PharmaSUG 2023 - Paper RW-050

Real World Evidence in Distributed Data Networks

Lessons from a Post-Marketing Safety Study

Matthew T. Slaughter, Denis B. Nyongesa, John F. Dickerson, and Jennifer L. Kuntz,
Kaiser Permanente Center for Health Research

ABSTRACT

As a case study to illustrate both opportunities and challenges in distributed data networks, this paper will focus on the implementation of a post-marketing safety study in the Healthcare Systems Research Network (HCSRN) via the associated Virtual Data Warehouse (VDW) common data model. In response to an FDA post-marketing requirement, this study establishes the incidence of angioedema in chronic heart failure patients treated with sacubitril/valsartan and incorporates data from multiple distributed data networks, including HCSRN.

Distributed data networks present exciting opportunities for gathering real-world evidence by pooling standardized datasets across institutions. Common data models facilitate the efficient allocation of programming work to conduct analysis while allowing participating sites to retain control of their own data. However, large-scale and high-quality data collection combining data from disparate health data systems presents technical, administrative, and scientific challenges.

In addition to describing programming, data management, and validation techniques used by HCSRN analysts in this study, we will compare design choices made by the various data networks involved in the project and explore their practical consequences.

INTRODUCTION

DISTRIBUTED DATA NETWORKS

A distributed (or federated) data network is a system for cooperative data collection and analysis without the need for a centralized repository. Each member institution in the network retains control of their own data protected behind their respective firewalls and agrees to curate a set of standardized datasets according to a common data model [6]. The network facilitates pooling of data across sites for analysis by the distribution of standardized code which local programmers execute at each site. IRB approval is required for research data extraction and contractual agreements between sites are prerequisite for data transfer [8]. This structure safeguards compliance with human subjects research regulations and respects institutional interests.

Government agencies such as the Centers for Disease Control and Prevention (CDC) (Vaccine Safety Datalink [2]) and the US Food and Drug Administration (FDA) (Sentinel [1][7]) lead distributed data networks for safety surveillance monitoring, while other networks are formed by self-organized coalitions of healthcare organizations to provide larger, more representative populations and higher quality data for research.

The Health Care Systems Research Network (HCSRN) is a consortium of 20 health care systems and their associated research programs, formed to facilitate public domain health and health services research (https://hcsrn.org/). The HCSRN VDW common data model currently consists of 23 tables for standard clinical data elements such as diagnoses, medication dispenses, laboratory test, procedures, and demographics (see appendix). Studies conducted in the HCSRN benefit from regular QA performed by VDW work groups in the various data domains, while insights generated by individual studies are incorporated into the VDW to benefit future work.

SACUBITRIL/VALSARTAN (ENTRESTO®)

Angiotensin converting enzyme (ACE) inhibitors are commonly prescribed to treat heart failure, while angiotensin receptor blockers (ARBs) may be prescribed as an alternative for patients who cannot tolerate the side effects of ACE inhibitors. Novartis developed sacubitril/valsartan (or Entresto®), a drug combining an ARB (valsartan) with a neprilysin inhibitor (sacubitril), as a novel treatment for chronic heart failure with reduced ejection fraction (HFrEF). The PARADIGM-HF trial, which compared sacubitril/valsartan to an ACE inhibitor (enalapril), stopped early after finding reduced risk of death or heart failure hospitalization among patients receiving sacubitril/valsartan [5].

At the time of FDA approval in July 2015, due to results of the PARADIGM trial and a known increased risk of ACEI-associated angioedema among Black patients, the FDA issued a post-marketing requirement (PMR) to conduct an epidemiologic study to evaluate the incidence of angioedema among Black patients with heart failure receiving sacubitril/valsartan compared to a control drug. To fulfill the PMR, Novartis initiated a database cohort study to assess the risk of serious angioedema following sacubitril/valsartan or ACEI initiation. The HCSRN was a logical setting to conduct such a study due to large populations across diverse sites with rich data on patient characteristics and healthcare utilization [3]. Researchers from four Kaiser Permanente regions and the Henry Ford Health System agreed to participate[1].

However, due to slow uptake of sacubitril/valsartan and the low rate of serious angioedema events, even this large group proved to possess an insufficient sample size, and the study was expanded to include data from 5 network partners within the Innovation in Medical Evidence and Surveillance (IMEDS) Distributed Database Network (https://reaganudall.org/programs/research/about-imeds) and Center for Medicaid and Medicare Services (CMS) Medicare fee-for-service (FFS) data. The combination of data from these sources in an overall meta-analysis offers a unique opportunity to explore differences between distributed data networks resulting from design decisions, ethos, and the nature of underlying data sources.

MULTISITE COLLABORATION IN HCSRN

In a typical research project in the HCSRN, one site takes the role of the “lead site”, coordinating data collection among the sites participating in the study and taking primary responsibility for managing the relationship with the study sponsor. The lead site establishes data use agreements (DUAs) or other contracts with the contributing sites to govern data transfer and stewardship of contributing sites’ data for the duration of the project. In the HCSRN, DUAs typically allow for the transfer of limited patient-level datasets to the lead site for analysis, but data transfer may be restricted to deidentified aggregate data or may include full PHI depending on the needs of the study in question. Study dataset received from the contributing sites are either archived or destroyed at the end of a project, as stipulated in the DUA.

Study staff at each site minimally includes an investigator and a programmer, and often a project manager as well. The expertise of local investigators and programmers is critical as they are most familiar with their population, health system, and any idiosyncrasies of the local data. Local programmers are required to run any distributed code and validate the results, before returning them to the lead site.

PROGRAMMING FOR MULTISITE DATA COLLECTION

For an HCSRN multisite research project, the lead site programmer is expected to perform the bulk of programming for data collection and analysis. While developing programs for data collection utilizing the lead site’s VDW, all code needs to be written so that it can be distributed to and run at the contributing sites with minimal changes. This means avoiding programming techniques which depend on knowledge of the local operating system, hardware, or use of a specific SAS® interface. In the HCSRN, the various sites mostly either run SAS on Windows or Unix, which have different conventions for directory and file paths. Techniques for in-memory analytics, such as the hash object should be used only with caution, as the lead site programmer has no knowledge of the size of available memory on computers at the contributing sites. Interface-specific features, such as the DM statement (specific to the SAS Windowing

[1] These sites were members of the Cardiovascular Research Network (CVRN), a subset of HCSRN sites participating in NIH-funded cardiovascular research [4].

environment) cannot be included in distributed code, which may need to run in batch mode or through another SAS interface at other sites because non-universal statements will cause the program to fail.

Because the VDW data model does not enforce the use of a specific DBMS platform, programmers must rely on PROC SQL as a vendor-neutral query engine. This means explicit passthrough should be avoided [10] and SQL should be written with the goal of achieving implicit passthrough when possible [11]. In other words, when writing SQL for an unknown DBMS in SAS, use simple queries which avoid using SAS functions to encourage SAS to pass processing to the DBMS as much as possible. In cases where it is not known whether the target database is an Oracle platform, programmers should keep in mind that some columns expected to be dates may instead be returned as datetimes in SAS.

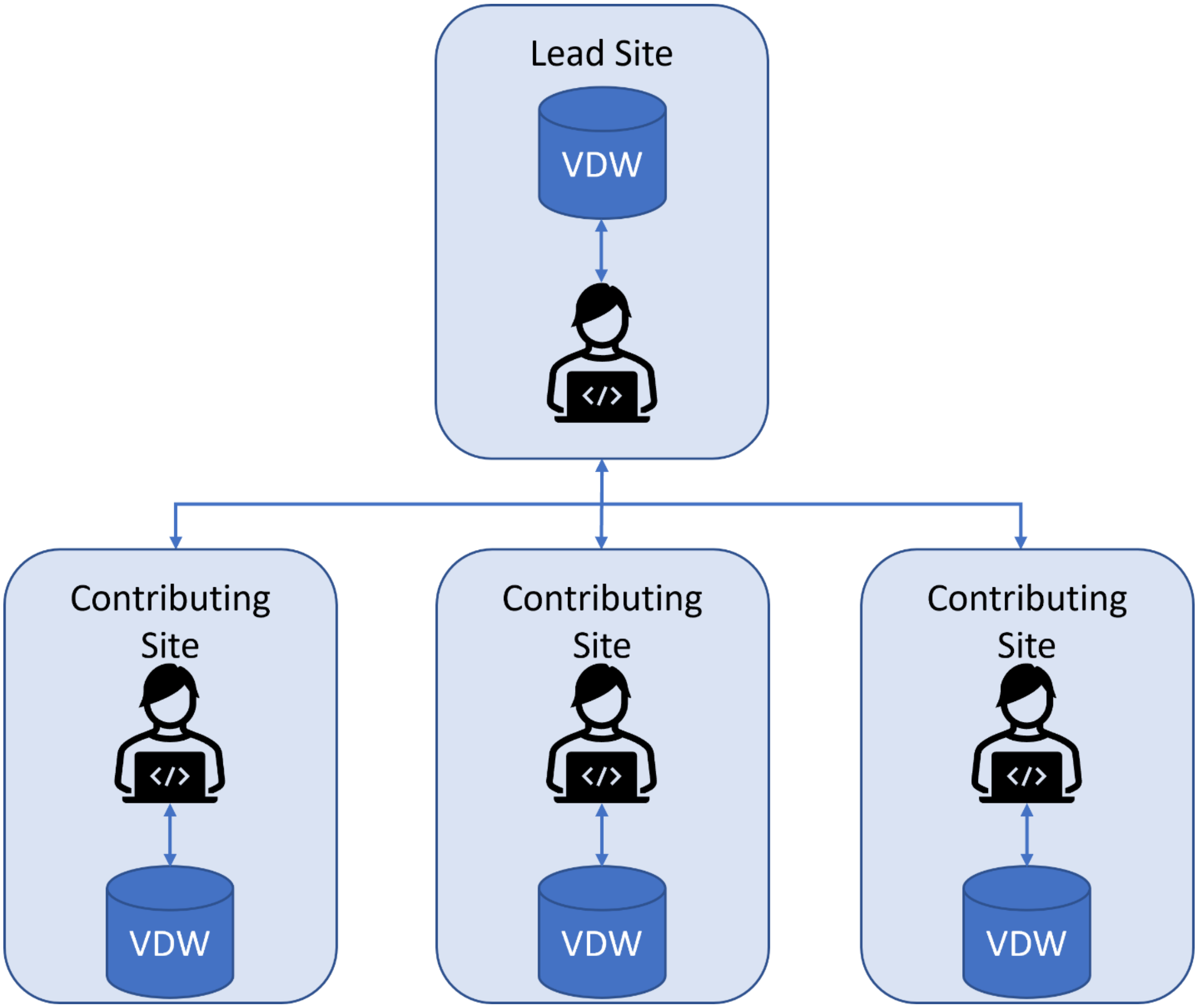


Figure 1. Sharing code and data in the HCSRN.

BEYOND THE VDW

Ultimately, the lead site relies on local programmers at the contributing sites to validate the results, report and fix unanticipated problems, and highlight any issues at their site which the lead site programmer fails to anticipate. The site programmers also play a key role in collecting additional data not included in the VDW data model, such as clinical text notes. For instance, on this project, site programmers collected data on ejection fraction among heart failure patients, an important measure not available in the VDW common data model but of particular interest since sacubitril/valsartan was originally approved for heart failure patients with reduced ejection fraction.

CHALLENGES AND LESSONS

DOUBLE PROGRAMMING

In a distributed data network study, programs to define cohorts and collect associated baseline and outcomes data run separately at each site. Two programs which produce identical results locally can potentially produce divergent results when run remotely on other sites' data not used in program development. Thus, to assess the results of a double programming exercise in a distributed cohort study, the results need to be collated across sites before comparison between the work of the two programmers can be fully assessed. If any discrepancies are discovered, resolution may require the participation of the local programmers at the contributing sites because neither the lead site programmer nor the validation programmer have direct access to the underlying data.

Lead site programmers can (and do) go through the normal double programming process locally first, before distributing code to the other sites. The validation process is not considered complete until data collection has been validated across all sites, and it is reasonable to expect that this process may extend the timeline of validation.

COHORT DEFINITION AND ATTRITION

Any change to the cohort definition has the potential to require distribution of new code to all of the contributing sites, and this is not surprising although it may be inconvenient. What is less obvious is that seemingly minor changes to reporting of metadata about the process of cohort definition can also force the full process to be repeated across all sites. For example, the addition of a separate cohort attrition diagram counting only Black patients was impossible without running a new, modified cohort program across all sites, even though the target cohort did not change at all, simply because the original program did not anticipate this reporting requirement. By contrast, division of the cohort into additional new subgroups for analysis was a trivial task because it required no new data or metadata from the contributing sites. Changes which could be trivial to perform locally could become time consuming when the participation of programmers at all sites is required.

CLINICAL PRACTICE TRUMPS DATA INFRASTRUCTURE

As previously noted, clinical trial evidence for the benefits of sacubitril/valsartan among patients with reduced ejection fraction was overwhelming, and the drug was widely expected to be wildly successful. In reality, adoption of sacubitril/valsartan in the HCSRN sites proved slow, only ramping up after the adoption of changing recommendations from the American College of Cardiology and the American Heart Association [12]. This led to smaller than anticipated samples sizes and a need to incorporate data from IMEDS and Medicare FFS, with particularly strong uptake of sacubitril/valsartan in the latter.

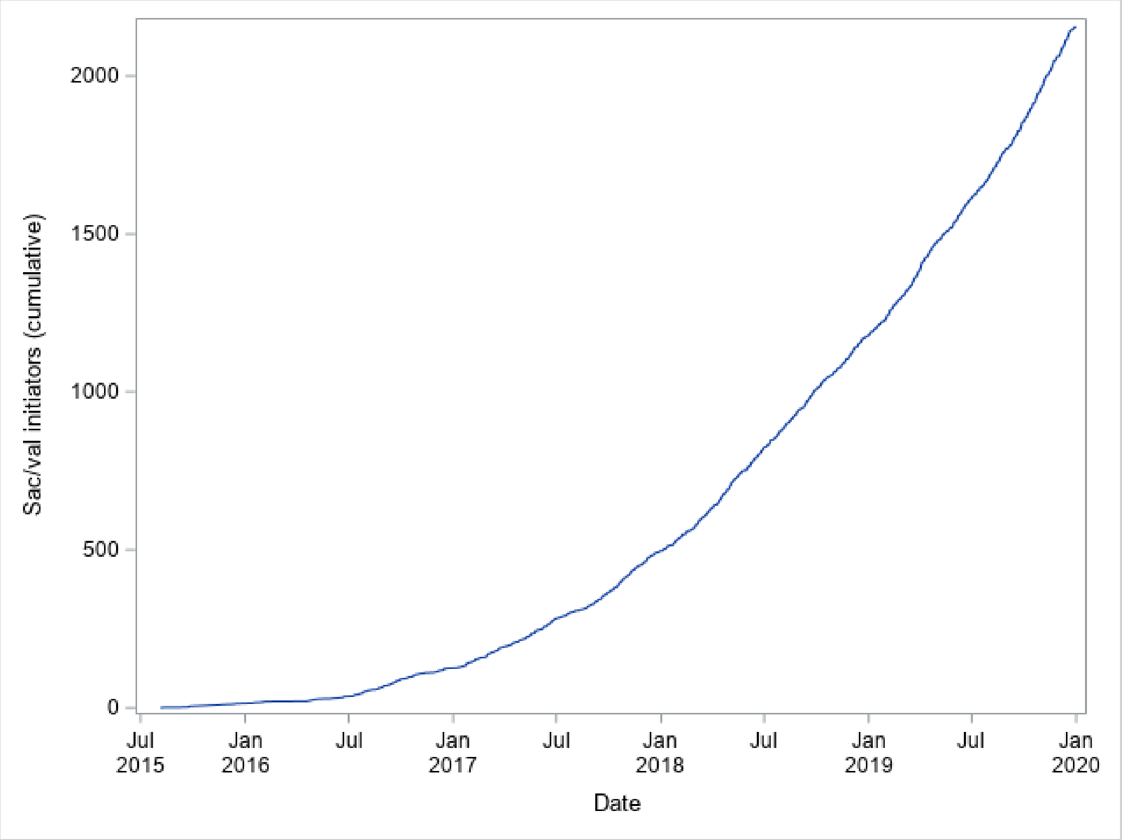


Figure 2. Cumulative counts of sacubitril/valsartan initiators among HCSRN sites over time

THE CONSEQUENCES OF NETWORK DESIGN AND MEMBERSHIP

Differences between the HCSRN VDW and the Sentinel common data model used by IMEDS do have some practical consequences, as do differences in the health systems supplying the data. Unlike HCSRN, Sentinel data is always stored as SAS data sets at each site. This gives the local data management teams less flexibility to choose their preferred tools, but has advantages for study programmers in that it allows for a higher level of standardization. IMEDS programmers never have to wonder about whether or not they might be querying an Oracle database which might return an errant datetime value where one was not expected. Also unlike HCSRN, distributed code in IMEDS typically only returns aggregate data to the lead site. This can make it more difficult to make changes without repeating the process of running the distributed program, but it also means more sites may be willing to participate in a study which might otherwise be reluctant to share individual level data.

Drawing mostly from integrated healthcare system EHRs, HCSRN was the only site to provide certain data elements such as ejection fraction and clinical text notes. IMEDS, by contrast, largely draws data from claims but provided a larger sample size. Medicare FFS, while providing by far the largest sample size, was required to censor rates of angioedema in some sub-cohorts due to low cell counts. This presents a conundrum; high sample sizes are required to study rare events, but the data source with the largest sample sizes was often required not to report those same rare event counts.

CONCLUSION

Distributed data networks are ideal for large cohort studies due to large sample sizes and standardized, high quality data. However, the distributed nature of these networks provides practical challenges not present in other programming environments. Relying on federated data from multiple sites can increase time and effort to complete certain types of validation such as double programming and increases the cost of making retroactive changes to the cohort definition or reporting on the data collection process.

REFERENCES

[1] Curtis LH, Weiner MG, Boudreau DM, Cooper WO, Daniel GW, Nair VP, et al. “Design considerations, architecture, and use of the Mini-Sentinel distributed data system.” *Pharmacoepidemiology and drug safety*. 2012;21 Suppl 1:23-31.

[2] DeStefano F, Vaccine Safety Datalink Research G. The “Vaccine Safety Datalink project.” *Pharmacoepidemiology and drug safety*. 2001;10(5):403-6.

[3] Kuntz JL, Connolly J, Rodriguez-Watson C, Lem JA. “Database cohort study to assess the risk of serious angioedema in association with LCZ696 (sacubitril/valsartan; Entresto®) use in Black patients with heart failure in the United States – Final report, v 1.0”, 12JUL2022.

[4] Magid DJ, Gurwitz JH, Rumsfeld JS, Go AS. “Creating a research data network for cardiovascular disease: the CVRN.” *Expert review of cardiovascular therapy*. 2008;6(8):1043–5.

[5] McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, Investigators PARADIGM-HF, Committees, et al. “Angiotensin-neprilysin inhibition versus enalapril in heart failure.” *N Engl J Med*. 2014;371(11):993–1004.

[6] Popovich, JR. “Multi-site Public and Population Health Research: Analytic Lessons From Distributed Data Networks” Proceedings of the 2019 SAS Global Forum. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3270-2019.pdf Accessed 27MAR2023.

[7] Popovich, JR “Programming in a Distributed Data Network Environment: A Perspective from the Mini-Sentinel Pilot Project” *Proceedings of the 2014 SAS Global Forum*. https://support.sas.com/resources/papers/proceedings14/1772-2014.pdf Accessed 27MAR2023.

[8] Ross, Tyler R., Daniel Ng, Jeffrey S. Brown, Roy Pardee, Mark C. Hornbrook, Gene Hart, and John F. Steiner. 2014. “The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration.” *EGEMS* 2 (1): 1049. doi:10.13063/2327-9214.1049.

[9] SAS Institute. “DM Statement.” *Dictionary of SAS Global Statements*. Accessed 30MAR2023. Available at https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/lestmtsglobal/p1puvpmmlcwvfqn1mvfkz3e0qvan.htm

[10] SAS Institute. “Using Explicit Pass-Through.” *FedSQL Reference*. Accessed 27MAR2023. Available at https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/fedsqlref/p1t7brk6e1lguwn1j4icn4s058h6.htm

[11] SAS Institute. “Implicit SQL Pass-Through.” *FedSQL Reference*. Accessed 27MAR2023. Available at https://documentation.sas.com/doc/en/spdsug/5.5/n095mdd1wof6ogn1neyglan77ghc.htm#p0nrt0oa3dx4bmn1km90a7vlshyi

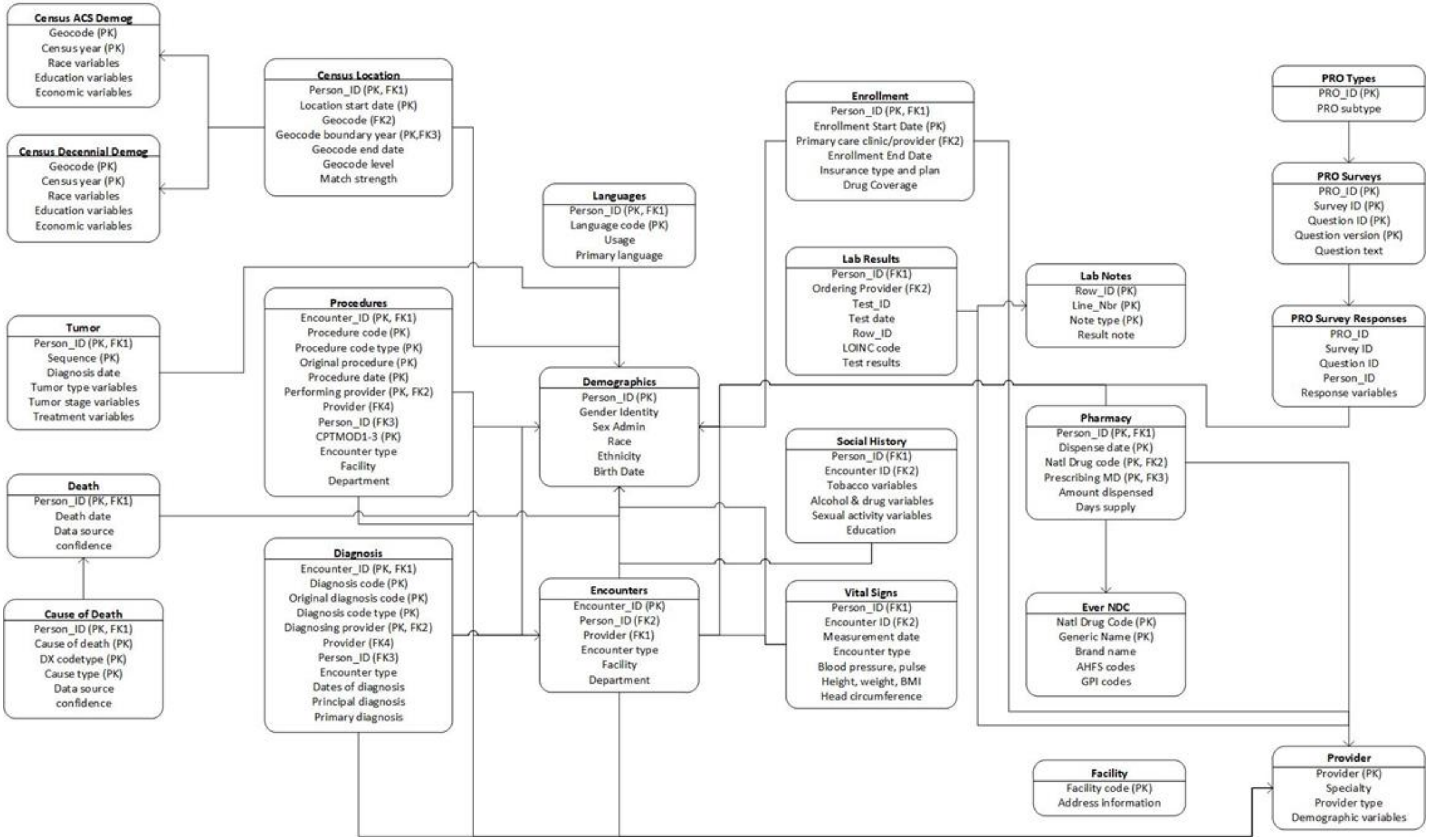
[12] Yancy CW, Jessup M, Bozkurt B, Butler J, CaseyJr DE, Colvin MM, Drazner MH, et al. “Focused Update on New Pharmacological Therapy for Heart Failure” *Circulation* 134 (13) https://doi.org/10.1161/CIR.0000000000000435

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matthew T. Slaughter
Kaiser Permanente Center for Health Research
Matthew.T.Slaughter@kpchr.org

APPENDIX


Census ACS Demog
Geocode (PK)
Census year (PK)
Race variables
Education variables
Economic variables
Census Decennial Demog
Geocode (PK)
Census year (PK)
Race variables
Education variables
Economic variables
Census Location
Person_ID (PK, FK1)
Location start date (PK)
Geocode (FK2)
Geocode boundary year (PK,FK3)
Geocode end date
Geocode level
Match strength
Languages
Person_ID (PK, FK1)
Language code (PK)
Usage
Primary language
Enrollment
Person_ID (PK, FK1)
Enrollment Start Date (PK)
Primary care clinic/provider (FK2)
Enrollment End Date
Insurance type and plan
Drug Coverage
PRO Types
PRO_ID (PK)
PRO subtype
PRO Surveys
PRO_ID (PK)
Survey ID (PK)
Question ID (PK)
Question version (PK)
Question text
PRO Survey Responses
PRO_ID
Survey ID
Question ID
Person_ID
Response variables
Lab Results
Person_ID (FK1)
Ordering Provider (FK2)
Test_ID
Test date
Row_ID
LOINC code
Test results
Lab Notes
Row_ID (PK)
Line_Nbr (PK)
Note type (PK)
Result note
Tumor
Person_ID (PK, FK1)
Sequence (PK)
Diagnosis date
Tumor type variables
Tumor stage variables
Treatment variables
Procedures
Encounter_ID (PK, FK1)
Procedure code (PK)
Procedure code type (PK)
Original procedure (PK)
Procedure date (PK)
Performing provider (PK, FK2)
Provider (FK4)
Person_ID (FK3)
CPTMOD1-3 (PK)
Encounter type
Facility
Department
Demographics
Person_ID (PK)
Gender Identity
Sex Admin
Race
Ethnicity
Birth Date
Pharmacy
Person_ID (PK, FK1)
Dispense date (PK)
Natl Drug code (PK, FK2)
Prescribing MD (PK, FK3)
Amount dispensed
Days supply
Social History
Person_ID (FK1)
Encounter ID (FK2)
Tobacco variables
Alcohol & drug variables
Sexual activity variables
Education
Death
Person_ID (PK, FK1)
Death date
Data source
confidence
Cause of Death
Person_ID (PK, FK1)
Cause of death (PK)
DX codetype (PK)
Cause type (PK)
Data source
confidence
Diagnosis
Encounter_ID (PK, FK1)
Diagnosis code (PK)
Original diagnosis code (PK)
Diagnosis code type (PK)
Diagnosing provider (PK, FK2)
Provider (FK4)
Person_ID (FK3)
Encounter type
Dates of diagnosis
Principal diagnosis
Primary diagnosis
Encounters
Encounter_ID (PK)
Person_ID (FK2)
Provider (FK1)
Encounter type
Facility
Department
Vital Signs
Person_ID (FK1)
Encounter ID (FK2)
Measurement date
Encounter type
Blood pressure, pulse
Height, weight, BMI
Head circumference
Ever NDC
Natl Drug Code (PK)
Generic Name (PK)
Brand name
AHFS codes
GPI codes
Facility
Facility code (PK)
Address information
Provider
Provider (PK)
Specialty
Provider type
Demographic variables