CJC Open 3 (2021) 639–645

Study Design

CREATE: A New Data Resource to Support Cardiac Precision Health

Seungwon Lee, MPH,[a,b,c,d] Bing Li, MA,[a,c] Elliot A. Martin, PhD,[a,c] Adam G. D'Souza, PhD,[a,c] Jason Jiang, MSc,[a,c] Chelsea Doktorchik, MSc,[a,b] Danielle A. Southern, MSc,[a,b] Joon Lee, PhD,[a,b,d,e] Natalie Wiebe, RN,[a,b] Hude Quan, MD, PhD,[a,b] and Cathy A. Eastwood, RN, PhD[a,b]

[a] *Centre for Health Informatics, University of Calgary, Calgary, Alberta, Canada*

[b] *Department of Community Health Sciences, University of Calgary, Calgary, Alberta, Canada*

[c] *Alberta Health Services, Calgary, Alberta, Canada*

[d] *Data Intelligence for Health Lab, University of Calgary, Calgary, Alberta, Canada*

[e] *Department of Cardiac Sciences, University of Calgary, Calgary, Alberta, Canada*

ABSTRACT

Background: The initiatives of precision medicine and learning health systems require databases with rich and accurately captured data on patient characteristics. We introduce the **C**linical **R**egistry, **A**dminis-**T**rative Data and **E**lectronic Medical Records (CREATE) database, which includes linked data from 4 population databases: **A**lberta **P**rovincial **P**roject for **O**utcome **A**ssessment in **C**oronary **H**eart Disease (APPROACH; a national clinical registry), Sunrise Clinical Manager (SCM) electronic medical record (city-wide), the Discharge Abstract Database (DAD), and the National Ambulatory Care Reporting System

RÉSUMÉ

Contexte : Les initiatives en matière de médecine de précision et les systèmes de santé apprenants ont besoin de bases de données riches et exactes sur les caractéristiques des patients. Nous présentons ici la base de données CREATE (***C**linical **R**egistry, **A**dministrative Data and **E**lectronic Medical Records*), qui regroupe les données couplées de quatre bases de données populationnelles : le registre clinique national APPROACH (***A**lberta **P**rovincial **P**roject for **O**utcome **A**ssessment in **C**oronary **H**eart Disease*), le système de gestion des dossiers médicaux électroniques SCM (Sunrise Clinical Manager, utilisé à

Precision medicine, as part of a learning health care system, is an emerging medical framework proposing that medical decisions and treatments should be tailored to each individual patient, choosing optimal interventions while minimizing resource utilization.[1] Ideally, learning health care systems should have the ability to leverage the massive amounts of patient data stored in electronic medical records (EMRs), to pick the optimal care pathway for a given patient.[2] However, this has yet to come to fruition because of difficulties associated with processing EMR data, especially in the near real-time manner required for clinical decision support. For example, without some form of high-throughput automated phenotyping of patient characteristics, it is challenging to group similar patients in real time to predict the optimal care pathway for an individual.

To achieve this aim, comorbidities and disease characteristics data must be captured to pinpoint the effects and outcomes associated with particular therapies. Current population-based administrative health databases lack sufficient sociodemographic (eg, income and education level) and clinical details (eg, disease stage, severity, and treatment), making it difficult to tailor individualized treatments and prevention strategies to each patient.[3] For example, the popular Charlson comorbidity scores used in coded administrative databases do not allow consideration of disease stages. Additionally, coded data are not abstracted until after inpatient visits, making them unavailable for real-time clinical decision-making in Canada.

Received for publication July 27, 2020. Accepted December 8, 2020.

Ethics Statement: The research reported has adhered to the relevant ethical guidelines under Conjoint Health Research Ethics Board (REB19-0088).

Corresponding author: Dr Cathy A. Eastwood, Department of Community Health Sciences, Centre for Health Informatics, Cumming School of Medicine, University of Calgary, TRW Building, 3280 Hospital Drive NW, Calgary, Alberta T2N 4Z6, Canada.

E-mail: caeastwo@ucalgary.ca

See page 644 for disclosure information.

https://doi.org/10.1016/j.cjco.2020.12.019

(NACRS). The intent of this work is to introduce a cardiovascular-specific database for pursuing precision health activities using big data analytics.

Methods: We used deterministic data linkage to link SCM electronic medical record data to APPROACH clinical registry data using patient identifier variables. The APPROACH-SCM data set was subsequently linked to DAD and NACRS to obtain inpatient and outpatient cohort data. We further validated the quality of the linkage, where applicable, in these databases by comparing against the Alberta Health Insurance Care Plan registry database.

Results: We achieved 99.96% linkage across these 4 databases. Currently, there are 30,984 patients with 35,753 catheterizations in the CREATE database. The inpatient cohort contained 65.75% (20,373/30,984) of the patient sample, whereas the outpatient cohort contained 29.78% (9226/30,984). The infrastructure and the process to update and expand the database has been established.

Conclusions: CREATE is intended to serve as a database for supporting big data analytics activities surrounding cardiac precision health. The CREATE database will be managed by the Centre for Health Informatics at the University of Calgary, and housed in a secure high-performance computing environment.

l'échelle municipale), la Base de données sur les congés des patients (BDCP), et le Système national d'information sur les soins ambulatoires (SNISA). Notre objectif est d'offrir une base de données portant précisément sur les maladies cardiovasculaires, afin de soutenir les activités en santé de précision nécessitant l'analyse de mégadonnées.

Méthodologie : Nous avons utilisé une méthode de couplage déterministe pour apparier les données du système SCM à celles du registre APPROACH à l'aide de variables d'identification des patients. L'ensemble de données SCM-APPROACH a ensuite été couplé aux données de la BDCP et du SNISA, afin d'obtenir les données des cohortes des patients hospitalisés et des patients ambulatoires. Lorsque c'était possible, nous avons en outre validé la qualité du couplage en comparant les données à celles de la base de données du Régime d'assurance maladie de l'Alberta.

Résultats : Nous avons obtenu un taux de couplage de 99,96 % pour les quatre bases de données. À l'heure actuelle, la base de données CREATE compte 30 984 patients ayant subi 35 753 cathétérismes. La cohorte des patients hospitalisés représente 65,75 % (20 373/30 984) de l'échantillon, tandis que la cohorte des patients ambulatoires représente 29,78 % (9226/30 984). L'infrastructure et le processus de mise à jour et d'expansion de la base de données ont été définis.

Conclusions : La base de données CREATE est destinée à soutenir les activités d'analyse de mégadonnées nécessaires à la santé cardiaque de précision. Elle sera gérée par le Centre for Health Informatics de l'Université de Calgary et hébergée dans un environnement informatique à haut rendement sécurisé.

EMR systems collect and digitize patient health information in real time,[4] expediting clinical decision-making for individual patients to improve their quality of care. In Canada, EMRs are gradually being adopted[3,4] and are creating repositories of timely and comprehensive clinical information. Leveraging EMR data is crucial for achieving translational research in cardiovascular sciences,[5,6] and for cardiovascular precision health.[5,7] EMR data enable point-of-care clinical decision support, patient safety alerts, and cardiovascular risk estimations.[5-7] Numerous EMR-based research databases have been established internationally, often by linking databases. Examples include the **C**linic**a**l Disease Research Using **Li**nked **B**espoke Studies and **E**lectronic Health **R**ecords (CALIBER) database from the United Kingdom, which links the Clinical Practice Research Datalink (CPRD), Hospital Episode Statistics (HES), and Myocardial Ischemia National Audit Project (MINAP)[8] data, and the **E**lectronic **Me**dical **R**ecords and **Ge**nomics (eMERGE) network data in the United States, which links inpatient EMR data to genomics data.[9] These databases have been used in numerous cardiovascular studies.[10-15]

Existing stand-alone Canadian databases contain restricted details of physician claims, laboratory tests, medications, specialist consultation letters, and hospital discharges (eg, the Canadian Primary Care Sentinel Surveillance Network [CPCSSN] and **E**lectronic **M**edical **R**ecord **A**dministrative Data **L**inked **D**atabase [EMRALD]).[8,16] Clinical cardiovascular registries have informed numerous studies to date, but rely on data collected on patients' baseline characteristics, previous treatments, and outcomes.[17] Detailed clinical outcome data from EMRs are consequently needed for developing enhanced individualized risk predictions and scoring systems that can advance translational research in cardiovascular sciences. EMR and clinical registry data are often held separately, and bringing them together would provide valuable information on disease occurrence and progression.

The intent of this work is to introduce a cardiovascular disease-specific database called **C**linical **Re**gistry, **A**dminis**T**rative Data and **E**lectronic Medical Records (CREATE) and discuss potential applications for precision medicine and population public health.

Methods

This section describes the data sources and linkage process behind the CREATE database.

Data sources

EMR: Sunrise Clinical Manager. Sunrise Clinical Manager (SCM) EMR data are collected as part of routine health care practice, and the data have so far not been widely used for research purposes. SCM provides patient-level clinical information including medical and nursing orders, medication records, clinical documentation, and diagnostic imaging and lab results. SCM also includes key demographic and visit information, as well as historical chart data. SCM supports inpatient locations, outpatient clinics, and emergency departments. Details of SCM data elements are described in a previous publication.[18] The data elements of SCM included in CREATE are listed in Supplemental Table S1.

Clinical registry: APPROACH. The **A**lberta **P**rovincial **Pro**ject for **O**utcome **A**ssessment in **C**oronary **H**eart Disease (APPROACH) project has prospectively collected and assembled 3 sequential cohorts comprising approximately

170,000 patients between 1998 and 2017 in Alberta: (1) patients who underwent cardiac catheterization; (2) patients who received percutaneous coronary intervention (PCI); and (3) patients who received coronary artery bypass grafting (CABG). The cardiac catheterization cohort is the largest. A subset of this cohort consists of patients who later underwent PCI, or CABG. This cohort includes inpatients and outpatients. Extensive details of APPROACH data can be found in other journal articles[19,20] and from the APPROACH website, www.APPROACH.org. [17] A condensed list of data elements incorporated in CREATE is provided in Supplemental Table S2.

Administrative data: DAD and NACRS. The Discharge Abstract Database (DAD) captures administrative, clinical, and demographic information on inpatients at the time of discharge.[21,22] The National Ambulatory Care Reporting System (NACRS) captures all hospital-based and community-based ambulatory care data including day surgery, outpatient, urgent care, and emergency department visits.[23] The content and quality of these 2 national administrative databases are managed by the Canadian Institute for Health Information. Diagnoses are coded using the International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Canada (ICD-10-CA), and procedures and interventions using Canadian Classification of Health Interventions codes. In this work, DAD served as an inpatient indicator database and NACRS served as an outpatient indicator database. The list of data elements used for CREATE is provided in Supplemental Table S3.

Study population

We selected all cardiac catheterization patients in Calgary between April 1, 2011 and March 31, 2017. The patient data were captured from the APPROACH clinical registry,[19] in which clinical information on Albertans with diagnostic cardiac catheterization and/or revascularization procedures and patients admitted to cardiac wards is collected.

The total size of our study population was determined by the operational history of the SCM EMR, and by cardiac catheterization practices in Alberta. The SCM EMR system has been maintained and operated in the Calgary zone of Alberta Health Services (AHS), the single health authority in the province of Alberta, since 2006. It has undergone multiple design changes before full implementation in 2011.[18] There are 3 acute care hospitals in Alberta that provide cardiac catheterization: Foothills Medical Centre in Calgary, and the Royal Alexandria and University of Alberta Hospitals in Edmonton. Calgary accounts for 46.5% of the catheterizations in Alberta. SCM does not capture data from the Edmonton hospitals.

Data exploration and data cleaning

The accuracy of sex and date of birth variables in SCM and APPROACH, used as linkage variables for these 2 data sets, was checked by comparing against the Alberta Health Insurance Care Plan (Registry) data,[24] which stores information about all people who have or have previously been registered for Alberta health insurance coverage.

The SCM database contains more than 3000 tables, and approximately 700 tables contain pertinent patient information. The complex data structure was understood by: (1) making use of an assembled internal SCM dictionary; (2) consulting data users who were familiar with SCM EMR data; and (3) gaining access to the production copy of SCM EMRs to track the flow of data between the front end and the back end, and to develop the queries used to pull and link the data.

We traced the flow of SCM data from when it was entered by clinicians in the front end of the SCM EMR system (the user interface), to when it arrived in the back end (the database tables). The front end and back end of SCM are described in further detail in a previous publication.[18] AHS analysts created example records from the front end of the system and tracked the same document to the production copy of the back end data. This process allowed time stamps to be correctly tracked and ensure that final versions of the free-text documentation and other structured elements would be pulled. Structured Query Language, Python 3.8.5 (Python Software Foundation, Beaverton, OR), and SAS 9.4 (SAS Institute Inc, Cary, NC) were used for these processes.

Data linkage

Linkage of SCM EMR and APPROACH Clinical Registry. The Calgary APPROACH data were first linked with the SCM client table on the basis of the patients' personal health number (PHN), last name, first name, sex, and date of birth.

Linkage of SCM EMR and APPROACH Clinical Registry to Administrative Databases (NACRS and DAD). We linked the SCM EMR and APPROACH Clinical Registry (SCM-APPROACH) with administrative databases. For DAD, which captures data on inpatient catheterizations, we required the following criteria: (1) a match on the patient's PHN; and (2) the catheterization date had to fall between the hospital admission date minus 3 days, and discharge date plus 3 days, with the 3-day tolerance chosen to account for observed discrepancies in time stamps between the 2 sources.

In SCM-NACRS, which captures data on outpatient catheterizations, the service end date is not a mandatory data requirement for outpatient clinic visits. However, the service start date (visit date) was mandatory and was 100% complete. Therefore, we linked APPROACH and NACRS using the following criteria: (1) a match on the patient's PHN; and (2) the catheterization date falling between 3 days before and 3 days after the visit date. Figure 1 shows the CREATE cohort selection and data linkage process.

Results

Study population

The Alberta-wide APPROACH cohort between 1998 and 2017 totaled approximately 170,000 patients who received catheterization. This sample size was reduced to 31,007 patients when the patient cohort was narrowed to those who

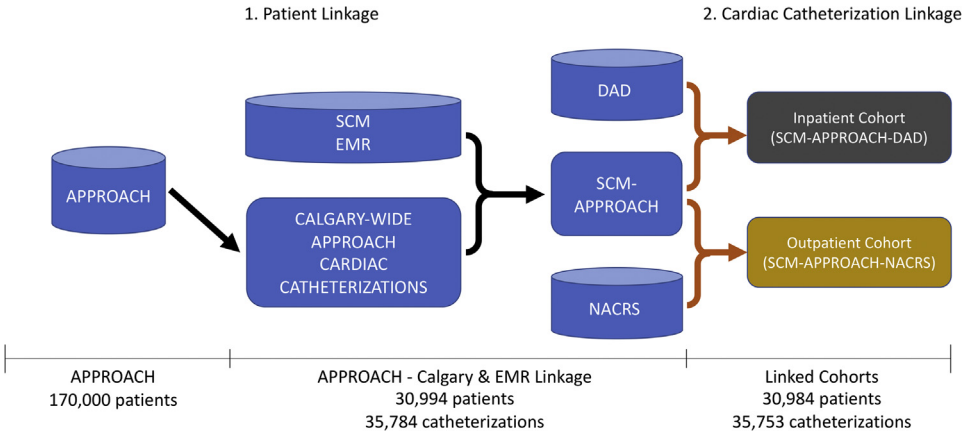

Figure 1. **C**linical **RE**gistry, **A**dminis**T**rative Data and **E**lectronic Medical Records (CREATE) cohort selection and linkage process. The **A**lberta **P**rovincial **Pr**oject for **O**utcome **A**ssessment in **C**oronary **H**eart Disease (APPROACH) clinical registry was linked to the Sunrise Clinical Manager (SCM) electronic medical record (EMR) using patient details. This linked database (SCM-APPROACH) was subsequently linked to administrative databases (Discharge Abstract Database [DAD]/National Ambulatory Care Reporting System [NACRS]) using catheterization episodes. **Cylinders** represent relational databases and **rectangles** represent flat tables.

received the procedure within acute care facilities in Calgary from 2011 to 2017.

Data exploration

Before linkage, we checked the quality of linkage variables by comparing against the registry. There were 268 patients who had inconsistencies on either sex or date of birth in the SCM and/or APPROACH compared with the registry. These records were updated with registry data.

Data linkage

Linkage of the SCM to APPROACH. Among the 31,007 patients who received a total of 35,784 catheterizations in Calgary acute care facilities and whose data were recorded in the APPROACH database, 30,994 (99.96%) were linked between APPROACH and the SCM EMR. Only 13 patients from APPROACH were not linked to the SCM EMR because 7 did not match on PHN (first level identifier) and 6 matched on PHN but did not match on the combination of last name, first name, sex, or date of birth. There were 35,768 catheterizations for those 30,994 patients.

Linkage of SCM-APPROACH to DAD and NACRS. The linkage rate using catheterization events was 99.96% (35,753/35,768) between SCM-APPROACH and DAD/NACRS. In the APPROACH registry, 24,603 catheterization records were successfully linked with DAD and 11,150 catheterization records with the NACRS database. Of the 15 catheterization records that were unlinked, 6 did not match on PHN, and 9 matched on PHN but the procedure date did not meet the criteria. The linkage steps are depicted in Figure 1.

Overview of the CREATE cohort

The major demographic and clinical characteristics of the CREATE cohort are summarized in Table 1. A total of 20,373 patients (65.7%) received inpatient catheterizations and 9226 (29.8%) patients received catheterizations in outpatient settings. Most patients from each cohort (90.1% of inpatients, and 95.9% of outpatients) had only 1 catheterization episode. Men accounted for 70% of each cohort. Mean and median ages were 65 years for both cohorts. A total of 1389 (5.0% of total cohort) patients had catheterization in inpatient and outpatient settings on different admission dates.

Hypertension and hyperlipidemia were the most common comorbidities in both cohorts, experienced by more than 60% of patients. A total of 11.3% (n = 2306 patients) of the inpatients had previous PCI, and 4.9% (n = 1005) had previous CABG. Among outpatients, 12.3% (n = 1134 patients) and 5.4% (n = 500) had previous PCI and previous CABG, respectively.

Discussion

Building toward precision medicine

Integrating EMR data with other data modalities is crucial for achieving precision medicine initiatives and building a learning health system. Linking these 3 data sources led to the creation of a large catheterization cohort that has the potential to allow for big data analytics. Each of the 3 linked data modalities contain distinct clinically relevant information. We briefly recap some of the most salient elements from these data sets, and then describe 3 potential research applications illustrating the power of using them together for EMR data analytics.

Approach. Patients' cardiovascular conditions and their comorbidities are documented in APPROACH, but not necessarily in the EMR when they undergo catheterization by clinicians.[19,20] Furthermore, APPROACH might document comorbidity status more comprehensively than administrative data.[25] Consequently, APPROACH data could serve as a reference standard for EMR analytics. For example, the data could be used for developing and validating EMR data-based phenotyping algorithms for accurately defining cardiovascular

Table 1. Characteristics of inpatient and outpatient CREATE cohorts on index catheterization encounter in APPROACH

Characteristic	Inpatient cohort	Outpatient cohort
Total unique patients	20,373	9226
1 Catheterization	18,360 (90.1)	8850 (95.9)
2 Catheterizations	1674 (8.2)	356 (3.9)
3 or more catheterizations	339 (1.7)	20 (0.2)
Male sex	14,476 (71.1)	6357 (68.9)
Median age (IQR), years	64.1 (55.8 -73.2)	65.7 (58.0-73.1)
Diagnoses and comorbidities		
Cerebral vascular disease	1124 (5.5)	576 (6.2)
Congestive heart failure	3052 (15.0)	1143 (12.4)
Chronic obstructive pulmonary disease	3217 (15.8)	1710 (18.5)
Diabetes	5573 (27.4)	2826 (30.6)
Hypertension	13,803 (67.8)	6783 (73.5)
Hyperlipidemia	12,250 (60.1)	6427 (69.7)
Liver disease	409 (2.0)	278 (3.0)
Malignancy	1039 (5.1)	520 (5.6)
Myocardial infarction	2024 (9.9)	882 (9.6)
Peripheral vascular disease	1438 (7.1)	764 (8.3)
Procedures		
Previous PCI	2306 (11.3)	1134 (12.3)
Previous CABG	1,005 (4.9)	500 (5.4)

Data are n or n (%)

APPROACH, **A**lberta **Pro**vincial **P**roject for **O**utcome **A**ssessment in **C**oronary **H**eart Disease; CABG, coronary artery bypass grafting; IQR, interquartile range; PCI, percutaneous coronary intervention.

conditions. Using clinical registries such as APPROACH can save considerable human and financial resources compared with the traditional method of conducting manual chart reviews to establish a reference standard.

DAD/NACRS. International Classification of Diseases-coded administrative data are often used for defining cohorts and risk adjustment.[26,27] The routinely collected data are widely available, but undercoded or miscoded for numerous conditions.[28] DAD and NACRS are crucial for dividing the CREATE database into outpatient and inpatient cohorts, because EMR data associated with these admissions will vary and will affect future EMR data-based studies and their methodology protocols. These databases also provide insights on resource utilization and outcome data, and enable long-term follow-up for events that require hospitalization.

SCM. EMRs contain comprehensive accounts of patients' medical conditions, including free-text narrative documents (eg, clinical notes and discharge summaries). However, these types of data are challenging for downstream analytics applications. Natural language processing, often in conjunction with machine learning or deep learning, are required to analyze EMRs.[5-7] However, EMR data are messy to work with, and other data sources containing high-quality structured data elements are valuable for cohort identification and for informing study designs.

Our high data linkage of 99.96% ensured that the Calgary catheterizations cohort was successfully created using the APPROACH, EMR, and administrative databases. Our high linkage rate is because of Alberta's single-payer health insurance system and the existence of PHNs as a unique lifetime identifier. Our team was previously successful in conducting a linkage study between a provincewide surveillance system and an administrative database using similar methods.[29]

The CREATE database enables many different kinds of EMR-based analytical studies that support precision cardiac medicine. We discuss 3 potential research applications of the CREATE database.

Application 1: EMR phenotyping. The area of research that uses EMR data to develop comorbidity definitions is known as EMR phenotyping, and EMR-based phenotypes often show superior performance compared with traditional methods that use other data sources. For example, Nath et al. developed the EchoInfer application, which can extract clinical information from echocardiography reports housed in an EMR system.[30] Xu et al. developed an EMR data-based congestive heart failure phenotyping algorithm, by applying natural language processing on discharge summaries and structured EMR data elements.[31] EMR phenotyping development requires an initial reference standard (eg, comorbidity status from APPROACH). When EMR phenotypes with good validity are developed, they can be deployed on longitudinal EMR data to extract comorbidity progression over time. Such enhanced comorbidity identification can be integrated into existing clinical support processes and systems to enable individualized clinical decision-making. This can be achieved by enhancing existing risk scoring systems such as the LACE (Length of stay, Acuity of the admission, Comorbidity of the patient, and emergency department use in the duration of 6 months before admission) index[32] and the Framingham Risk Score[33] for improving patient outcomes[34] and supporting learning health systems.[5,6,35]

Application 2: developing individualized treatment approaches. A second potential application area is to develop individualized approaches to treatment using comprehensive longitudinal clinical data. Evidence from previous studies shows that EMR data can improve risk estimates. Kennedy et al. predicted the risk of cerebrovascular and cardiovascular death, and compared the Framingham Risk Score with regression methods using longitudinal clinical EMR data. The results indicated that longitudinal regression methods outperformed the Framingham score.[36] Zhao et al. extracted aggregated longitudinal features from EMR data and showed improved cardiovascular event prediction.[37] Panahiazar et al. used a hierarchical clustering algorithm to develop a heart failure therapy response medication recommendation plan on the basis of a patient similarity index developed using EMR data.[38] Additionally, CREATE can support direct comparison of these methods with traditional risk adjustment methods that rely on International Classification of Diseases-coded data.

Application 3: Building synergistic human-artificial intelligence hybrid systems for clinical decision support. A third potential application is designing artificial intelligence (AI)-based algorithms that will work in tandem with human clinicians, leading to improved data-informed decision-making.[39] AI technologies excel at detecting patterns in complex, high-dimensional data, but can only learn from the data that are recorded electronically. In contrast,

human clinicians collect a wide array of qualitative data that might inform optimal treatment decisions. There is the potential for developing interactive human-AI processes that synergistically combine clinical knowledge and intuition with predictive AI, to develop systems and tools for informing personalized treatment decisions, thereby supporting the goals of precision medicine and precision public health.[40] For example, it was recently shown that AI-assisted clinicians were better able to diagnose skin cancer than either AI or individual clinicians alone.[41] Building machine learning prediction models on cardiac outcomes using cardiac-specific clinical data, guided by clinical knowledge from clinicians (ie, cardiologists), might allow development of similar clinical decision support tools. Access to databases that are rich in clinical information, such as CREATE, will be essential for such endeavours.

Additional data linkage and infrastructure

The CREATE database represents a major milestone in building an enriched cardiovascular disease-specific database. The CREATE database can be linked to additional databases, such as genomics data, and has the potential to bridge the gap between biological sciences and applied health sciences.

Data extraction, storage, and management agreements with health authorities are in place for maintaining, updating, and expanding the databases to answer specific research questions. The CREATE database will be housed on the Medical Advanced Research Computing (MARC) cluster. MARC is a secure high-performance computing infrastructure, which is approved by AHS to store sensitive data (eg, SCM EMRs) at the University of Calgary. The University of Calgary has been designated as the Information Manager in collaboration with AHS for the CREATE database on this high-performance computing infrastructure. Having health system buy-in will be critical to turn CREATE into a real-time data set that supports a fully operational learning health system.

Limitations

This study has some limitations. First, highly restricted access was required by health system authorities under the terms of the information management agreement, because of the sensitive content and the difficulty of deidentifying unstructured data. As such, the CREATE database is not publicly available. We are presently exploring deidentification protocols[42,43] on the CREATE database to remedy this. Second, the CREATE database does not cover the entire population of cardiac patients in Calgary. This cohort is limited to patients who have undergone diagnostic coronary angiogram and revascularization procedures. The patients not covered and their outcomes have been previously studied.[44,45] Finally, the outpatient cohort does not contain equally comprehensive documentation, laboratory results, and problem lists as the inpatient cohort. Primary care EMR data should ideally be linked, because patients are often referred to primary care physicians after cardiac procedures.

Conclusion

CREATE was established to pursue cardiac precision medicine activities and facilitate big EMR data analytics. Future research applications will focus on: (1) EMR phenotyping; (2) developing individualized clinical and treatment approaches; and (3) building AI-based assistive decision-making tools.

Acknowledgements

The authors thank Kevin Lonergan and Gary Ruta from AHS Data & Analytics for their assistance with SCM EMR exploration.

For more information about data sources within CHI, contact chi@ucalgary.ca.

Funding Sources

This research was funded by Canadian Institutes of Health Research Foundation Grant FDN-167272, awarded to Dr Hude Quan.

Disclosures

The authors have no conflicts of interest to declare.

References

1. Ginsburg GS, Phillips KA. Precision medicine: from science to value. Health Aff (Millwood) 2018;37:694-701.
2. Deans KJ, Sabihi S, Forrest CB. Learning health systems. Semin Pediatr Surg 2018;27:375-8.
3. Cadarette SM, Wong L. An introduction to health care administrative data. Can J Hosp Pharm 2015;68:232-7.
4. Ambinder EP. Electronic health records. J Oncol Pract 2005;1:57-63.
5. Hemingway H, Asselbergs FW, Danesh J, et al. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Heart J 2018;39:1481-95.
6. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and cardiovascular care: a scientific statement from the American Heart Association. Circulation 2017;135:e826-57.
7. Banerjee A. Challenges for learning health systems in the NHS. Case study: electronic health records in cardiology. Future Healthc J 2017;4:193-7.
8. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26:1545-59.
9. McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011;4:13.
10. Birtwhistle RV. Canadian Primary Care Sentinel Surveillance Network: a developing resource for family medicine and public health. Can Fam Physician 2011;57:1219-20.
11. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood) 2014;33:1178-86.
12. Fleurence RL, Curtis LH, Califf RM, et al. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21:578-82.

13. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol 2015;44:827-36.

14. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035.

15. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: building a novel node in the learning health care system. Health Aff (Millwood) 2014;33:1187-94.

16. Williamson T, Green ME, Birtwhistle R, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Ann Fam Med 2014;12:367-72.

17. Clinical Research Unit, University of Calgary, Cumming School of Medicine. Alberta Provincial Project for Outcomes Assessment in Coronary Heart Disease (APPROACH). Available at: http://www.approach.org. Accessed February 28, 2020.

18. Lee S, Xu Y, D'Souza AG, et al. Unlocking the potential of electronic health records for health research. Int J Popul Data Sci 2020;5:1123.

19. Ghali WA, Knudtson ML. Overview of the Alberta Provincial Project for Outcome Assessment in Coronary Heart Disease. On behalf of the APPROACH investigators. Can J Cardiol 2000;16:1225-30.

20. Southern DA, Norris CM, Quan H, et al. An administrative data merging solution for dealing with missing data in a clinical registry: adaptation from ICD-9 to ICD-10. BMC Med Res Methodol 2008;8:1.

21. Quan H, Smith M, Bartlett-Esquilant G, et al. Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. Can J Cardiol 2012;28:152-4.

22. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130-9.

23. Gibson D, Richards H, Chapman A. The national ambulatory care reporting system: factors that affect the quality of its emergency data. Int J Inf Qual 2008;2:97-114.

24. Government of Alberta. Alberta Health Care Insurance Plan (AHCIP). Available at: https://www.alberta.ca/ahcip.aspx. Accessed April 23, 2020.

25. Jiang J, Southern D, Beck CA, et al. Validity of Canadian discharge abstract data for hypertension and diabetes from 2002 to 2013. CMAJ Open 2016;4:E646-53.

26. Skull SA, Andrews RM, Byrnes GB, et al. ICD-10 codes are a valid tool for identification of pneumonia in hospitalized patients aged > or = 65 years. Epidemiol Infect 2008;136:232-40.

27. Sundararajan V, Romano PS, Quan H, et al. Capturing diagnosis-timing in ICD-coded hospital data: recommendations from the WHO ICD-11 topic advisory group on quality and safety. Int J Qual Health Care 2015;27:328-33.

28. Quan H, Li B, Saunders LD, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health Serv Res 2008;43:1424-41.

29. Lee S, Ronksley P, Conly J, et al. Using data linkage methodologies to augment healthcare-associated infection surveillance data. Infect Control Hosp Epidemiol 2019;40:1144-50.

30. Nath C, Albaghdadi MS, Jonnalagadda SR. A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS One 2016;11:e0153749.

31. Xu Y, Lee S, Martin E, et al. Enhancing ICD code-based case definition for heart failure using electronic medical record data. J Card Fail 2020;26:610-7.

32. Robinson R, Hudali T. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. PeerJ 2017;5:e3137.

33. Bosomworth NJ. Practical use of the Framingham risk score in primary prevention: Canadian perspective. Can Fam Physician 2011;57:417-23.

34. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010;48:981-8.

35. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013;20:e206-11.

36. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 2013;51:251-8.

37. Zhao J, Feng Q, Wu P, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Sci Rep 2019;9:717.

38. Panahiazar M, Taslimitehrani V, Pereira NL, Pathak J. Using EHRs for heart failure therapy recommendation using multidimensional patient similarity analytics. Stud Health Technol Inform 2015;210:369-73.

39. Shameer K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiovascular medicine: are we there yet? Heart 2018;104:1156-64.

40. Lee J. Is artificial intelligence better than human clinicians in predicting patient outcomes? J Med Internet Res 2020;22:e19918.

41. Tschandl P, Rinner C, Apalla Z, et al. Human—computer collaboration for skin cancer recognition. Nat Med 2020;26:1229-34.

42. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24:596-606.

43. Neamatullah I, Douglass MM, Lehman LW, et al. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.

44. Southern DA, Ngo J, Martin BJ, et al. Characterizing types of readmission after acute coronary syndrome hospitalization: implications for quality reporting. J Am Heart Assoc 2014;3:e001046.

45. O'Neill DE, Southern DA, Norris CM, et al. Acute coronary syndrome patients admitted to a cardiology vs non-cardiology service: variations in treatment & outcome. BMC Health Serv Res 2017;17:354.

Supplementary Material

To access the supplementary material accompanying this article, visit *CJC Open* at https://www.cjcopen.ca/ and at https://doi.org/10.1016/j.cjco.2020.12.019.