PCORnet-CMS Pilot Linkage Project

Process & Results

Authors
Bradley G. Hammill[1,2], Yinghong Zhang[1], Laura G. Qualls[1], L. Russell Waitman[3], Matthew J. Resnick[4], Russell L. Rothman[4], Lesley H. Curtis[1,2]

Affiliations
[1] Duke Clinical Research Institute
[2] Duke School of Medicine
[3] Kansas University Medical Center
[4] Vanderbilt University Medical Center

Duke Clinical Research Institute

Table of Contents

Tables and Figures

Introduction

The goal of this pilot project was to develop, test, and evaluate processes for using Medicare claims data to supplement PCORnet Clinical Data Research Network (CDRN) data. This work was undertaken to support the ADAPTABLE study, which was not yet fully underway when this work began, since linked Medicare data was proposed as one of the supplemental data sources to be used for complete ascertainment of trial events.

This project was conducted in multiple phases. The first phase involved transforming existing Medicare claims data, housed by Duke Clinical Research Institute (DCRI), into the PCORnet Common Data Model (CDM) v3.0. These data were then used to provide an initial feasibility assessment of this larger objective by using claims data from a random 5% sample of Medicare beneficiaries to both (a) identify patients within Medicare fee-for-service claims data who would be eligible for the ADAPTABLE trial, and (b) to ascertain outcomes of interest within this population using claims data.

The second phase of the project involved partnering with sites from 2 PCORnet CDRNs and consulting with key representatives from the Centers for Medicare and Medicaid Services (CMS) to establish processes and data flows that would enable multi-institution research using Medicare data for patients identified within PCORnet. The partner sites—Vanderbilt (Mid-South) and University of Kansas Medical Center (GPC)—joined the DCRI in submitting a multi-institution Data Use Agreement (DUA) with CMS with the purpose of obtaining Medicare claims data for a cohort of patients that were, again, similar to those being enrolled in ADAPTABLE trial. The Medicare data obtained from this phase of the project were also transformed into the PCORnet CDM and queried.

Phase 1 / Transforming Medicare Data to the PCORnet Common Data Model

The most straightforward way to incorporate Medicare data into an analysis of PCORnet data is to put that Medicare data into a PCORnet CDM-compliant DataMart so that it can be queried using the same SAS or SQL code that is used to query a PCORnet EHR-based DataMart. By doing this, any results generated from the Medicare data will have the same structure as results generated from PCORnet EHR data, which enables an analyst to easily merge or concatenate these files. Since Medicare data are not structured as a PCORnet CDM-compliant DataMart by default, the initial work of this project phase was to develop a transparent and repeatable process that would enable this transformation.

The data we transformed were Medicare data from 2010 to 2013 for a random 5% sample of all Medicare beneficiaries. Specifically, we obtained a DUA from CMS to use the following Research Identifiable Files for this project: Master Beneficiary Summary File (MBSF), inpatient claims file, outpatient claims file, carrier (physician/professional) claims file, and Part D (prescription medication) event file. The MBSF contains information about beneficiary demographics, mortality, and enrollment periods. The inpatient and outpatient claims files contain institutional claims for services provided in a hospital inpatient or hospital outpatient setting. The carrier claims contain physician and other professional service claims (e.g., laboratory testing, ambulance services) for services provided in all healthcare settings. The Part D event file contains information about outpatient dispensed prescription medications. Other types of claims files (e.g., home health service, durable medical equipment) are available and are planned to be incorporated into future work.

Data transformation specifications and programming

The first step in this process was the development of specifications that described all the mappings needed to get from the existing Medicare data tables, fields, and value sets to the PCORnet CDM table, fields, and value sets. At the time this project was undertaken, PCORnet CDM v3.0[1] was the version implemented throughout the network. **Table 1** shows which Medicare files informed which PCORnet CDM tables. Note that not all PCORnet CDM tables could be populated using Medicare data. For example, Medicare data contain no information on lab results, prescribed medications, patient-reported outcomes, etc.

Mapping of many Medicare fields and values to the PCORnet CDM was straightforward. For example, most demographic information involved simple mapping; death information from the MBSF went easily in the Death table; and diagnosis codes and procedure codes found on claims went directly into the appropriate tables. Missing data were handled according to the HL7 standards articulated in the CDM. The full mapping specification, involving more than 500 field and value set relationships, is posted on GitHub[2] (current version, with changes described later) and on iMeet Central[3] (original and current versions). The iMeet Central location also includes the relevant PCORnet Annotated Data Dictionary.

There was some information within the Medicare data, however, that did not map simply into the PCORnet CDM. Three examples are noteworthy. First, periods of prescription medication enrollment were not able to be included in the Enrollment table. As initially conceived, the Enrollment table was designed to capture periods of time during which a person was expected to have complete data capture, without the notion that different healthcare delivery domains may have different enrollment periods. In Medicare, enrollment for coverage of healthcare utilization services is separate from enrollment for coverage of prescription medication services. We chose to populate the PCORnet CDM Enrollment table with the utilization enrollment periods, since this is arguably the more important enrollment window to record.

Second, there were many professional service claims for physician services provided within an institutional setting. It was not clear how these should be included in the PCORnet CDM Encounter table. This was an issue because the PCORnet CDM is encounter-based, while the Medicare data is not. Medicare data are based on the billing entity, and the same encounter can generate bills from many different healthcare providers. As an example, an outpatient surgical procedure may generate a bill from the hospital, a bill from the anesthesiologist, and a bill from the surgeon. Unfortunately, there is no easy way to reliably combine those claims to completely define an encounter. Therefore, no attempt was made to reconcile different claims into the same encounter. For inpatient or outpatient hospital encounters, the institutional bill was the one that we used to define the encounter in the PCORnet CDM Encounter table, since these were the claims that had information about admission date, discharge date, discharge disposition, etc. Physician bills for services rendered in an institutional setting were coded as "other" encounter types.

Both of these issues led to changes in PCORnet CDM v3.1[1]. New items were added to value sets for existing PCORnet CDM fields that allowed more meaningful mapping of claims data to the PCORnet CDM. The additions are described in **Table 2**.

A mapping issue that was not easily resolved involved the Medicare concept of race. Within Medicare data, race and Hispanic ethnicity are both captured in a single field that does not allow multiple selection, whereas the PCORnet CDM allows designation of both race and Hispanic ethnicity as separate fields. This means that, within Medicare data, a person of Hispanic ethnicity has to either be classified as Hispanic, ignoring their race, or classified as a specific race, ignoring their ethnicity. **Table 3** shows the imperfect decisions that were made in how to translate this information into the separate fields available within the PCORnet CDM. This type of mapping cannot be fixed by changes to the PCORnet CDM, but would rather require changes to data collection within Medicare.

After these specifications were approved, SAS code was written to perform the extraction, transformation and loading (ETL) of Medicare data to the PCORnet CDM. When the first version of the Medicare-based DataMart was ready, the PCORnet data curation query was run against it. Query results were used to evaluate the DataMart's foundational data quality by applying a series of data quality checks addressing data model conformance, data plausibility, and data completeness (see PCORnet Data Quality Checks v1[1]). This began an iterative process whereby ETL code was changed based on the results of the DC query. The ETL code was considered final when there was satisfaction with the foundational data quality[4] of the DataMart. Evaluated against 13 data checking rules and 498 data quality measures, the DataMart had no data check exceptions.

ADAPTABLE feasibility study

To inform the ADAPTABLE study, we used these transformed Medicare data to identify a cohort of patients with coronary artery disease (CAD). Then, within this cohort, we identified clinical events similar to those specified as efficacy or safety endpoints within the ADAPTABLE trial.

We identified an ADAPTABLE-like cohort of CAD patients using the computable phenotype developed for the trial[5]. We used January 1, 2011 as the cohort identification date and, in addition to requiring that patients meet the listed trial inclusion and exclusion criteria, required that beneficiaries were alive and 65 years of age or older as of that date. We additionally required that beneficiaries had been enrolled in fee-for-service Medicare for all of 2010 and had at least 2 or more inpatient, outpatient, or ambulatory encounters in 2010. This was done to simulate a CDRN selecting patients who had recently visited their clinic and may be likely to return for future healthcare services.

ADAPTABLE efficacy endpoints included hospitalization for myocardial infarction, hospitalization for ischemic stroke, hospitalization for hemorrhagic stroke, coronary revascularization, and a composite of all of these. Hospitalization for major bleeding was the ADAPTABLE safety endpoint. We also reported on other endpoints of potential interest, including death and any hospitalization. All endpoints were assessed over a 3-year follow-up period, until December 31, 2013.

We reported baseline characteristics of the cohort by reporting demographics—age, sex, race, ethnicity—and some basic comorbid conditions related to the inclusion criteria—prior myocardial infarction, prior PCI, prior CABG, diabetes mellitus, cerebrovascular disease, peripheral artery disease, systolic heart failure. We also reported a few enrollment-related metrics, such as loss to follow-up and whether or not patients had information about prescription medication in 2010. Within Medicare data, loss to follow-up occurs primarily when a beneficiary enrolls in a Medicare managed care plan. And prescription medication is only known for beneficiaries enrolled in Medicare Part D. Note that we did not limit the cohort identification to patients enrolled in Medicare Part D. This may result in misclassification of some patients as eligible when they are, in fact, ineligible due to concurrent medication use. For example, patients currently on warfarin, but not enrolled in Medicare Part D, will not be excluded.

See **Table 4** for a report of the baseline characteristics and 3-year events for this ADAPTABLE feasibility study cohort. Event rates were substantially higher here than those anticipated within the trial. This may be because the entire population was aged 65 years and older. Additionally, because only 1 year of claims history was searched, there will be many more high-risk patients with a known recent CABG, PCI, or MI in this population than in the trial-enrolled population. This should result in a cohort that is more likely to have the events of interest.

Updated transformation specifications and programming

One of the last things we did within this project was to update the Medicare data transformation specifications and SAS code to reflect changes and additions made within PCORnet CDM v3.1[1]. We plan to continue maintaining this effort to keep up with future versions of the PCORnet CDM and future changes to the Medicare data, if applicable. We also plan to expand the Medicare data utilized within this transformation to include the Skilled Nursing Files (SNF) files.

We wrote up documentation to accompany the SAS programs and created a package that other sites could use with Medicare data they may have ordered. As of May 2018, 2 different sites had tested this code and reported success in using that to transform Medicare data into the PCORnet CDM. The SAS program package and the data transformation specifications are available on GitHub[2].

Phase 2 / Linking PCORnet patients to Medicare data

The second phase of the project involved establishing and testing processes and data flows that would enable multi-institution research using Medicare data for patients identified within PCORnet. For this work, the DCRI, as a coordinating center, partnered with 2 PCORnet sites—Vanderbilt, a member of the Mid-South CDRN, and University of Kansas Medical Center (KUMC), part of the Greater Plains Collaborative CDRN—to submit a multi-institution DUA with CMS with the purpose of obtaining Medicare claims data for a cohort of patients that were similar to those being enrolled in ADAPTABLE trial. Vanderbilt and KUMC were chosen, in part, because they were able to obtain permission to share patient-identifiable information with DCRI for research purposes. This identifiable information was the basis by which patients were to be linked to Medicare beneficiaries. As in the prior project phase, the Medicare data obtained for these linked patients were transformed into the PCORnet CDM and queried.

Governance requirements

Before any data work could begin, we first mapped out the data flows between the partner sites, the coordinating center at DCRI, and GDIT, the CMS contractor responsible for distributing Medicare data to researchers. We then needed to get the required approvals and agreements in place that would allow these sensitive data transfers. These approvals and agreements included data sharing agreements (DSA) between each site and DCRI, a multi-institution DUA submitted jointly by the partner sites and DCRI, Institutional Review Board (IRB) protocols at each site and DCRI, and a social security number (SSN) security exception request at DCRI.

The data flow diagram is shown in **Figure 1**. This diagram shows how patient identifiable information for a specific cohort of patients was sent from both sites to the coordinating center. The coordinating center then created and sent multiple finder files to GDIT, which created a crosswalk between these local identifiers and Medicare identifiers and extracted the relevant Medicare data. The final crosswalks and data extracts were then sent back to the original sites and to the DCRI. The details of the data transferred at each step are described later.

To support these data flows, we first needed to set up data transfer agreements between each of the partner sites and DCRI that allowed the transfer of patient-level data between institutions and to GDIT for processing. This was required since the existing PCORnet data sharing agreements that Vanderbilt and KUMC had with DCRI, as one of the PCORnet coordinating centers, did not cover the sharing of patient-level data.

VUMC reported that establishing a data sharing agreement with DCRI occurred with few barriers. Factors facilitating this agreement included prospective discussion of the flow of data between institutions, the transparent documentation of exactly what data would be shared, and the commitment to efforts to minimize the risk of data loss through only transferring the minimum necessary data to support the project. KUMC reported more difficulty obtaining internal approval for this data sharing agreement, since the request included the transfer of sensitive patient identifiers like SSN and Medicare health insurance claim (HIC) numbers.

Next, each of the 3 institutions involved all needed to obtain IRB approval for this work. At DCRI, this process triggered an additional internal approval process to cover the receipt and storage of the SSN and HIC numbers required for linkage within this project.

The DUA amendment was more involved for this project than is typical for single-institution research, since all 3 participating institutions were going to be receiving Medicare data. Because of this, each institution needed to be a signatory to the DUA and needed to submit separate data management plans to CMS. These data management plans outlined the information technology and security in place to safeguard Medicare data at each site. The DUA process took about 9 months to complete due to this complexity. By way of comparison, a single-site DUA typically takes about 4 months to complete.

Medicare data

We obtained Medicare data—including the Master Beneficiary Summary File, inpatient claims file, outpatient claims file, and carrier claims file—covering the period from January 1, 2014 to September 30, 2015 for a cohort of patients identified by VUMC and KUMC. (Selection criteria for these cohorts are described below.) These data were to be ordered as physical files delivered to each of the 3 institutions party to the DUA.

It is important to note that CMS currently offers 2 methods for researchers to access research-identifiable Medicare data for a known cohort of patients, as was done in this project. The first involves shipping the requested data via physical media to the researcher. The second involves the researcher accessing the requested data via a secure remote computing environment called the Virtual Research Data Center (VRDC).

The key difference between these methods is the actual location of the data. Researchers receiving physical media securely store and analyze the data locally. Researchers using the VRDC access and analyze the data remotely. A key restriction of working in the VRDC is that, for security reasons, only summary-level results or data may be downloaded. Researchers are not permitted to download beneficiary-level data. This means that, when working in the VRDC, any data that needs to be merged or used alongside the Medicare data must be uploaded into the remote environment for analysis.

The costs associated with each method differ as well. When using the VRDC, costs reflect, primarily, access to the data on an annual basis. The actual amount of data accessed is not important for the price of this access. When receiving physical files, on the other hand, the size of the cohort, the number of different data files requested, and the number of years of data requested all are part of the cost. ResDAC (Research Data Assistance Center), a CMS contractor that provides assistance to researchers using Medicare data, offers cost estimates for either method.

Both methods have some common requirements. First, the requesting researcher must have a signed DUA with CMS. The DUA delineates the specific files requested, the proposed use of those files, and the security in place to protect those files, among other things. Second, for each beneficiary whose data are being requested, either their Medicare HIC number or their SSN (partial or full) must be sent to CMS along with some basic demographic information, like date of birth and gender. This allows the CMS data contractor to identify the beneficiary and crosswalk their HIC or SSN to the encrypted beneficiary identifier used for data distributions. While there are encounter-based linking methods that can used for identifying beneficiaries in Medicare claims data without direct identifiers, the data required to make that type of link are broader than what is discussed here.

Site data

Each of the 2 partner sites selected a cohort of patients with coronary artery disease, using criteria similar to those proposed for the ADAPTABLE trial. The selection process at each site differed in important ways. Details of these processes are below.

VUMC cohort description

Given the uncertainty surrounding the completeness of EHR data at VUMC to identify a population of patients with definitive CAD, the VUMC team felt that a more sensitive (i.e., inclusive) definition would be most appropriate for this project. To this end, VUMC defined their cohort using four case definitions for CAD (**Table 5**). All definitions required the subject to be 30 years or greater, and for the relevant service(s) to have been delivered within the past 5 years.

Furthermore, identified patients required the availability of either HIC or SSN for linkage. While characterization of CAD, using the above case definitions, was straightforward within EHR data, characterization of the HIC presented significant challenges. Data characterizing individuals' specific insurance policies (i.e., policy ID, HIC) are not stored in the research derivative files used to build the VUMC cohort, and were inconsistently documented within the source EHR. Ultimately, this problem was solved through building internal crosswalks of files containing individual-level HIC ID to individual-level EHR data. Nonetheless, there remain important unanswered questions surrounding the accuracy and completeness of VUMC HIC data.

KUMC cohort description

KUMC focused on identifying the CAD cohort most likely to participate in ADAPTABLE via electronic recruitment. Among the patients identified by applying the ADAPTABLE computable phenotype to their EHR data, they selected those who had used the patient portal or for whom they had a recent email on file. They further restricted this cohort to patients known to have Medicare insurance coverage. For these patients, they extracted, and sent to the DCRI team, SSNs and HICs—including identifiers labeled as HICs, but that may not have had the typical HIC structure.

Finder files

For the selected patients, both sites generated a patient-level finder file containing SSN (where known), Medicare HIC (where known), date of birth (DOB), sex, and a site-generated patient identifier. Details about these finder files are reported in **Table 6**. Of note, SSNs were universally available for these patients, while HICs were not. Of course, we do not expect patients who are not enrolled in Medicare to have a HIC; and patients enrolled in a Medicare managed care plan may have that managed care plan provider's insurance identifier on file instead of a HIC. Also, even when a value for HIC was included in the finder file, it was not always a proper or correctly formatted HIC.

These files were sent to DCRI using the same format and protocol that is used for sending finder files to GDIT. Specifically, we requested that finder files be created as comma-delimited text files with variable names in the first row, then zipped and encrypted using AES-256 encryption based on a password that was at least 10 characters in length and that contained a combination of letters, numbers, and symbols. The file was then to be burned onto a CD or DVD and shipped using a courier with tracking capabilities.

Upon receipt, DCRI merged the information from both sites and created 3 separate finder files for submission to GDIT. One finder file included HICs; one finder file included full 9-digit SSNs; and one finder file included the last 4 digits of SSNs. All finder files additionally included date of birth, sex, a local patient identifier, and a site identifier. The site identifier distinguished KUMC patients from VUMC patients, which allowed GDIT to split the generated crosswalks and extracted Medicare data appropriately for each site. Different finder files were submitted so that we could evaluate the yield of each different identifier for linkage.

Crosswalk files

GDIT used each of these finder files to identify patients within the Medicare enrollment files. Each finder file generated a crosswalk file that associated the local patient identifier with an encrypted Medicare beneficiary identifier that would identify patients within the extracted Medicare data.

The HIC- and SSN-based crosswalks returned by GDIT require some post-processing, since not all of the linkages in that file are robust. It's true that all records returned in the crosswalk required that HIC, or SSN, as appropriate, matched between data sources, but this is not always enough information to call a link definitive. Reasons for this include the fact that spouses often receive benefits under the same SSN and both can link to a submitted SSN. There also may be data errors in the finder file that lead to erroneous links. To establish robust links, researchers should use the "matched DOB" and "matched sex" flags provided by GDIT in the crosswalk file, and should also cross-check the submitted DOB and sex information against the DOB and sex information in the MBSF. There are six different ways (shown as rules in **Table 7**) that patient DOB and sex data can match between the finder file and CMS data, even when HIC or SSN matches. Not all of these rules lead us to have the same confidence in the link. Clearly, if SSN matches, but neither patient DOB nor sex match, it is difficult to have confidence that the data in each source represent the same patient. Therefore, when determining final links to use for analysis, we limited ourselves to records that matched using Rules 1, 2, or 3.

The results of the 3 linkage methods are shown in **Table 8**. For the HIC finder file, the match rate to any Medicare beneficiary was very high. In all but a few cases, the patients with matching HICs also had matched DOB and sex. For the full SSN finder file, the match rate to any Medicare beneficiary was also high, but there were more matches where DOB and/or sex did not align. The difference in match rates associated with SSNs between KUMC and Vanderbilt appears to be due to KUMC's selection of patients known to be enrolled in Medicare. For the partial SSN finder file, the match rate to any Medicare beneficiary was substantially lower—about 25% lower—than the match rate based on full SSN. This was because all three linking fields—partial SSN, DOB, and sex—were required to match exactly for GDIT to return a link, and this combination was not always unique within the Medicare enrollment file. For easier comparison, the yield of each method—limiting only Rules 1, 2, and 3 for the SSN- and HIC-based links—is shown in **Table 9**.

To compare the linking methods directly, **Table 10** shows the numbers of patients identified or not identified by each the different finder files. These counts demonstrates that linking on the full SSN seems most robust. There were very few patients not identified with a full SSN who were only identified by linking on the HIC and/or partial SSN, whereas many patients without known HIC information were identified by their SSN. For subsequent analysis, a single crosswalk file was created by including patients with high-quality links to CMS beneficiaries via any method—HIC, full SSN, or partial SSN.

Site confirmation of crosswalk

Based on our experience reconciling these crosswalk files to establish definitive links for patients in the submitted finder files, we asked, and provided guidance to, Vanderbilt and KUMC to process the 3 finder files as we had. For multi-site research using Medicare data, this step is essential, as it establishes a common crosswalk to be used across participating sites. There was more complexity here than there would be typically, however, since we asked them to both process and reconcile potential links across 3 crosswalk files. Typically, sites will only be dealing with a single crosswalk file. **Table 11** shows the results of this confirmatory work. Most links were identical, but there were some inconsistencies. For example, at both sites there were patients who had viable links to multiple beneficiary identifiers; and in some of these cases, we chose a different link than the site chose. Other reasons for mismatches include the site selecting a link from one crosswalk when a better link existed from a different crosswalk file.

Linked Medicare data

For patients in our study cohort with a high quality link to a Medicare beneficiary, **Table 12** shows some basic facts about their Medicare data. First, not all of these patients had a demographic information. This is a common situation that arises because GDIT links patients from finder files using the most current Medicare enrollment information available. If someone first enrolled in Medicare at some point after the dates covered by the data request, the requester will likely not have any MBSF records for that person. Second, many linked patients appear to have no enrollment information. This is due to the different types of Medicare programs available to beneficiaries. We only received claims data for patients while they are enrolled in fee-for-service Medicare, so entries in the PCORnet Enrollment table reflect these periods. We did not have any utilization information for patients while they are enrolled in a Medicare managed care plan, so no enrollment record was be generated to cover these periods. It is important to know that we did have information about mortality for all patients, however, regardless of the type of Medicare program(s) they have enrolled in. Third, for the event rate calculation, we set a start date at 01-Jan-2014, so only the patients enrolled in fee-for-service Medicare at that time were included. Finally, most patients included in the analysis had complete data through Sept 30, 2015. The few patients without complete data had transitioned from a fee-for-service plan to a managed care plan. This low switching rate is fairly common in Medicare data.

Table 13 shows the demographics and 21-month event rates for this linked cohort of patients with CAD.

Future work / Medicare data for ADAPTABLE

This project has successfully established processes that will enable the use of Medicare data within ADAPTABLE and other pragmatic clinical trials. For patients at participating PCORnet sites enrolled in the ADAPTABLE trial, we will be collecting identifiable information to send to GDIT to enable linkage with Medicare beneficiaries. We will then, for linked patients, receive Medicare data, which will be transformed into the PCORnet CDM. The same SAS query that will be sent to sites to ascertain trial events will also be run against these data. The results will be sent to the trial statistician for incorporation into the main trial dataset.

Recommendations

For PCORnet studies or sites that plan on using Medicare data, we offer the following recommendations:

(1) *Transform the Medicare data into the PCORnet CDM format using the programs and specifications developed within this project*

The information and programs we have released are a product of many years of experience working with these data. Transforming Medicare data enables researchers familiar with the PCORnet CDM to easily analyze Medicare data. It also allows researchers to run the same query on the Medicare data that is run within the PCORnet Distributed Research Network.

(2) *Allow ample time for preparation and processing of the CMS Data Use Agreement*

Current processing times of 6 months are not uncommon.

(3) *For identification of patients within the Medicare data, Social Security numbers (SSN) are most reliable*

Medicare HICs are also reliable, if sites are able to locate that information.

Acknowledgements

This work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award for development of the National Patient-Centered Clinical Research Network, known as PCORnet.

The views presented in this work are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee or other participants in PCORnet.

We thank the entire project team that made this work possible. In addition to the authors, this includes Chantelle Hardy at the Duke Clinical Research Institute; Brittany Zschoche, Nathan Graham, Steve Fennel, and Sravani Chandaka at the Kansas University Medical Center; and Jacqueline Kirby, Srinivas Chiguluri, Katherine Whorley at the Vanderbilt University Medical Center.

We also thank Maryan Zirkle and Shayna Barbash at the Patient-Centered Outcomes Research Institute.

References

[1] PCORnet Common Data Model Specifications and Data Quality Checks. http://pcornet.org/pcornet-common-data-model/

[2] Medicare to PCORnet Data Transformation Package, GitHub. https://github.com/PCORnet-DRN-OC/Medicare-Data-Transformation

[3] CMS Linkage Pilot Data Transformation Development Information, iMeet Central. https://pcornet.imeetcentral.com/cmslinkagepilot/folder/WzIwLDYzOTI4Njld/

[4] Qualls, L.G., Phillips, T.A., Hammill, B.G., Topping, J., Louzao, D.M., Brown, J.S., Curtis, L.H. and Marsolo, K., 2018. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®). *eGEMs*, *6*(1). http://doi.org/10.5334/egems.199

[5] ADAPTABLE Computable Phenotype, GitHub. https://github.com/ADAPTABLETRIAL/PHENOTYPE

Table 1. Relationship between Medicare Data Files and PCORnet CDM Tables

	Source Medicare Data File				
Destination PCORnet CDM Table	Master Beneficiary Summary File	Inpatient Claims	Outpatient Claims	Carrier Claims (Physician/ Professional)	Part D Events (Prescription Medication)
Demographic	X				
Enrollment	X				
Encounter		X	X	X	
Diagnosis		X	X	X	
Procedures		X	X	X	
Dispensing					X
Death	X				

Table 2. Changes to PCORnet CDM v3.1 informed by this project

PCORnet Table	PCORnet Field	New Value Set Item	New Value Set Item Description
Enrollment	Enrollment basis [ENR_BASIS]	D = Outpatient prescription drug coverage	The start and stop dates are based on enrollment where the health plan has any responsibility for covering outpatient prescription drugs for the member during this enrollment period.
Encounter	Encounter type [ENC_TYPE]	IC = Institutional professional consult	Permissible substitution when services provided by a medical professional cannot be combined with the given encounter record, such as a specialist consult in an inpatient setting; this situation can be common with claims data sources.

Table 3. Example mapping of Medicare RACE field to PCORnet CDM RACE and HISPANIC fields

Medicare RACE	Mapping to PCORnet…	
	RACE	HISPANIC
Asian	Asian	Other
Black	Black or African American	Other
Hispanic	Other	Yes
North American Native	American Indian or Alaska Native	Other
White	White	Other
Other	Other	Other
Unknown	Unknown	Unknown
[Missing]	No Information	No Information

Table 4. Baseline Characteristics and 3-Year Events for ADAPTABLE-like Cohort from 5% Medicare Data

Variable	*Overall*
N	81,748
Demographics	
Patient age, Median (Q1, Q3)	77.0 (72.0, 83.0)
Sex	
Female	35,529 (43.5)
Male	46,219 (56.5)
Race	
American Indian or Alaska Native	342 (0.4)
Asian	896 (1.1)
Black or African American	4653 (5.7)
White	73,874 (90.4)
Other/Unknown	1983 (2.4)
Hispanic Ethnicity	1090 (1.3)
Medical History	
Prior CABG	32,047 (39.2)
Prior MI	40,245 (49.2)
Prior PCI	31,689 (38.8)
Cerebrovascular disease	26,638 (32.6)
Diabetes	34,881 (42.7)
LV Systolic Dysfunction	8751 (10.7)
Peripheral arterial disease	20,235 (24.8)
Events, 3 years	
Myocardial infarction	3940 (5.0)
Hemorrhagic stroke	455 (0.6)
Ischemic stroke	2434 (3.1)
Death	19,636 (24.0)
Composite MI, stroke, death	23,472 (29.5)
Revascularization	6870 (8.6)
Bleeding	2160 (2.7)
Any hospitalization	48,730 (60.9)
Other info	
Censored before 31-Dec-2013	4620 (5.7)
Incomplete or missing 2010 Rx info	38,757 (47.4)

Results presented as N (%) unless otherwise noted

Table 5. VUMC CAD Case Definitions

Case Definition	Criteria
1	ICD-9-CM diagnosis code for MI or CAD paired with an outpatient CPT code for E&M (evaluation and management) service (e.g., office visit, ED visit)
2	ICD-9-CM procedure code or CPT code for percutaneous revascularization
3	ICD-9-CM procedure code or CPT code for coronary artery bypass graft(s)
4	ICD-9-CM diagnosis code for MI paired with an inpatient admission of longer than 48 hours

Table 6. Information about records received

	Site A	Site B
Records received	1869	39,843
Missing SSN	0 (0.0)	0 (0.0)
Missing HIC	932 (49.9)	2987 (7.5)
Missing DOB	0 (0.0)	0 (0.0)
Missing or Unknown Sex	0 (0.0)	4 (<0.1)
Valid format, SSNs	1869 (100)	39,843 (100)
Valid format, HICs	924 (49.4)	24,463 (61.4)
Distinct patient IDs	1868	30,571
Age ≥ 65 years	1208 (64.7)	26,245 (85.9)
Distinct SSNs	1859	30,564
Distinct HICs, any	937	36,844
Distinct HICs, correct format	924	24,462
Distinct Patient ID + SSN	1868	30,571
Distinct Patient ID + HIC, any	937	36,846
Distinct Patient ID + HIC, correct format	924	24,463

Results presented as N or N (%)

Table 7. Matching rules for assessing linkage quality

	Comparison between data sources for…	
Matching Rule	**Date of Birth**[a]	**Sex**
1	Exactly matched	Matched
2	Inexactly matched	Matched
3	Exactly matched	Unmatched
4	Inexactly matched	Unmatched
5	Unmatched	Matched
6	Unmatched	Unmatched

[a] An exact match on DOB is where day, month, and year are all equal between data sources. An inexact match on DOB is where any two of the three components—month, day, or year of birth—were equal between data sources. Any other situation—1 component equal or no components equal—was considered unmatched.

Table 8. Results of linkage using different information

	Site A	Site B
Linkage using CMS HIC		
Distinct Patient ID + HIC, any	937	36,846
Matched HIC	923 (99.5)	24,296 (65.9)
Distinct Patient ID + HIC, correct format	924	24,463
Matched HIC	923 (99.9)	24,296 (99.3)
Among matched HICs		
Rule 1: Match DOB (exact) & sex	921 (99.8)	24,054 (99.0)
Rule 2: Match DOB (inexact) & sex	2 (0.2)	160 (0.7)
Rule 3: Match DOB (exact) only	0 (0.0)	20 (0.1)
Rule 4: Match DOB (inexact) only	0 (0.0)	1 (<0.1)
Rule 5: Match sex only	0 (0.0)	46 (0.2)
Rule 6: Unmatched DOB & Sex	0 (0.0)	15 (<0.1)
Linkage using Full SSN		
Distinct Patient ID + SSN	1868	30,571
Matched SSN	1353 (72.4)	30,427 (99.5)
Among matched SSNs		
Rule 1: Match DOB (exact) & sex	1344 (99.3)	30,078 (98.9)
Rule 2: Match DOB (inexact) & sex	3 (0.2)	202 (0.7)
Rule 3: Match DOB (exact) only	0 (0.0)	22 (0.1)
Rule 4: Match DOB (inexact) only	0 (0.0)	0 (0.0)
Rule 5: Match sex only	4 (0.3)	75 (0.2)
Rule 6: Unmatched DOB & Sex	2 (0.1)	50 (0.2)
Linkage using Partial SSN		
Distinct Patient ID + SSN	1868	30,571
Matched Partial SSN & DOB & Sex	1024 (54.8)	22,303 (73.0)

Results presented as N or N (%)

Table 9. Patient-level results of different linkage methods

Method	Site A	Site B
Distinct patients in finder file	1868	30,571
HIC (Rules 1–3 only)	923 (49.4)	23,853 (78.0)
SSN (Rules 1–3 only)	1347 (72.1)	30,302 (99.1)
SSN-4 (Exact DOB & sex only)	1024 (54.8)	22,303 (73.0)
Linked by any of the 3 above methods	1398 (74.8)	30,500 (99.8)

Results presented as N or N (%)

Table 10. Counts of patients identified by different combinations of linking methods

Patient ID linked by…				
HIC	**Full SSN**	**Partial SSN**	**Site A**	**Site B**
No	No	Yes	44	42
No	Yes	No	128	1790
No	Yes	Yes	303	4815
Yes	No	No	1	99
Yes	No	Yes	6	57
Yes	Yes	No	245	6308
Yes	Yes	Yes	671	17,389

Table 11. Comparison of Patients in Coordinating Center and PCORnet Site Crosswalks

Site	*Coord Ctr +* *PCORnet +*	*Coord Ctr +* *PCORnet –*	*Coord Ctr –* *PCORnet +*	*Coord Ctr –* *PCORnet –*
A	1306	5[a]	94[b]	0
B	30,071	15[c]	31[d]	0
Overall	31,377	20	125	0

[a] Site A: Coord Ctr + / PCORnet –

- 4 = Acceptable link for HIC, but not SSN; Site chose unacceptable SSN link
- 1 = Multiple possible SSN-BENE_ID links; Different link chosen from coordinating center

[b] Site A: Coord Ctr – / PCORnet +

- 87 = No Medicare denominator information; Coordinating center did not include these patients in the Medicare DataMart
- 6 = Unacceptable SSN links accepted
- 1 = Multiple possible SSN-BENE_ID links; Different link chosen from site

[c] Site B: Coord Ctr + / PCORnet –

- 15 = Multiple possible SSN-BENE_ID links; Different link chosen from coordinating center

[d] Site B: Coord Ctr + / PCORnet –

- 15 = Multiple possible SSN-BENE_ID links; Different link chosen from site
- 16 = Acceptable links based on partial SSN; Not used by site

Table 12. Crosswalk and enrollment records for PCORnet-CMS cohorts

Count of patients…	**Site A**	**Site B**
…in CMS crosswalk[1]	1398	30,499
…with a CDM Demographic table record[2]	1311	30,102
…with any CDM Enrollment records[3]	995	21,898
…enrolled on 01-Jan-2014 (start of CMS data)	845	20,671
…continuously enrolled from 01-Jan-2014 to 30-Sep-2015 (end of CMS data)[4]	827	19,740

[1] Includes patients who linked to CMS beneficiaries via any method—HIC, full SSN, partial SSN. Links were limited to those deemed to be of high quality—with Rule 1 (Exact DOB + sex), Rule 2 (Inexact DOB + sex), and Rule 3 (Exact DOB only) only (*see other document*).
[2] A small number of patients matched to CMS beneficiaries who enrolled in Medicare for the first time after 30-Sep-2015. These patients would not appear in the Demographic table.
[3] Enrollment is defined as being enrolled in fee-for-service Medicare Parts A & B. A substantial proportion of CMS beneficiaries identified were never enrolled in fee-for-service Medicare between 01-Jan-2014 and 30-Sep-2015. Patients not enrolled in fee-for-service Medicare Parts A & B may be enrolled in either a Medicare managed care plan or in fee-for-service Part A (hospital insurance) only.
[4] Continuous enrollment includes patients who died while enrolled during this time period.

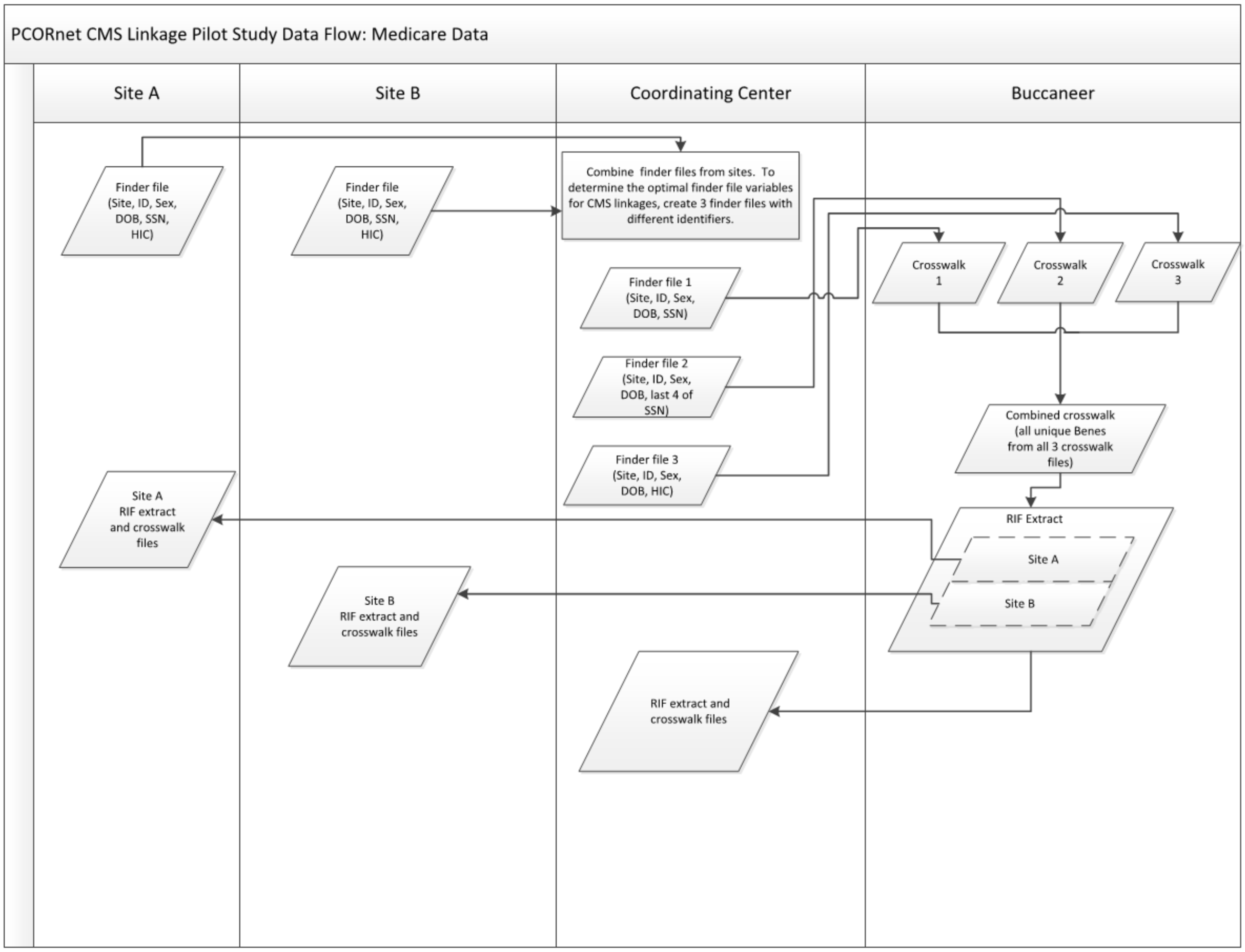
Table 13. Demographics and 21-Month Events for a CAD cohort Enrolled in FFS Medicare

Variable	Site A	Site B
Enrolled on 01-Jan-2014, N	845	20,671
Demographics		
Patient age, Median (Q1, Q3)	71.0 (66.0, 77.0)	72.0 (67.0, 79.0)
Sex		
Female	31.4	38.7
Male	68.6	61.3
Race		
American Indian or Alaska Native	≤1.2	0.1
Asian	≤1.2	0.3
Black or African American	6.3	6.4
White	89.9	92.0
Other/Unknown	2.6	1.3
Hispanic Ethnicity	≤1.2	0.2
Events, 21 months		
Myocardial infarction	4.2	4.3
Hemorrhagic stroke	≤1.2	0.5
Ischemic stroke	≤1.2	2.0
Death	2.4	14.4
Composite MI, stroke, death	7.4	19.0
Revascularization	11.6	8.2
Bleeding	≤1.2	2.1
Any hospitalization	45.2	48.7
Other info		
Censored before 30-Sep-2015	2.1	5.5

Results presented as % unless otherwise noted

[*] Ns are not displayed since many cells included ≤10 patients, which cannot be explicitly reported, per CMS rules.

Figure 1. Data flows within the PCORnet-CMS linkage project


PCORnet CMS Linkage Pilot Study Data Flow: Medicare Data
Site A
Site B
Coordinating Center
Buccaneer
Finder file (Site, ID, Sex, DOB, SSN, HIC)
Finder file (Site, ID, Sex, DOB, SSN, HIC)
Combine finder files from sites. To determine the optimal finder file variables for CMS linkages, create 3 finder files with different identifiers.
Finder file 1 (Site, ID, Sex, DOB, SSN)
Finder file 2 (Site, ID, Sex, DOB, last 4 of SSN)
Finder file 3 (Site, ID, Sex, DOB, HIC)
Crosswalk 1
Crosswalk 2
Crosswalk 3
Combined crosswalk (all unique Benes from all 3 crosswalk files)
RIF Extract
Site A
Site B
Site A RIF extract and crosswalk files
Site B RIF extract and crosswalk files
RIF extract and crosswalk files