IJC Heart & Vasculature 39 (2022) 100974

Contents lists available at ScienceDirect

IJC Heart & Vasculature

journal homepage: www.sciencedirect.com/journal/ijc-heart-and-vasculature

Applying computable phenotypes within a common data model to identify heart failure patients for an implantable cardiac device registry

Jove Graham [a,1,*], Andy Iverson [b,1], Joao Monteiro [b], Katherine Weiner [b], Kara Southall [b], Katherine Schiller [b], Mudit Gupta [c], Edgar P. Simard [b]

[a] *Center for Pharmacy Innovation and Outcomes, Geisinger Clinic, Danville, PA, USA*
[b] *Global Clinical Research and Analytics Medtronic, Inc., Minneapolis, MN, USA*
[c] *Phenomic Analytics and Clinical Data Core, Geisinger Clinic, Danville, PA, USA*

ARTICLE INFO

Keywords:
Comorbidities heart failure
Clinical trial
Registry
Electronic health record
Computable phenotype
Common data model

ABSTRACT

Background: Use of existing data in electronic health records (EHRs) could be used more extensively to better leverage real world data for clinical studies, but only if standard, reliable processes are developed. Numerous computable phenotypes have been validated against manual chart review, and common data models (CDMs) exist to aid implementation of such phenotypes across platforms and sites. Our objective was to measure consistency between data that had previously been manually collected for an implantable cardiac device registry and CDM-based phenotypes for the condition of heart failure (HF).

Methods: Patients enrolled in an implantable cardiac device registry at two hospitals from 2013 to 2018 contributed to this analysis wherein registry data were compared to PCORnet CDM-formatted EHR data. Seven different phenotype algorithms were used to search for the presence of HF and compare the results with the registry. Sensitivity, specificity, predictive value and congruence were calculated for each phenotype.

Results: In the registry, 176 of 319 (55%) patients had history of HF, compared with different phenotypes estimating between 96 (30%) and 188 (59%). The least-restrictive phenotypes (any diagnosis) had high sensitivity and specificity (90%/80%), but more restrictive phenotypes had higher specificity (e.g., code present in problem list, 94%). Differences were observed using time-based criteria (e.g., days between visit diagnoses) and between participating hospitals.

Conclusions: Consistency between manually-collected registry data and CDM-based phenotypes for history of HF was high overall, but use of different phenotypes impacted sensitivity and specificity, and results may differ depending on the medical condition of interest.

1. Introduction

The expanding volume of available electronic health record (EHR), claims, and other administrative data continues to drive interest in leveraging such real world data for evaluating the safety and effectiveness of medical products. In the United States, since the 2016 passage of the 21st Century Cures Act, the U.S. Food and Drug Administration (FDA) has issued guidance on how such data can support FDA-regulated clinical investigations [1,2]. Updated FDA draft guidances for drugs and biological products were issued in September and October 2021 that give much more detailed considerations regarding study design and data quality [3,4], and the FDA uses the term “eSourcing“ to describe direct capture of clinical data elements from existing electronic sources [5]. However, there are still knowledge gaps regarding how to ensure data reliability, which have contributed to slow adoption of this approach for data used in regulatory submission.

Baseline patient information in a clinical study is typically entered into a study database ‘manually’ by dedicated personnel who gather information from direct patient interviews or by reviewing paper or EHRs. Studies have reported reductions in data capture time and transcription errors by eSourcing discrete elements like age, sex, or race [6]. For more complex concepts like medical history, both manual approaches require considerable time and effort, and both have shortcomings: in direct interviews, patients may have recall bias, be

[*] Corresponding author at: 100 N Academy Ave, MC 44-01, Danville, PA 17822-2402, USA.
E-mail address: jhgraham1@geisinger.edu (J. Graham).

[1] Contributed equally to this work.

https://doi.org/10.1016/j.ijcha.2022.100974
Received 13 January 2022; Received in revised form 9 February 2022; Accepted 11 February 2022

unfamiliar with condition-specific terminology, misinterpret, or contradict their true medical history, and in manual chart reviews, there may be multiple places to look for information and unclear criteria for what constitutes a confirmed case [7–10]. Methods to eSource medical history information for clinical trials could potentially be highly valuable for that reason.

Partnerships such as the Electronic Medical Records and Genomics (eMERGE) [11] and Observational Health Data Sciences and Informatics (OHDSI) [12] have developed libraries of electronic "computable phenotypes" that are regularly used in observational and precision medicine research to identify patients with conditions of interest. These phenotypes translate human-readable definitions of a disease or condition into rules that look for the presence of matching elements in electronic source data [13,11]. The lack of standardization across EHRs and difficulties in linking EHRs across provider networks has meant that phenotypes have often been shared merely as descriptions of logic that must then be customized to the local data source. Multiple research networks, however, have developed common data models (CDMs) which are blueprints to transform disparate sources of healthcare data into a common structure of tables and fields [14]. Data from multiple source systems (e.g., EHR, billing, imaging) can be combined, and the primary benefit of CDMs is that they provide a non-proprietary, software-agnostic format for users to load data into a common structure of tables and fields for which standard queries and programs can be written. Users can develop code and analyses and distribute these across a network of sites without directly interfacing with the EHR or sharing identifiable data. Developers have therefore advocated for, and developed tools for, the implementation of phenotypes in CDMs [15–17].

Computable phenotypes are typically validated by comparing results with manual chart reviews that have usually been conducted expressly as a research exercise for the purpose of that validation, and then by reporting metrics such as sensitivity. We were interested to see if these validation metrics would be different when a representative group of computable phenotypes were compared with data that had been pragmatically collected by clinical personnel in the context of an implantable cardiac device registry. We realize that different use cases (e.g., identifying incident events vs. prevalent cases) might require different phenotypes and levels of sensitivity or specificity. The objective of this study was to measure the consistency between how the medical history of a single condition was defined via computable phenotypes implemented via a CDM versus how it had been identified during data collection for an implantable cardiac device registry. We chose heart failure (HF) as our example condition and performed this study in a population of patients at a large integrated health system with a long-established EHR and widely-adopted CDM (PCORnet) [18].

2. Methods

2.1. Data collection

All data for this study originated from two hospitals within Geisinger, an integrated health system in Pennsylvania, USA, serving over 500,000 patients per year with seven hospitals, 138 primary and specialty clinics, and a single EHR platform (Epic, Verona, WI). Patients in this study were seen at two large hospitals in central and northeastern Pennsylvania with similar underlying patient populations. Geisinger participates in several multicenter research networks, for which it implements multiple CDMs including PCORnet, Virtual Data Warehouse (VDW), and OHDSI. The study was reviewed and approved by Geisinger's Institutional Review Board (IRB).

The goal of the study was to compare the data in the registry to data that were sourced by applying computable phenotypes to the EHR data, via a CDM. The registry used in this analysis (Product Surveillance Registry, ClinicalTrials.gov Identifier NCT01524276) is currently active, has been reported on elsewhere, is patient-centric, and is intended "to provide continuing evaluation and periodic reporting of safety and effectiveness of Medtronic market-released products." [19–21]. To be eligible for registry enrollment, patients must be implanted with an eligible Medtronic device within a defined timeframe and give informed consent, and patients are only excluded if they are inaccessible for follow-up, excluded per local law, or enrolled in a concurrent study that could confound results. Following enrollment in the registry, clinical personnel at the hospital site document the presence or absence of many conditions in the medical history including HF; this determination is based on a review of various clinical documentation sources within the EHR including problem list, medical history, physical history, or notes from cardiac studies and office visits. The clinical site personnel's designation of HF as recorded in the registry database (as a yes/no flag) was considered the gold standard for the current study.

Patients who were age 18–89 when they enrolled in the registry prior to May 29, 2018 were considered eligible for this current study, with each patient's registry consent date considered the index date for purposes of computing HF phenotypes. Because this was a secondary review of data already collected, the IRB issued a waiver of informed consent for the present study. To further protect privacy, patients who were alive but outside the age range of 18–89 at the time of data extraction (September 2019) were excluded.

2.2. Computable phenotypes

Computable phenotypes were implemented via the PCORnet CDM format for EHR data. Geisinger had previously completed a PCORnet CDM v4.1 EHR-only implementation with records from August 1996 to March 2018. Implementation of the phenotypes via a CDM was done to gain the advantage of making our approach more interoperable for multi-site clinical research, with the tradeoff that the EHR's unstructured data could not be used. Application of phenotypes was based on the assumption that structured data in the EHR related to HF would be mapped to the PCORnet tables that contain encounter diagnoses (DIAGNOSIS), problem list diagnoses (CONDITION), and lab results (LAB_RESULT_CM). Other relevant HF metrics including ejection fraction, QRS duration, and New York Heart Association Classification [22] were not structured fields in the PCORnet implementation and therefore not used for this study.

As there are many HF phenotypes in the literature (and freely available via eMERGE and OHDSI), but no single phenotype universally accepted as best, we chose to test seven HF phenotypes, labeled HF1-HF7, representative of those in the prior literature [11,23,24]. The number of phenotypes (seven) was arbitrary and chosen in order to investigate a reasonable variety of different previously-reported principles or philosophies of phenotype building while limiting the study to a manageable scope. The primary differences among the seven phenotypes, which were not directly copied from previous work but adapted based on their design principles, were the location(s) of diagnosis codes in the EHR, the number and frequency of codes, and the presence of abnormal labs in addition to diagnoses. For each patient, phenotypes were only applied to data generated in the EHR on or before that patient's registry consent date. HF1 defined heart failure as the presence of any HF diagnosis code in either an encounter or problem list. HF2 required a HF diagnosis code to be associated with an encounter. HF3, HF4, and HF5 took a more restrictive approach by requiring that two encounters with HF diagnosis codes appeared at least 30, 60, and 90 days apart, respectively. HF6 required a diagnosis specifically in the patient's problem list since this location offers the most readily available diagnosis information to a clinician or clinical staff. Finally, HF7 required an inpatient or problem list diagnosis of HF and also an NT-proBNP-type Natriuretic Peptide (NT-proBNP) lab result [25] that had been flagged as abnormal. Definitions of the phenotypes and their relevant codes are summarized in Table 1.

All diagnosis codes for HF were from the International Classification of Diseases, Ninth/Tenth Revisions, Clinical Modification (ICD-9/10-CM) [26,27], and Logical Observation Identifiers Names and Codes

Table 1
Definitions of heart failure clinical phenotypes, referred to as HF1-HF7, and diagnosis/lab codes used for these phenotypes.

Phenotype	Description
HF1	Any encounter or problem list heart failure diagnosis code.
HF2	Any encounter heart failure diagnosis code.
HF3	Two encounters with a heart failure diagnosis code, >30 days apart.
HF4	Two encounters with a heart failure diagnosis code, >60 days apart.
HF5	Two encounters with a heart failure diagnosis code, >90 days apart.
HF6	Any heart failure diagnosis code on the problem list.
HF7	An abnormal NT-proBNP lab result flag AND a heart failure diagnosis code either on the problem list or an inpatient encounter.
Code System	**Codes**
ICD9-CM (Heart Failure)	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 428.0, 428.1, 428.20, 428.21, 428.22, 428.23, 428.30, 428.31, 428.32, 428.33, 428.40, 428.41, 428.42, 428.43, 428.9
ICD10-CM (Heart Failure)	I09.81, I11.0, I13.0, I13.2, I50.1, I50.20, I50.21, I50.22, I50.23, I50.30, I50.31, I50.32, I50.33, I50.40, I50.41, I50.42, I50.43, I50.9, I50.810, I50.811, I50.812, I50.813, I50.814, I50.82, I50.83, I50.84, I50.89
LOINC (NT-proBNP)	33762-6, 33763-4, 71425-3, 77621-1, 77622-9, 83107-3, 83108-1

(LOINC) from Regenstreif [28] were used to identify NT-proBNP lab results for HF7. Note that other authors have compared phenotypes that use different sets diagnosis codes (e.g., ICD-9 428.* versus others) [29], but for internal comparison purposes, we chose to apply the full list of all ICD9/10 HF codes from the Chronic Conditions Data Warehouse (CCW) Algorithm to all seven phenotypes [13].

2.3. Statistical analysis

Our goal was to compare a binary variable (presence or absence of HF history) between the registry and each of the seven phenotypes applied via the PCORnet CDM. For each of the comparisons, we calculated five metrics—congruence, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)–considering the registry to be the gold standard source of truth. The five metrics are defined in more detail in Table 2. Each metric was expressed as a percentage, along with empiric 95% confidence intervals based on bootstrapping with 50,000 repetitions. Because we had no *a priori* hypotheses about how these phenotypes compared, no hypothesis testing or causal inference was conducted other than descriptive examinations of the confidence intervals.

We also conducted a secondary analysis stratifying all metrics by site/hospital, labeled as Site A or Site B, with each individual patient belonging to one site or the other. Our rationale for this secondary analysis was that registry designations required review and interpretation of medical records and manual entry; therefore, there could be differences in metrics depending on differences in EHR search strategies, knowledge, clinician documentation, or site-specific workflows between the sites. We examined these stratified results to explore how phenotype performance could have been impacted by these factors. All statistical analysis was performed using SAS (SAS Institute, Cary, NC) or R (The R Group, Vienna, Austria) statistical software.

3. Results

There were 319 patients enrolled in the registry from February 2013 to May 2018 and eligible for this analysis. Median age was 73 years old (range 29–89), 65% of subjects were male (206/319), and the cohort was 99% Caucasian (315/319) and non-Hispanic (318/319), reflecting the demographics of the region. The number of patients with a history of HF at baseline reported in the registry was 176 (55%) of 319. In comparison, the seven phenotypes categorized between 96 (30%) and 188 (59%) patients as having a history of HF at baseline.

Table 2
Definitions of the five performance metrics used to compare registry vs. PCORnet CDM heart failure history, for each of the seven phenotypes.

Performance Metric	Interpretation	Formula
Congruence	Percent of Registry patients whose Registry and CDM HF status agree	$\frac{(\text{N with same HF status in Registry and CDM})}{(\text{Total N in Registry})}$
Sensitivity	Percent of patients with HF in Registry who also have HF in CDM	$\frac{(\text{N with same HF status in Registry and CDM})}{(\text{N with HF in Registry})}$
Specificity	Percent of patients without HF in Registry who also are without HF in CDM	$\frac{(\text{N without HF in either Registry or CDM})}{(\text{N without HF in Registry})}$
PPV	Percent of patients with HF in CDM who also have HF in Registry	$\frac{(\text{N with HF in both Registry and CDM})}{(\text{N with HF in CDM})}$
NPV	Percent of patients without HF in CDM who are also without HF in Registry	$\frac{(\text{N without HF in either Registry or CDM})}{(\text{N without HF in CDM})}$

Table 3 displays the five performance metrics for each phenotype. The two least restrictive phenotypes (HF1 and HF2) gave identical results and demonstrated the highest sensitivity (90.3%) and NPV (87.0%), as well as relatively high congruence (85.6%), specificity (79.7%) and PPV (84.6%). For the phenotype variants (HF3 through HF5) that required diagnoses codes at multiple encounters separated by increasing amounts of time (30, 60 and 90 days), a slight improvement in specificity (from 92.3% to 93.0%) as the separation time increased was offset by much larger decreases in congruence (82.5% to 79.0%), sensitivity (74.5% to 67.6%) and NPV (74.6% to 70.0%). PPV was the least affected by varying the time criterion, with identical values at 30 and 90 days and only a slight decrease at 60 days (92.2%, 92.2% and 91.9%, respectively). The phenotype (HF6) that specifically required a diagnosis on the problem list had the highest specificity (94.4%) and PPV (92.9%), but much poorer sensitivity and NPV (59.7% and 65.5%, respectively). The phenotype (HF7) that required an abnormal NT-proBNP laboratory indicator showed high specificity (93.0%) and PPV (89.6%) but had the lowest estimates for all other performance metrics.

Some phenotypes performed significantly better at one hospital site than the other. Table 4 shows the absolute differences in each performance metric between the two sites, with individual sites' metrics in parentheses and shaded vs. unshaded cells indicating which metrics were higher at Site A or B, respectively. Sensitivity and PPV were higher at Site A for all phenotypes, while specificity and NPV were higher at Site B. Differences between the sites also decreased as the phenotypes became more specific: for example, the sensitivity differed by 13.1% between sites for HF1 (any diagnosis) vs. only 0.9% for HF6 (diagnosis specifically on the problem list).

Table 3
Five performance metrics with 95% confidence intervals for the seven heart failure phenotypes.

Phenotype	Description	Congruence (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
HF1	Any diagnosis	85.6% (81.5%, 89.3%)	90.3% (85.8%, 94.5%)	79.7% (73.0%, 86.0%)	84.6% (79.3%, 89.5%)	87.0% (81.1%, 92.6%)
HF2	Any encounter diagnosis	85.6% (81.5%, 89.3%)	90.3% (85.7%, 94.5%)	79.7% (72.9%, 86.2%)	84.6% (79.3%, 89.6%)	87.0% (81.0%, 92.5%)
HF3	Multiple encounter diagnoses > 30 days apart	82.5% (78.2%, 86.5%)	74.5% (67.9%, 80.8%)	92.3% (87.6%, 96.4%)	92.2% (87.6%, 96.3%)	74.6% (68.0%, 80.9%)
HF4	Multiple encounter diagnoses > 60 days apart	80.6% (76.2%, 85.0%)	71.0% (64.2%, 77.6%)	92.3% (87.7%, 96.4%)	91.9% (87.1%, 96.2%)	72.1% (65.5%, 78.6%)
HF5	Multiple encounter diagnoses > 90 days apart	79.0% (74.3%, 83.4%)	67.6% (60.5%, 74.5%)	93.0% (88.5%, 96.9%)	92.2% (87.3%, 96.6%)	70.0% (63.3%, 76.4%)
HF6	Problem list	75.2% (70.5%, 79.9%)	59.7% (52.3%, 66.9%)	94.4% (90.3%, 97.9%)	92.9% (87.8%, 97.3%)	65.5% (59.0%, 71.9%)
HF7	(Problem list OR inpatient diagnosis) AND abnormal NT-proBNP lab	68.6% (63.6%, 73.7%)	48.8% (41.4%, 56.2%)	93.0% (88.5%, 96.9%)	89.6% (83.1%, 95.3%)	59.6% (53.1%, 65.9%)

Table 4
Absolute difference in % for each metric and phenotype between clinical Sites A and B.

	Congruence	Sensitivity	Specificity	PPV	NPV
HF1	3.6 (87.5 vs 83.9)	13.1* (95.9 vs 82.8)	16.6* (68.2 vs 84.8)	6.5 (87.2 vs 80.7)	1.7 (88.2 vs 86.5)
HF2	3.5 (87.5 vs 84.0)	13.1* (96.0 vs 82.9)	16.7* (68.1 vs 84.8)	6.5 (87.2 vs 80.7)	1.6 (88.2 vs. 86.6)
HF3	2.2 (81.2 vs 83.4)	8.3 (78.0 vs 69.7)	5.3 (88.6 vs 93.9)	4.1 (93.9 vs 89.8)	16.2* (63.9 vs 80.1)
HF4	2.6 (79.1 vs 81.7)	9.1 (74.9 vs 65.8)	5.3 (88.6 vs 93.9)	4.5 (93.7 vs 89.2)	17.2* (60.9 vs 78.1)
HF5	3.5 (77.0 vs 80.5)	7.8 (70.9 vs 63.1)	3.0 (90.9 vs 93.9)	5.8 (94.6 vs 88.8)	18.9* (57.9 vs 76.8)
HF6	10.5* (69.4 vs 79.9)	0.9 (60.0 vs 59.1)	5.1 (90.8 vs 95.9)	1.9 (93.7 vs 91.8)	25.4* (49.9 vs 75.3)
HF7	11.2* (62.4 vs 73.6)	0.3 (49.9 vs 48.6)	0.2 (93.1 vs 92.9)	10.2 (94.2 vs 84.0)	25.7* (44.5 vs 70.2)

Numbers in parentheses indicate the stratified performance metrics at Site A and Site B, respectively. Shaded cells represent instances where the performance metric was higher at Site A, unshaded cells represent instances where the metric was higher at Site B, and asterisks indicate where the 95% confidence interval of the difference between sites did not include zero.

4. Discussion

In this study of 319 patients enrolled in an implantable cardiac device registry at a large, integrated health system, we found that applying computable phenotypes to EHR data via a CDM implementation showed strong agreement with clinical personnel's prior assessments of whether patients in that registry had a history of heart failure. We compared performance among seven computable phenotypes derived from prior literature, and the least restrictive of these (using any HF diagnosis) demonstrated 85.6% congruence, 90.3% sensitivity, and 79.7% specificity. Requiring a diagnosis to be on the problem list yielded the highest specificity and PPV but much poorer sensitivity and NPV. The only phenotype to require evidence of an abnormal lab result (NT-proBNP) showed good specificity and PPV but had the lowest performance otherwise. Phenotypes requiring diagnoses at multiple time points improved the specificity but at the expense of the other measures, and increasing the length of the time period between diagnoses did not have much effect on specificity or PPV. Differences in metrics were seen between the two participating hospital sites, reinforcing the idea that human decisions and workflows contribute to present-day gold standard practices.

The relative performance of the seven phenotypes indicates that comprehensive queries of a patient's structured EHR data for any diagnosis of heart failure (HF1 and HF2) were more comparable with site personnel's registry data collection than phenotype algorithms that utilized more complex logic or focused on a specific EHR location. In their 2014 systematic review of HF phenotype studies, McCormick et al. concluded that, across 19 studies, HF diagnosis codes in administrative data were highly predictive of true HF cases, though they failed to capture as many as 30% of true cases, particularly less severe cases [23]. The authors' four recommendations for future studies generally advocated for loosening restrictions, e.g., using both primary and secondary diagnoses, using both inpatient and outpatient diagnoses, and searching problem lists and unstructured text for mentions of heart failure. Our findings generally support this theme (except that we could not include unstructured text), as the least restrictive phenotypes gave us not only the highest sensitivity but highest total congruence between registry and CDM as well.

We were not expecting phenotypes HF1 and HF2 to yield identical results as they did. Subsequent investigation revealed that during Geisinger's implementation of the PCORnet CDM model, programmers chose to exclude problem list diagnosis entries if they could not be linked to a specific EHR encounter, a decision that made the first two phenotypes (HF1 and HF2) identical since all diagnoses were tied to encounters. Decisions like this one are consistent with, and necessary for, transforming raw EHR data into a CDM format to gain the benefits of a standardized structure, but highlight the importance of nuances. In 2019, Hripcsak et al. reported on the implementation of phenotypes in a common data model across ten participating sites [30]. Their assessment was that using CDMs, particularly at sites that have already populated that model, can drastically speed phenotype implementation, and we believe that our study adds to the literature by demonstrating use of a different widely-used CDM (PCORnet) for heart failure. We recognize that using the CDM format was not strictly needed for this study inside a single health system but was done to enable the extension of this work across additional health systems or provider networks. We also note the caveat here that use of a CDM alone does not guarantee consistency of results among participating sites unless those sites are also following similar data collection procedures and implement data quality reviews or other processes to improve comparability.

The decision to exclude some problem list entries from the PCORnet CDM may also have contributed to the low sensitivity of HF6, which required a diagnosis to be on the problem list. Problem lists record active health conditions and are easily accessed by several members of a patient's care team which makes them a logical first place for an algorithm or clinical site personnel to search for a condition. Previous studies have cautioned, however, that the problem list is often incomplete or may contain outdated patient information [20,31]. In our case, requiring a problem list diagnosis yielded the highest specificity and PPV, but much poorer sensitivity and NPV than the other phenotypes. This finding may support the argument that computable phenotypes have the advantage over manual data collection that the entire health record can be queried just as easily as a single location like a problem list.

The phenotype (HF7) that required an abnormal NT-proBNP laboratory value indicator showed the poorest congruence, sensitivity and NPV. We note that this phenotype relied on the CDM data containing an abnormal indicator (i.e., flag), and in a secondary analysis we noted that the abnormal indicator was not always applied in the original EHR source data consistently, such as readings in the 450–899 pg/mL range for patients age 50–74 that were flagged as normal [30]. Other authors have included NT-proBNP > 500 pg/mL in their phenotype definitions [32,33], but our results highlighted that lab values presented several challenges, including patients not having the lab taken, and possible conflicting interpretations of abnormal flags.

We are not aware of a previous study investigating the application of HF criteria for multiple encounters over increasingly long periods of time, as in HF3-HF5. Increasing the period of time between diagnosis codes significantly reduced sensitivity and moderately increased specificity. Multiple encounter diagnoses are also common search criteria to identify patients with an actively managed health condition, because they offer the face validity of ruling out spurious or erroneous diagnosis codes that appear only once in the record and never recur. Our results, however, suggest that such criteria did not provide obvious benefits for the purpose of a registry.

For all phenotypes, sensitivity was higher for Site A and specificity was higher for Site B. The greatest discrepancies in metrics were for the least restrictive phenotypes, HF1 and HF2, where sensitivity and specificity varied between sites by 13.1 and 16.6 percentage points, respectively. These differences in phenotype performance between the two hospital sites could be attributed to differences in a number of human or process-related factors, including but not limited to: (1) differences in codes used; (2) differences in physician coding practice; (3) differences in EHR workflows; (4) differences in abstractor knowledge or experience; or (5) difference in training/guidance given to abstractors. It is important to emphasize that site personnel were not reviewing the same patients. The fact that the differences in sensitivity and specificity between the two sites narrowed as the phenotypes became more restrictive, however, suggest that if the sites had reviewed the same patient, they would have been in better agreement with each other for a patient who met the more restrictive phenotypes' criteria for HF. For patients whose evidence of heart failure was less obvious, however, Site A's personnel were more likely than Site B's to label patients as having heart failure.

To examine the non-concordant patients more closely, we subsequently performed additional manual chart review of a random sample (approximately 25%) of patients who were 'false negatives' (i.e., heart failure in registry but not CDM) or 'false positives' (i.e., heart failure in CDM but not in registry) according to phenotypes HF1 and HF2. Approximately 20 min per patient was dedicated solely to searching for evidence of heart failure in the EHR, including areas not specified in the phenotypes. For all false negative patients reviewed, history of heart failure was only found to be referenced in progress notes, discharge instructions, or other 'free text' areas of the EHR, which explains why they were detected by the clinical site but not by the phenotypes. For the review of false positive results, there was a more diverse mixture of patients with heart failure on the problem list or in documentation of past encounters. It is therefore unknown why the clinical sites did not classify these patients as having heart failure. It could be either because information was missed in their search or because they had access to additional information outside of the EHR that ruled out a history of heart failure. Given the variation observed between two hospitals in one system, we recognize that use of phenotypes and CDMs across multiple health systems may not be able to address or overcome differences in how data are collected. These results suggest that using phenotypes earlier in the process, as part of a data collection workflow, might lead to more consistency in data collection between sites, particularly for instances where less restrictive phenotypes are needed.

The primary strength of this study was our ability to leverage an existing implantable cardiac device registry in an EHR-integrated health system where patient information could be compared side-by-side with data from an already-implemented CDM. Main limitations were the sample size and the fact that the study was performed at a single health system on a single chronic disease, which could limit generalizability. We also acknowledge that our use case, the documenting of a prevalent condition in the medical history, may differ from other use cases such as identification of patients or incident events. For the purposes of studying HF specifically, we recognize that we could have added phenotypes that considered medication prescribing but chose not to, in order to keep the scope more manageable, and we were also limited by the fact that ejection fraction (a common surrogate measure for disease) was not available in the CDM data and therefore could not be leveraged in phenotype definitions. We recognize that diseases of varying complexity may be documented differently and could impact the congruence between manual data collection and electronically sourced data. Finally, although we defined the registry data as gold standard for the purpose of computing metrics, we did not have an additional third-party adjudication of each patient's heart failure status as a true standard. Our results, particularly those that showed differences in CDM performance when compared with the registry information recorded at two different sites, highlight the fact that non-automated data capture may be influenced by subjective decision-making.

5. Conclusion

In conclusion, these results provide further evidence that combining computable phenotypes with CDM-structured data for the purpose of implantable cardiac device registries is promising and could reduce repetitive work and unwanted variance while still yielding high agreement with data generated by manual chart review. Several limitations and questions remain as areas for further study, however. More information is needed on how results may vary across conditions, as some diseases have more algorithm-resistant clinical nuances than others. Given the differences in how medical history is collected across sites during implantable device registry studies, a better understanding is also needed of how clinically impactful those differences are to study results. Finally, more information is needed on whether traditional manual entry of clinical data into device registries should indeed always be considered a gold standard of accuracy or if there are instances where well-defined clinical phenotypes are more appropriate. It is important for users of device registry (or other clinical study) data to understand these complex issues and their impact on results when performing observational research or deploying direct data capture methods.

Financial disclosures

Several authors (AI, JM, KW, KS, KS, ES) are employed by and own shares in Medtronic plc.

Funding

Funding was provided by Medtronic Inc.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Several authors (AI, JM, KW, KS, KS, ES) are employed by and own shares in Medtronic plc.].

Acknowledgements

The authors wish to acknowledge Lorraine Tusing and Christina Gregor for chart review and project management at Geisinger. The authors also wish to thank Nichole Carslon, Nishi Bhadra, Susan Lang, and Julia Yang for facilitating the data transfer, and Linda Nelson for guidance and manuscript revision and review.

References

[1] Food and Drug Administration, “Guidance for Industry: Electronic Source Data in Clinical Investigations,” September 2013. [Online]. Available: http://www.fda.gov/media/85183/download. [Accessed 02 09 2022].

[2] Food and Drug Administration, “Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff,” 12 05 2017. [Online]. Available: https://www.fda.gov/media/99447/download. [Accessed 12 05 2021].

[3] Food and Drug Administration, “Data Standards for Drug and Biological Product Submissions Containing Real-World Data,” October 2021. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data. [Accessed 22 11 2021].

[4] Food and Drug Administration, “Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. Guidance for industry: draft guidance,” September 2021. [Online]. Available: http://fda.gov/media/152503/download. [Accessed 9 2 2022].

[5] C. Shore, A. Wagner Gee, B. Kahn and E. Hammers Forstag, “Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series,” 06 02 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539609/. [Accessed 12 05 2021].

[6] A.H. Nordo, E.L. Eisenstein, J. Hawley, S. Vadakkeveedu, M. Pressley, J. Pennock, I. Sanderson, A comparative effectiveness study of eSource used for data capture for a clinical research registry, Int. J. Med. Inform. 103 (2017) 89–94.

[7] A.M. Kelstrup, P. Juillerat, J. Korzenik, The accurary of self-reported medical history: a preliminary analysis of the promise of internet-based research in Inflammatory Bowel Diseases, J. Chrons. Colitis 8 (5) (May 2014) 349–356.

[8] S. Chung, T. Rosewall, R. Menezes, T. Kalliomäki, “I'm just guessing these answers!” An evaluation of the (in)accuracy of patient-reported medical history collected as part of a breast imaging program, J. Med. Imaging Rad. Sci. 49 (4) (2018) 390–396.

[9] A. Iliceto, S.L. Berndt, J.H. Greenslade, W.A. Parsonage, C. Hammett, M. Than, T. Hawkins, K. Parker, S. O'Kane, L. Cullen, Agreement between patient-reported and cardiology-adjudicated medical history in patients with possible ischemic chest pain: an observational study, Crit. Pathw. Cardiol. 15 (3) (2016) 121–125.

[10] J.K. Schmier, M.T. Halpern, Patient recall and recall bias of health state and health status, Expert. Rev. Pharmacoecon. Outcomes. Res. 4 (2) (2004) 159–163.

[11] Vanderbilt University, “PheKB a knowledgebase for discovering phenotypes from electronic medical records,” 2017. [Online]. Available: https://phekb.org/phenotypes. [Accessed Jun 2020].

[12] Observational Health Data Sciences and Informatics, “Phenotype Library,” [Online]. Available: https://data.ohdsi.org/PhenotypeLibrary/. [Accessed 12 May 2021].

[13] Centers for Medicare & Medicaid Services, “Condition Categories,” Feb 2019. [Online]. Available: https://www2.ccwdata.org/web/guest/condition-categories. [Accessed June 2020].

[14] J. Weeks, R. Pardee, Learning to share health care data: a brief timeline of influential common data models and distributed health data networks in U.S. health care research, EGEMS 7 (1) (2019) 1–7.

[15] G. Hripcsak, N. Shang, P.L. Peissig, L.V. Rasmussen, Facilitating phenotype transfer using a common data model, J. Biomed. Informat. 96 (103253) (2019).

[16] J. Pacheco, L. Rasmussen, R. Kiefer, T. Campion, A case study evaluating the portability of an executable computable phenotype algorithm across mulitple institutions and electronic health recrd environments, J. Am. Med. Inform. Assoc. 25 (11) (2018) 1540–1546.

[17] Observational Health Data Sciences and Informatics, “ATLAS-a unified interface for the OHDSI tools,” [Online]. Available: https://www.ohdsi.org/atlas-a-unified-interface-for-the-ohdsi-tools/. [Accessed 12 05 2021].

[18] R.L. Fleurence, L.H. Curtis, R.M. Califf, R. Platt, J.V. Selby, J.S. Brown, Launching PCORnet, a national patient-centered clinical research network, JAMIA 21 (4) (2014) 578–582.

[19] NIH U.S. National Library of Medicine, “ClinicalTrials.gov,” 21 April 2021. [Online]. Available: https://clinicaltrials.gov/ct2/show/record/NCT01524276. [Accessed 11 May 2021].

[20] W. Ceusters, J. Blaisure, Caveats for the use of the active problem list as ground truth for decision support, Stud. Health Technol. Inform. 255 (2018) 10–14.

[21] J.P. Singh, Y.-M. Cha, M. Lunati, E.S. Chung, S. Li, P. Smeets, D. O'Donnell, Real-world behavior of CRT pacing using the AdaptivCRT algorithm on patient outcomes: Effect on mortality and atrial fibrillation incidence, J. Cardiovasc. Electrophysiol. 31 (4) (2020) 825–833.

[22] M. Dolgin, A.C. Fox, R. Gorlin, R.I. Levin, Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels, ew York Heart Association. Criteria Committee. Nomenclature and criteria for diagnosis of diseases of the heart and great vessels, (1994) 9(567).

[23] N. McCormick, D. Lacaille, V. Bhole, J.A. Avina-Zubieta, Y. Guo, Validity of Heart Failure Diagnoses in Administrative Databases: A Systematic Review and Meta-Analysis, PLoS ONE 9 (8) (2014).

[24] S.S. Cohen, V.L. Roger, S.A. Weston, R. Jiang, N. Movva, A.A. Yusuf, A. M. Chamberlain, Evaluation of claims-based computable phenotypes to identify heart failure patients with preserved ejection fraction, Pharmacol. Res. Perspect. 8 (6) (2020), https://doi.org/10.1002/prp2.v8.610.1002/prp2.676.

[25] A. Maisel, C. Mueller, K. Adams, S.D. Anker, N. Aspromonte, J.G.F. Cleland, A. Cohen-Solal, U. Dahlstrom, A. DeMaria, S. Di Somma, G.S. Filippatos, G. C. Fonarow, P. Jourdain, M. Komajda, P.P. Liu, T. McDonagh, K. McDonald, A. Mebazaa, M.S. Nieminen, W.F. Peacock, M. Tubaro, R. Valle, M. Vanderhyden, C.W. Yancy, F. Zannad, E. Braunwald, State of the art: Using natriuretic peptide levels in clinical practice, Eur. J. Heart Fail. 10 (9) (2008) 824–839.

[26] Centers for Medicare & Medicaid Services, “ICD-9-CM Diagnosis and Procedure Codes: Abbreviated and Full Code Titles,” [Online]. Available: https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/codes. [Accessed Jun 2020].

[27] Centers for Medicare & Medicaid Services, “ICD-10,” [Online]. Available: https://www.cms.gov/Medicare/Coding/ICD10/index. [Accessed Jun 2020].

[28] Regenstrief Institute Inc., “LOINC from Regenstrief,” [Online]. Available: https://loinc.org/. [Accessed Aug 2020].

[29] V. Roger, S. Weston, M. Redfield, J. Hellermann-Homan, J. Killian, B. Yawn, S. Jacobsen, Trends in heart failure incidence and survival in a community-based population, JAMA 292 (3) (2004) 344–350.

[30] G. Hripcsak, N. Shang, P.L. Peissig, L.V. Rasmussen, C. Liu, B. Benoit, R.J. Carroll, D.S. Carrell, J.C. Denny, O. Dikilitas, V.S. Gainer, K.M. Howell, J.G. Klann, I. J. Kullo, T. Lingren, F.D. Mentch, S.N. Murphy, K. Natarajan, J.A. Pacheco, W.-Q. Wei, K. Wiley, C. Weng, Facilitating phenotype transfer using a common data model, J. Biomed. Inform. 96 (2019) 103253, https://doi.org/10.1016/j.jbi.2019.103253.

[31] A. Wright, A.B. McCoy, T.-T. Hickman, D.S. Hilaire, D. Borbolla, W.A. Bowes, W. G. Dixon, D.A. Dorr, M. Krall, S. Malholtra, D.W. Bates, D.F. Sittig, Problem list completeness in electronic health records: a multi-site study and assessment of success factors, Int. J. Med. Inform. 84 (10) (2015) 784–790.

[32] F. Alqaisi, L.K. Williams, E.L. Peterson, D.E. Lanfear, Comparing methods for identifying patients with heart failure using electronic data sources, BMC Health Serv. Res. 9 (1) (2009), https://doi.org/10.1186/1472-6963-9-237.

[33] M. Rosenman, J. He, J. Martin, K. Nutakki, G. Eckert, K. Lane, I. Gradus-Pizlo, S. L. Hui, Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory, J. Am. Med. Inform. Assoc. 21 (2) (2014) 345–352.