature is to facilitate users to compare the SCCS results of different index events. Currently, SCCQ only supports SCCS analysis of a single index event, and SCCS analysis involving multiple index events will be supported in future work. The observation window size was defined in the observation window area (Figure 4.C). Eventually, the SCCS query in Figure 4 can be translated to "Find COVID-19 PCR-positive patients with the diagnosis of anosmia within 30 days before and after their first PCR-positive date between March 1st, 2020 and March 1st, 2021". SCCQ offered two alternative strategies for censored patients (i.e., patients without sufficient observational data): remove or adjust (Figure 4.C). The remove strategy simply removed incomplete data, and adjust strategy applied the inverse-probability-of-censoring weighting[42] for adjustment to achieve a more accurate assessment of variability across control and exposure periods.

The query results displayed the number of patients for each group (Figure 4.D). An R-Shiny application for cohort information visualization will be launched by clicking on the corresponding chart icon. The cloud-download icon allowed researchers to export the EHR data for patients in the corresponding group. We calculated two rates using group *Before* and group *After-Before* to provide rough estimates of the probability of an outcome event occurring before and after the index event, respectively. As shown in Figure 4, there were 15.44% patients diagnosed with

Figure 4: The SCCS query builder interface.

anosmia before the COVID-19 PCR-positive date, while 84.56% of patients were newly diagnosed with anosmia (i.e., no prior anosmia history) after the COVID-19 PCR-positive date. Such a high proportion of newly diagnosed anosmia patients indicated a high correlation between COVID-19 and anosmia.

4.3 R-Shiny interactive visualization

We developed two modalities to fulfill the most common visualization needs: discovering patterns/trends. Figure 5 shows the modality for single cohort information interactive visualization including demographic, top 20 diagnosis, top 20 procedures, and top 20 medications. In the demographic visualization, by clicking on the graph, the researcher can drill-down to a specific sub-cohort and all visualizations will be updated with the selected sub-cohort. For example, the bottom of Figure 5 shows demographic information for the subgroup that included only female Caucasians, while clicking on the top tab removes the selected condition. Figure 6 shows the modality for interactive visualization of cohort information comparison. Each column represented different cohorts, and each row was the item to be compared. As shown in Figure 6, it compared the demographic distributions, top 20 diagnosis codes and top 20 procedures between group *Before* and group *After*. All modalities are extendable and interchangeable so that additional visualizations can be implemented as needed.

Table 2: COVID-19 outcome case studies using SCCQ.

Disease/Symptom	ICD-10	(After-Before)/(After+Before)
Anosmia	R430	84.56%
Cough	R05	82.37%
Fever	R509	82.13%
Sore throat	J029	80.04%
Fatigue	R5383	79.38%
Headache	R51	79.31%
Renal failures	N179	78.83%
Myocardial infarction	I21	78.48%
Chest pain	R07	77.45%
Pulmonary fibrosis	J841	73.76%
Acute ischemic stroke	R07	63.80%
Eye problem	H579	62.16%
Random Selection	200 ICD-10 Codes	50.94%
Pregnancy	Z34	31.67%

4.4 COVID-19 outcome case study

We performed multiple COVID-19 outcome case studies with existing diseases or symptoms that have been demonstrated to be associated with COVID-19 to validate the effectiveness of our SCCQ query engine. Symptoms listed by Centers for Disease Control and Prevention (CDC)[43] included fatigue (ICD-10: R5383), cough (ICD-10: R05), fever (ICD-10: R509), chest pain (ICD-10: R07), anosmia (ICD-10: R430), headache (ICD-10: R51), and sore throat (ICD-10: J029). COVID-19 related diseases included pulmonary fibrosis (ICD-10: J841)[44], renal failure (ICD-10:

Figure 5: The cohort information visualization.

N179)[45], myocardial infarction (ICD-10: I21)[46], acute ischemic stroke (ICD-10: I63)[47,48], and eye diseases (ICD-10: H579)[49]. We randomly selected 200 ICD-10 codes as outcomes, kept all other settings unchanged, performed SCCS analysis, and took the average value as a reference for comparison. With such a reference, it facilitated the determination of whether the queried disease is associated with COVID-19. We also added pregnancy (ICD-10: Z34) as an event unrelated to COVID-19 as a negative test. As shown in Table 2, the results were consistent with existing published studies with higher rates of both COVID-19 symptoms and COVID-19 associated diseases. The lower rate of pregnancy indicated that it was not associated with COVID-19.

5 Discussion

Since the COVID-19 epidemic, the size of the dataset used in our study has increased rapidly, from 0.6 million patients in June 2020 to 7 million patients with the latest release, and the volume of raw data has grown from 200 GB to 2.5 TB. At the same time, there is a dramatic increase in research topics related to COVID-19. The primary goal is to utilize a large-scale longitudinal observational database containing time-stamped patient-level medical information (including COVID-19 exposure periods and diagnosis dates for millions of patients) to investigate the relationship between COVID-19 as time-varying and adverse events associated with health outcomes. Statistical methods for such large-scale dataset have to deal with computational challenges related to data size and also need to address confounding and other biases that can undermine estimates of effect sizes[19]. Since each individual serves as its own control, the SCCS indirectly controls for fixed baseline covariates. Furthermore, the analysis requires only exposed patients, which is computationally advantageous. Therefore, we applied the SCCQ on this OPTUM® COVID-19 EHR data.

In the traditional paradigm, once clinical investigators attempt to perform a preliminary analysis with EHR data, they need to go through multiple procedures, including defining the cohort, submitting a data request to the database administrators to export the data, and working with the data analysts to obtain the results. Each step may involve several rounds of revisions. The whole process can take days or even weeks. SCCQ seeks to create a paradigm that gives clinical investigators and data analysts a direct access to data and exploration tools and then conducts collaborative data exploration. For clinical investigators, it reduces the time it takes to access data, breaking the limitations of the previous data access paradigm, from what could take days or even weeks to having data available in minutes. On the other hand, in SCCQ, the entire process of generating cohorts based on the SCCS design is automated, which has greatly improved the ability to quickly release data to researchers with large-scale EHR data and reduced the burden on data analysis teams. The ability to quickly access data, combined with information visualization tools, can increase

Figure 6: The visualization for cohort information comparison.

the productivity of researchers in different studies, thus leading to shorter research cycles.

There is more than one strategy for visualizing information of EHR data. One alternative we have designed is to use existing libraries such as D3 (Data-Driven Documents)[50] or Highcharts[51] to develop visualization interfaces. However, the development of dedicated software is undoubtedly an expensive process. The difficulty of development increases exponentially as the number of visualization types increases and the need for more interactive visualizations arises, which is not scalable and extendable. We avoid the need for complicated programming visualizations for the web application by designing visualization modalities within the R-Shiny ecosystem, which removes the burden of designing visualizations compatible with different web browsers, platforms, and devices. This should improve the researchers' experience by providing a consistent and professional visualization environment. R-Shiny is widely used in the research field because it provides a professional visualization library, is easy to use, does not require a strong programming background, and is free and open source. Our visualization modalities based on R-Shiny can be reused for different studies, and other researchers can also use R-Shiny to do research-specific analysis on CSV files exported from SCCQ. In addition to visualization, the R framework also provides powerful statistical analysis capabilities. Leveraging these advantages, we will develop more powerful and integrated applications that combine visualization and statistical analysis in the future.

SCCQ is designed and implemented on top of ELII[24], which is specially designed for fast temporal query on large EHR-derived data sources and has a high scalability and generalizability. SCCQ inherits these features from ELII and is applicable to different structured EHR data.

Limitations. Many methods exist for designing SCCS for studies with different purposes, and we have implemented only one of them in our application. In future work, we will offer a variety of SCCS methods to be selected, add more options of index events, and add more statistics on the results in the future. For outcome events, to improve usability, we will build a special database and import all ICD codes into it. Users can search and select ICD codes by keywords instead of entering them directly. For censored patients, we will apply different adjustment methods. In terms of visualization, preliminary results only presented basic information, such as demographics. In future work, we will extract more clinical data such as BMI, HbA1c, metabolic panel, and blood pressure and visualize them through

different chart types such as box plots and stacked bar charts.

6 Conclusion

We introduced SCCQ, a query engine to meet data exploration needs for SCCS queries. SCCQ supports both interactive cohort information visualization and raw data records export to assist the researcher in effectively exploring and understanding EHR data. Our preliminary validation and evaluation experiments using a large-scale COVID-19 EHR dataset shows that SCCQ is a unique tool providing comprehensive support for cohort exploration, data extraction, and information visualization for structured EHR data, thus lowering the barrier for clinical and epidemiological research that uses the SCCS method.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) grant 2047001 and the National Institutes of Health (NIH) grants R01LM013335 and R01NS126690. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF or NIH.

References

1. Casey J. A, Schwartz B. S, Stewart W. F, and Adler N. E. Using electronic health records for population health research: a review of methods and applications. *Annual review of public health*. 2016; 37: 61–81.
2. Tao S, Lhatoo S, Hampson J, Cui L, and Zhang G.-Q. A bespoke electronic health record for epilepsy care (epitome): Development and qualitative evaluation. *Journal of Medical Internet Research*. 2021; 23(2): e22939.
3. Abul-Husn N. S and Kenny E. E. Personalized medicine and the power of electronic health records. *Cell*. 2019; 177(1): 58–69.
4. Meystre S. M, Lovis C, Bürkle T, Tognola G, Budrionis A, and Lehmann C. U. Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics*. 2017; 26(01): 38–52.
5. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates D. W, and Middleton B. The value of health care information exchange and interoperability: There is a business case to be made for spending money on a fully standardized nationwide system. *Health affairs*. 2005; 24(Suppl1): W5–10.
6. Thakkar M and Davis D. C. Risks, barriers, and benefits of ehr systems: a comparative study based on size of hospital. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*. 2006; 3.
7. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton S. C, and Shekelle P. G. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of internal medicine*. 2006; 144(10): 742–752.
8. Menachemi N and Collum T. H. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*. 2011; 4: 47.
9. Nicola M, Sohrabi C, Mathew G, Kerwan A, Al-Jabir A, Griffin M, Agha M, and Agha R. Health policy and leadership models during the covid-19 pandemic: A review. *International journal of surgery*. 2020; 81: 122–129.
10. Tran D. M, Thwaites C. L, Van Nuil J. I, McKnight J, Luu A. P, Paton C, and (VITAL) V. I. T. A. L. Digital health policy and programs for hospital care in vietnam: Scoping review. *Journal of medical Internet research*. 2022; 24(2): e32392.
11. Rind A, Wang T. D, Aigner W, Miksch S, Wongsuphasawat K, Plaisant C, and Shneiderman B. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*. 2013; 5(3): 207–298.
12. Dobbins N. J, Spital C. H, Black R. A, Morrison J. M, De Veer B, Zampino E, Harrington R. D, Britt B. D, Stephens K. A, Wilcox A. B, et al. Leaf: an open-source, model-agnostic, data-driven web application for cohort discovery and translational biomedical research. *Journal of the American Medical Informatics Association*. 2020; 27(1): 109–118.
13. Buka S. L, Rosenthal S. R, and Lacy M. E. Epidemiological study designs: traditional and novel approaches to advance life course health development research. *Handbook of life course health development*. 2018; p. 541–560.
14. Petersen I, Douglas I, and Whitaker H. Self controlled case series methods: an alternative to standard epidemiological study designs. *bmj*. 2016; 354.
15. Douglas I. J, Evans S. J, Pocock S, and Smeeth L. The risk of fractures associated with thiazolidinediones: a self-controlled case-series study. *PLoS Medicine*. 2009; 6(9): e1000154.
16. Chatziralli I, Kabanarou S. A, Parikakis E, Chatzirallis A, Xirou T, and Mitropoulos P. Risk factors for central serous chorioretinopathy: multivariate approach in a case-control study. *Current Eye Research*. 2017; 42(7): 1069–1073.
17. Sharma P, Chen V, Fung C. M, Troost J. P, Patel V. N, Combs M, Norman S, Garg P, Colvin M, Aaronson K, et al. Covid-19 outcomes among solid organ transplant recipients: a case-control study. *Transplantation*. 2021; 105(1): 128–137.
18. Bauer A. Z, Gore R, Sama S. R, Rosiello R, Garber L, Sundaresan D, McDonald A, Arruda P, and Kriebel D. Hypertension, medications, and risk of severe covid-19: a massachusetts community-based observational study. *The Journal of Clinical Hypertension*. 2021; 23(1): 21–27.

19. Simpson S. E, Madigan D, Zorych I, Schuemie M. J, Ryan P. B, and Suchard M. A. Multiple self-controlled case series for large-scale longitudinal observational databases. *Biometrics*. 2013; 69(4): 893–902.
20. Di Bartolomeo S, Marino M, Guastaroba P, Valent F, and De Palma R. Self-controlled case-series study to verify the effect of adherence to beta-blockers in secondary prevention of myocardial infarction. *Journal of the American Heart Association*. 2015; 4(1): e001575.
21. Whitaker H. J, Paddy Farrington C, Spiessens B, and Musonda P. Tutorial in biostatistics: the self-controlled case series method. *Statistics in medicine*. 2006; 25(10): 1768–1797.
22. Farrington C. Relative incidence estimation from case series for vaccine safety evaluation. *Biometrics*. 1995; p. 228–235.
23. Farrington C, Nash J, and Miller E. Case series analysis of adverse reactions to vaccines: a comparative evaluation. *American journal of epidemiology*. 1996; 143(11): 1165–1173.
24. Huang Y, Li X, and Zhang G.-Q. Elii: A novel inverted index for fast temporal query, with application to a large covid-19 ehr dataset. *Journal of Biomedical Informatics*. 2021; 117: 103744.
25. Zhang G.-Q, Li X, Huang Y, and Cui L. Temporal cohort logic. *AMIA Annual Symposium Proceedings*. 2022.
26. Chittaro L. Information visualization and its application to medicine. *Artificial intelligence in medicine*. 2001; 22(2): 81–88.
27. Shneiderman B, Plaisant C, and Hesse B. W. Improving healthcare with interactive visualization. *Computer*. 2013; 46(5): 58–66.
28. West V. L, Borland D, and Hammond W. E. Innovative information visualization of electronic health record data: a systematic review. *Journal of the American Medical Informatics Association*. 2015; 22(2): 330–339.
29. Harris D. R and Henderson D. W. i2b2t2: unlocking visualization for clinical research. *AMIA Summits on Translational Science Proceedings*. 2016; 2016: 98.
30. El-Gilany A. Self-controlled case series study (sccss): a novel research method; 2019.
31. Douglas I. J and Smeeth L. Exposure to antipsychotics and risk of stroke: self controlled case series study. *Bmj*. 2008; 337.
32. Musonda P, Paddy Farrington C, and Whitaker H. J. Sample sizes for self-controlled case series studies. *Statistics in medicine*. 2006; 25(15): 2618–2631.
33. Nie X, Xu L, Bai Y, Liu Z, Liu Z, Farrington P, and Zhan S. Self-controlled case series design in vaccine safety: a systematic review. *Expert Review of Vaccines*. 2022; 21(3): 313–324.
34. Farrington P, Whitaker H, and Weldeselassie Y. G. *Self-controlled case series studies: a modelling guide with R*. Chapman and Hall/CRC; 2018.
35. Self-controlled case series studies. https://sccs-studies.info/index.html.
36. RStudio, Inc. *shiny: Web Application Framework for R*; 2014. URL: `http://shiny.rstudio.com`.
37. Estiri H and Stephens K. Dqe-v: a database-agnostic framework for exploring variability in electronic health record data across time and site location. *eGEMs*. 2017; 5(1).
38. Beesley L. J, Fritsche L. G, and Mukherjee B. A modeling framework for exploring sampling and observation process biases in genome and phenome-wide association studies using electronic health records. *bioRxiv*. 2019; p. 499392.
39. Mayer D. A, Rasmussen L. V, Roark C. D, Kahn M. G, Schilling L. M, and Wiley L. K. Reviewr: A light-weight and extensible tool for manual review of clinical records. *medRxiv*. 2021.
40. Rawat P and Mahajan A. N. Reactjs: A modern web development framework. *International Journal of Innovative Science and Research Technology*. 2020; 5(11).
41. Hansson D. H and Team R. C. Ruby on rails. *Development*. 2009; 4: 0.
42. Dong G, Mao L, Huang B, Gamalo-Siebers M, Wang J, Yu G, and Hoaglin D. C. The inverse-probability-of-censoring weighting (ipcw) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring. *Journal of biopharmaceutical statistics*. 2020; 30(5): 882–899.
43. Symptoms of covid-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.
44. George P. M, Wells A. U, and Jenkins R. G. Pulmonary fibrosis and covid-19: the potential role for antifibrotic therapy. *The Lancet Respiratory Medicine*. 2020; 8(8): 807–815.
45. Adapa S, Aeddula N. R, Konala V. M, Chenna A, Naramala S, Madhira B. R, Gayam V, Balla M, Muppidi V, and Bose S. Covid-19 and renal failure: challenges in the delivery of renal replacement therapy. *Journal of clinical medicine research*. 2020; 12(5): 276.
46. Katsoularis I, Fonseca-Rodríguez O, Farrington P, Lindmark K, and Connolly A.-M. F. Risk of acute myocardial infarction and ischaemic stroke following covid-19 in sweden: a self-controlled case series and matched cohort study. *The Lancet*. 2021; 398(10300): 599–607.
47. Modin D, Claggett B, Sindet-Pedersen C, Lassen M. C. H, Skaarup K. G, Jensen J. U. S, Fralick M, Schou M, Lamberts M, Gerds T, et al. Acute covid-19 and the incidence of ischemic stroke and acute myocardial infarction. *Circulation*. 2020; 142(21): 2080–2082.
48. Yang Q, Tong X, George M. G, Chang A, and Merritt R. K. Covid-19 and risk of acute ischemic stroke among medicare beneficiaries aged 65 years or older: self-controlled case series study. *Neurology*. 2022; 98(8): e778–e789.
49. Hu K, Patel J, Swiston C, and Patel B. C. Ophthalmic manifestations of coronavirus (covid-19). *StatPearls [Internet]*. 2021.
50. Bostock M, Ogievetsky V, and Heer J. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*. 2011; 17(12): 2301–2309.
51. Mishra S. Getting started with highcharts. In *Practical Highcharts with Angular*; p. 1–14. Springer; 2020.