The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration

3-24-2014

The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration

Tyler R. Ross MA
Group Health Research Institute, Group Health Cooperative, Seattle, WA, ross.t@ghc.org

Daniel Ng MBA
Division of Research, Kaiser Permanente Northern California, Oakland, CA

Jeffrey S. Brown PhD
Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

Roy Pardee JD MA
Group Health Research Institute, Group Health Cooperative, Seattle, WA

See next pages for additional authors

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration

Abstract
The HMO Research Network (HMORN) Virtual Data Warehouse (VDW) is a public, non-proprietary, research-focused data model implemented at 17 health care systems across theUnited States. The HMORN has created a governance structure and specified policies concerning the VDW's content, development, implementation, and quality assurance. Data extracted from the VDW have been used by thousands of studies published in peer-reviewed journal articles. Advances in software supporting care delivery and claims processing and the availability of new data sources have greatly expanded the data available for research, but substantially increased the complexity of data management. The VDW data model incorporates software and data advances to ensure that comprehensive, up-to-date data of known quality are available for research. VDW governance works to accommodate new data and system complexities. This article highlights the HMORN VDW data model, its governance principles, data content, and quality assurance procedures. Our goal is to share the VDW data model and its operations to those wishing to implement a distributed interoperable health care data system.

Acknowledgements
Funding was provided by the Academy Health Electronic Data Methods (EDM) Forum, a project supported by the Agency for Healthcare Research and Quality (AHRQ) through the American Recovery & Reinvestment Act of 2009, Grant U13 HS19564-01. Additional funding and support was provided by U19 CA079689 and U24 CA171524 from the National Cancer Institute (NCI). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

Keywords
Informatics, Data Reuse, Health Information Technology, Research Networks, Standardized Data Collection

Disciplines
Databases and Information Systems | Health Information Technology | Public Health

Authors
Tyler R Ross, *Group Health Research Institute, Group Health Cooperative, Seattle, WA*; Daniel Ng, *Division of Research, Kaiser Permanente Northern California, Oakland, CA*; Jeffrey S Brown, *Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA*; Roy Pardee, *Group Health Research Institute, Group Health Cooperative, Seattle, WA*; Mark C Hornbrook, *The Center for Health Research, Kaiser Permanente Northwest, Portland, OR*; Gene Hart, *Group Health Research Institute, Group Health Cooperative, Seattle, WA*; John F Steiner, *Institute for Health Research, Kaiser Permanente Colorado, Denver, CO*.

The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration

Tyler R. Ross, MA;[i,ii] Daniel Ng, MBA;[iii] Jeffrey S. Brown, PhD;[iv,v] Roy Pardee, JD, MA;[i,ii] Mark C. Hornbrook, PhD;[vi] Gene Hart, MS;[i,ii] John F. Steiner, MD, MPH[vii]

Abstract

The HMO Research Network (HMORN) Virtual Data Warehouse (VDW) is a public, non-proprietary, research-focused data model implemented at 17 health care systems across the United States. The HMORN has created a governance structure and specified policies concerning the VDW's content, development, implementation, and quality assurance. Data extracted from the VDW have been used by thousands of studies published in peer-reviewed journal articles. Advances in software supporting care delivery and claims processing and the availability of new data sources have greatly expanded the data available for research, but substantially increased the complexity of data management. The VDW data model incorporates software and data advances to ensure that comprehensive, up-to-date data of known quality are available for research. VDW governance works to accommodate new data and system complexities. This article highlights the HMORN VDW data model, its governance principles, data content, and quality assurance procedures. Our goal is to share the VDW data model and its operations to those wishing to implement a distributed interoperable health care data system.

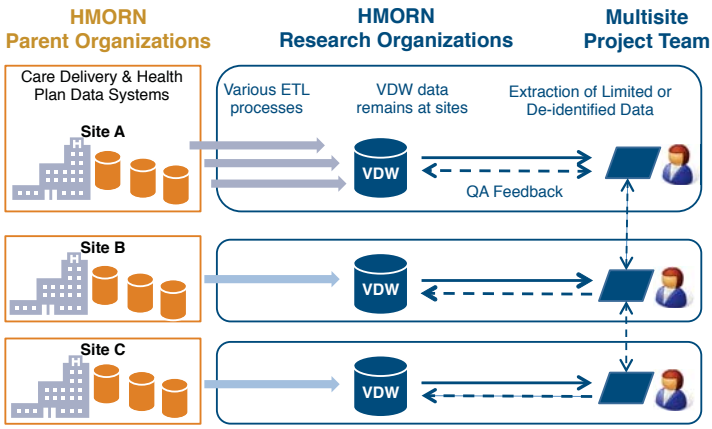
Introduction

The HMO Research Network (HMORN) Virtual Data Warehouse (VDW) removes duplicative work within a research center by maintaining single extract, transform, and load (ETL) processes for creating commonly used variables in single and multisite research studies. This allows research projects to focus data development efforts on data not yet included in the VDW. By documenting these pioneering efforts and following VDW documentation guidelines, investigators contribute to expanding VDW coverage. Other projects and other research centers can build on the work of individual projects to expand the VDW data model. Because VDW data files already exist at each site, data query tools may be used to obtain preparation-to-research data tabulations across multiple sites within days or even hours. Such queries enable investigators to expediently assess the feasibility of research questions and quickly compute statistical power levels. Implementing the VDW data model supports efficient data warehousing and analytic efforts on research projects from feasibility through final analysis.

Population-based health care research requires comprehensive, high quality data on large and diverse cohorts. Collecting these comprehensive data specifically for research purposes can be costly, even for relatively small numbers of individuals. To collect data efficiently on large populations, scientists frequently rely on information routinely collected in the course of delivering health care and generating bills and health insurance claims. These operational data systems and models, however, are not designed for research. For example, a claims adjudication system might organize its data to link insurance contract information with the services billed by health care providers. An electronic health record (EHR) might organize and store data in a way that facilitates the rapid display of all information about a particular patient to meet the needs of a clinician. For these reasons, data structures that support claims adjudication software or an EHR are often not conducive to population-level queries, for example assessing changes in blood pressure across all incident users of a therapeutic drug for a pharmacoepidemiology study. Special procedures and expertise are needed to create research-quality data from operational insurance and clinical data.

[i] Group Health Research Institute, [ii]Group Health Cooperative, [iii]Kaiser Permanente Northern California, [iv]Harvard Pilgrim Health Care Institute, [v]Harvard Medical School, [vi]Kaiser Permanente Center for Health Research-Northwest [vii]Kaiser Permanente Colorado Institute for Health Research

Figure 1. Data Flow in and out of the VDW



Data are generated in the course of care delivery and health plan administration. Each site transforms these data according to the standard VDW data model. Researchers extract data from the VDW and provide feedback to improve data quality.

To repurpose clinical, administrative, and claims data for research use, scientists must design data models specifically to facilitate research and then extract, transform, and load data from source systems to a research data model (Figure 1). The data model design and ETL process can be labor intensive and prone to error due to data system complexity and frequently changing data sources. Frequently, these ETL processes are performed on a project-by-project basis, which exacerbates cross-project inconsistency and duplicates costly programming. As demand for these project-specific research data sets has grown, research institutions have invested in research data warehouses and technical infrastructure that standardize their data extraction processes.1-4 This approach has several advantages. By identifying common data needs across research projects, the data model design and initial ETL processes can be done once and maintained centrally. Having many users of the same data warehouse greatly increases opportunities to develop analytic efficiencies and identify, document, and resolve quality issues, thereby benefiting both current and future projects. A research data warehouse enables a research center to develop its data assets strategically and intentionally. A research data model also enables the development of tools and resources that can be used across projects over time. For example, algorithms that define common research concepts such as "medication adherence," "type 2 diabetic patients," "high-deductible medical insurance," and "comorbidity index" can be created once and shared across projects. In addition, the standardized and stable data model facilitates the development of software tools, and standardized querying approaches can make use of the shared resource.

Unfortunately, even the largest health-plan data warehouses are often insufficient to address many research questions. To address this limitation, multicenter collaborations are formed to aggregate data across institutions to obtain larger and more representative populations.[5] Consortia such as the Centers for Disease Control and Prevention's (CDC) Vaccine Safety Datalink (VSD),[6] the Agency for Healthcare Research and Quality (AHRQ) Centers for Education and Research on Therapeutics (CERT),[7] the National Heart Lung and Blood Institute's (NHLBI) Cardiovascular Research Network,[8] and the National Cancer Institute (NCI) Cancer Research Network (CRN)[9] have designed data models, infrastructure, and governance that allow data aggregation across institutions for research purposes while ensuring that health care organizations maintain local control over their highly regulated and proprietary data.[10-12] Aggregating data across multiple and diverse health care organizations and systems increases the generalizability of research findings. By covering a greater population, aggregating data provides greater ability to study rare exposures and outcomes, conduct subgroup analyses, and improve statistical power.

In this article, we describe the development and governance of one such distributed data model, the HMORN VDW. The HMORN is an international consortium of 18 health care delivery systems with public domain research programs; 17 members use the VDW data model to support multicenter research (http://www.hmoresearchnetwork.org).[13-15] HMORN VDW data are used for studies on a broad range of topics in health services research and epidemiology, in observational studies and clinical trials.[16-21] We describe the principles that guide VDW development and the policies that result from those principles. Finally, we describe how these policies are put into practice through the VDW's many uses.

Approach to the Model

Principles

At inception, the HMORN VDW data model did not establish a set of principles to govern its development. The NCI CRN founded the VDW to achieve its study aims. These principles were formed organically while the CRN worked through challenges in data provenance, data governance, and regulations. While never having been codified in bylaws, the HMORN VDW data model adheres to the following principles:

The first principle of the HMORN VDW is that its primary purpose is to facilitate public domain health and health services research. A consequence of this principle is that the data model does not always conform to common relational database theory such as database normalization or enforced key constraints. When database theory and pragmatism are at odds during data model development or implementation, the practical solution is favored. For example, the VDW at times contains redundant data to expedite the extraction of large volumes of data, despite the risk of data inconsistency within the model.

A second principle is that the VDW, its tools, and governance policies should facilitate compliance with human subjects protection and privacy regulations as well as proprietary institutional interests. This principle compels us to implement the VDW as a federated database in which data reside locally at their respective institutions until the point of extraction for a specific approved purpose. Data extraction for research occurs only after Institutional Review Board (IRB) approval, and the transfer of protected health information among institutions occurs only after data use agreements (DUA) or other contractual agreements are executed. Tools and applications built on the VDW data model must consider these restrictions in their interactions with and use of the data. For instance, the HMORN has developed a tool that identifies

possible protected health information among queried VDW data and prompts data stewards to reconcile the data to be shared with applicable DUAs before transfer.[22]

A third principle of the VDW is that the data model is public and nonproprietary, although the data contained within the model are not. Any entity may implement the HMORN VDW or clone and adapt the model to serve its particular need. The U.S. Food and Drug Administration Mini-Sentinel Network initiative,[23] for example, modified the VDW model to develop the Mini-Sentinel Common Data Model for medical product safety surveillance. Increasingly, health systems that implement the VDW initially for research purposes are making use of the VDW for internal business reporting and decision support rather than modifying the model for their particular use. A consequence of this principle, however, is that the VDW cannot replicate specifications from proprietary data models or publish proprietary medical terminology.

A fourth principle is that the VDW data model should be flexible and extensible to accommodate a range of participating institutions, research interests, and data sources. As new institutions with different data environments implement the VDW, the governing body modifies specifications to integrate the new data. As an example, the addition of health care organizations not integrated with health plans meant broadening study definitions that determine populations at risk for an outcome or exposure. Previously, "at risk" was defined strictly as a person enrolled in a health care plan. Once the HMORN added new organizations, we changed this conception to include people based on their place of residence or by patterns of health care utilization. To accommodate the broadest range of research interests possible, the VDW data model covers a variety of content areas. Its design also allows researchers to supplement VDW data with narrower content-specific data sources. For instance, studies by the National Institute of Mental Health's Mental Health Research Network (MHRN)[24] use data on diagnoses, procedures, and pharmaceutical dispensings from the VDW; however, data on patient responses to the PHQ9 depression screening instrument[25] are not currently in the VDW. The VDW design allows MHRN investigators to collect these data separately and easily merge them with VDW data. Supplemented data likely to be used repeatedly can be incorporated as new data tables into the VDW model.

A fifth principle of the VDW is that the data model should be agnostic to source data systems. In practice, this means that VDW specifications are defined by data concepts rather than data sources. For example, the VDW defines race and ethnicity categories in accordance with National Institute of Health policy[26] rather than any data source system that collects race. Further, while replicating a data model specification that imitates a data source shared across multiple institutions might be expedient, this impairs other institutions from contributing data. Imitating source data models also potentially violates the third principle of the VDW as a public data model if the source data model is proprietary. For these reasons, the VDW data model is agnostic to source.

In summary, these five principles have led to a data sharing model that is tailored to the needs of researchers, protects the privacy and confidentiality of member data, can encompass new content areas, and facilitates aggregation of data across disparate health care delivery systems.

Policies and Governance

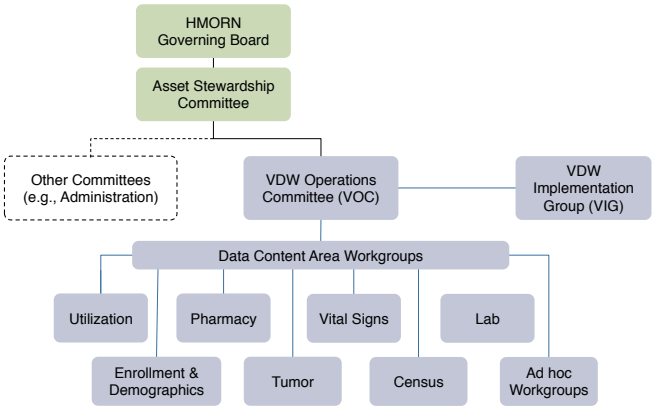
The HMORN VDW Operations Committee (VOC) (Figure 2) is the governing body responsible for VDW development, implementation, operations, and oversight. Workgroups under the VOC, each co-led by a scientist and a data programmer, conduct quality assurance, develop specification changes, and support implementation for specific content areas. VOC membership consists of workgroup leads and representatives from multisite consortia such as the MHRN, VSD, CVRN, Mini-Sentinel, and CRN. The VDW Implementation Group (VIG) consists of the workgroup leads as well as site data managers who are chiefly responsible for implementing the VDW at their respective institutions. The VOC, VIG, and workgroups each have monthly teleconference meetings. The VIG additionally has two in-person working meetings per year. VDW oversight is provided by the Asset Stewardship Committee (ASC) whose members include research center directors, investigators, and senior staff who develop and maintain tools useful for multiorganizational research in areas such as data development, procedures for human subjects review, and administrative efficiencies. The ASC in turn is overseen by the HMORN Governing Board, which consists of the directors of the participating research institutions and provides leadership and strategic direction to the HMORN.

Formalized processes exist for changing VDW specifications, implementing new VDW tables, and formulating new workgroups. These activities can be costly for VDW implementers and tool builders, so specification changes and table additions are first proposed by workgroups, then discussed by the VIG. Formal approval is by vote, frequently at in-person meetings. Each change to a VDW table results in an incremented table version number. VDW specifications are published on a public website,[27] while meeting minutes, implementation documentation, and HMORN policies are hosted on a private website. Data quality issues, identified either by VDW users or through dedicated quality assurance programs, are logged in an issue tracker viewable by HMORN members and collaborators and reviewed monthly at VIG meetings.

Quality Assurance

VDW data quality is assessed and improved through two mechanisms: dedicated quality assurance programming and crowdsourcing via the VDW user base. Workgroups are responsible for authoring quality assurance programs that assess adherence to the VDW data model and identify data anomalies. These quality checks range from simply verifying the existence of variables and assuring they contain only permissible values to more sophisticated analyses requiring clinical or scientific knowledge such as comparing rates and trends of events across institutions. Identified

Figure 2. HMORN Governance



The VDW Operations Committee (VOC) is the governing body responsible for the VDW. Workgroups under the VOC are responsible for specific data content areas. The Asset Stewardship Committee and HMORN Governing Board provide oversight over HMORN resources including the VDW. The VDW Implementation Group includes site data managers who implement the VDW at their respective site.

data anomalies can be the result of errors in ETL code, limitations in data availability from source systems at an institution, or true anomalies in the data.

Studies and research consortia that use the VDW are encouraged to begin their data analysis with basic descriptive assessments of variables key to their research. The prolific use of the VDW ensures that this crowdsourced quality assurance approach identifies most discovered data anomalies. This approach also prioritizes quality assessment based on the most commonly used variables. Quality assurance by users who are experts in their field means that evaluation is by the people who are most likely to identify data issues. Kaiser Permanente's Center for Effectiveness and Safety Research (CESR)[28] is an example of a consortium that has conducted extensive quality analysis on VDW data across participating sites and shared these programs with the VOC workgroups. Data anomalies, whether identified by VDW workgroups or users, are recorded in an issue tracking system on a private website. Site data managers investigate anomalies and report resolutions in the issue tracker. This critical knowledge management process helps prevent scientific errors and the inefficient use of study resources to reidentify data anomalies already discovered by prior researchers.

Application of the Model

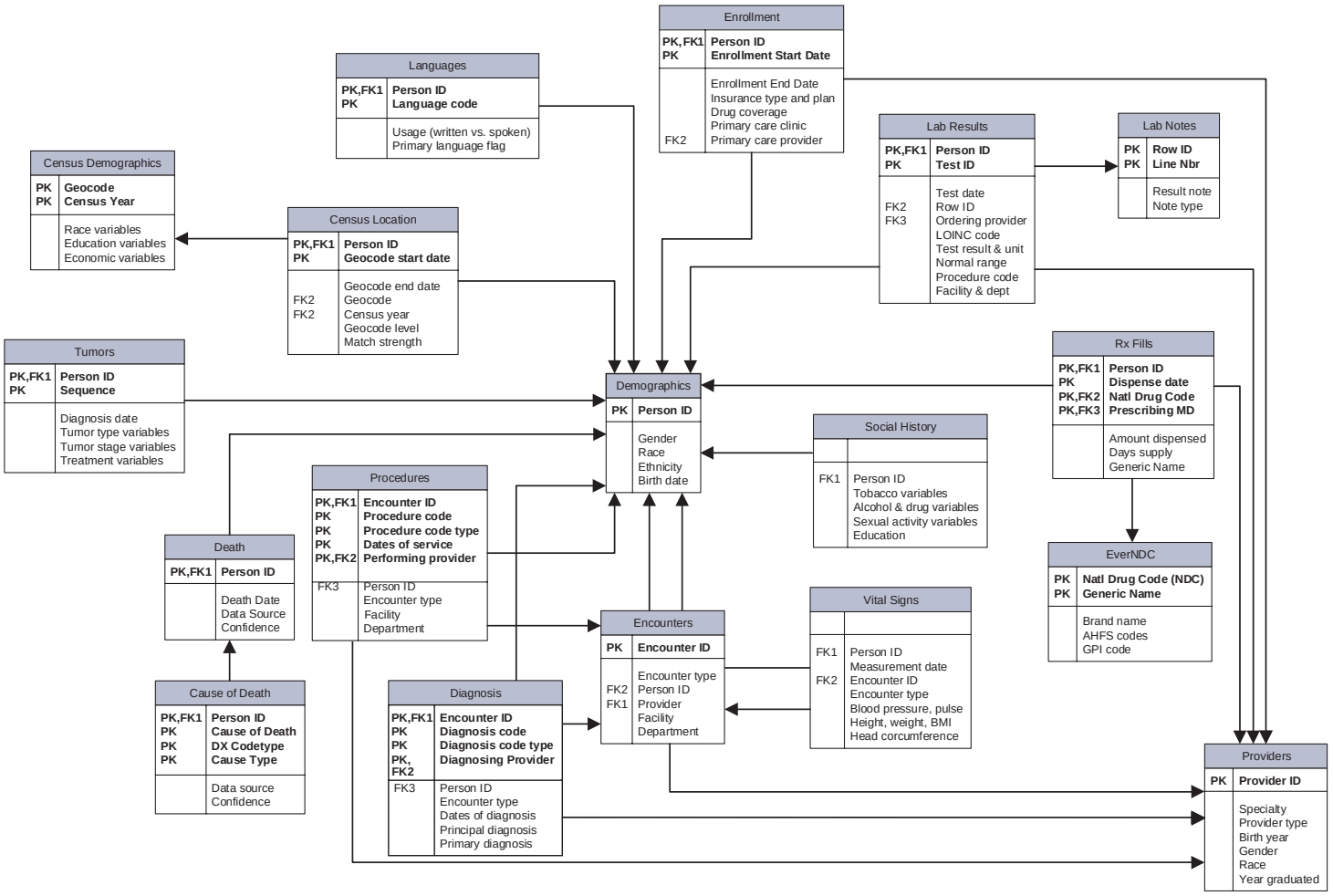
Content of the VDW Data Model

The VDW data model (Figure 3) has seven content areas: enrollment/demographics, utilization, laboratory, pharmacy, census, tumor registry, and vital signs/social history. These content areas comprise more than 450 variables among 18 tables. Variable availability is dependent on collection and preservation of data in local source data systems. Site data managers document the availability and missingness of data on a shared secure web portal as a part of VDW implementation. A rich set of tools, applications, and data models have successfully interfaced with the VDW. PopMedNet, an open-source distributed querying software application (http://www.popmednet.org), includes several query tools that operate against the VDW and is being used by several HMORN sites for general feasibility queries and by several HMORN-based projects to enable distributed querying. Installations of Informatics for Integrating Biology and the Bedside (i2b2, https://www.i2b2.org) software, a data warehouse and associated querying tool, have been established at select sites with data drawn from the VDW to facilitate proposed research project feasibility. The Mini-Sentinel Common Data Model is derived from the VDW data model through maintained ETL code that translates the VDW to the Mini-Sentinel data model. The VSD data model, with the exceptions of its mother-baby linkage and dedicated vaccine tables, may also be derived from the VDW. The VOC maintains a library of publicly available SAS macros (https://github.com/HMORN/vdw-macros) called "Standard Macros," for users to perform common research tasks such as identifying individuals continuously at risk for outcomes or exposures, calculating body mass index percentiles for children, and calculating comorbidity scores. The VOC also maintains a dictionary of references, called "Standard Vars," which simplifies data extraction by abstracting away site-specific environmental factors like directory names and database connection strings.

Context Among Other Data Models

The HMORN VDW is just one standardized data model currently used in clinical research.[6,23,29,30] Other standardized models include the Vaccine Safety Datalink, Observational Medical Outcomes Partnership (OMOP), and the Mini-Sentinel Common Data Model. The i2b2 is a commonly used data repository, but it does not enforce standard value sets; each implementation of i2b2 can by definition be unique, whereas each implementation of the other data models mentioned are intended to be comparable. These standardized data models were not developed independently, but rather were built on shared and learned processes, data concepts, and best practices. Differentiation among models exists principally because of their different users, the needs of the funders/organizer, and characteristics of the expected data contributors. The HMORN VDW's first principle is that it exists to facilitate public research by a large cohort of investigators across a wide range of clinical areas and research topics, ranging from comparative effectiveness and safety research to epidemiology to burden of illness. The VDW guiding principles also placed a high priority on analytic simplicity and data provenance (see principles section above). The Vaccine Safety Datalink data model is based on similar source data as the VDW but, in contrast, serves to support prospective near-real-time vaccine safety monitoring and therefore has developed a unique approach to organizing and using the data to support that mission. Mini-Sentinel's Common Data Model is a simplified derivative of the HMORN VDW designed specifically to support medical product safety surveillance for the U.S. Food and Drug Administration. The OMOP data model borrowed lessons from the VDW and developed its own model to support high-throughput methodologic research on comparative safety and effectiveness, with a specific need to accommodate medical data from a diverse range of data sources to enable comparisons across those disparate data types.

Figure 3 – The HMORN Virtual Data Warehouse Data Model



An entity-relationship (ER) model depicting the HMO Research Network Virtual Data Warehouse as it existed in 2014.

Moody and Shank[31] created a framework to evaluate data models that can be used to compare data model quality dimensions. While all dimensions are important in this framework and comparisons are difficult to make out of specific use case contexts, the HMORN VDW tends to have more complete data but less simplicity, so as to accommodate the broader set of expected uses. Flexibility, or what we have termed "extensibility," has greater emphasis than does correctness. Integrity, implementability, and integration are more critical than understandability. These emphases guide the VDW's development in contrasting ways to other clinical data models.

Implementation and Use

The HMORN VDW currently consists of 17 sites that together cover 13 million individuals; in total, the VDW has over 185 million person-years of data. Additional non-HMORN affiliated institutions in the Denver metropolitan area are currently implementing their own VDW nodes to promote federated research with HMORN sites. The VDW has been successfully implemented on a variety of platforms, including Teradata®, SQL Server®, Oracle®, and SAS® data sets in Windows® and UNIX®, while allowing interoperability and distributed querying. In practice, VDW data extraction is performed using SAS, so accessibility via SAS is an important consideration in site implementation configuration.

Appendix A provides a partial list of the hundreds of peer-reviewed journal articles that have used data extracted from the VDW organized by network consortia. Examples from recent studies using the VDW include the CRN's finding that the increased use of computed tomography in the pediatric population has resulted in children receiving high doses of radiation known to cause cancer.[32] The MHRN, based on health care utilization and mortality data from the VDW, concluded that there is great opportunity for suicide prevention through targeted improvements in primary care.[33] Also using the VDW, the Cardiovascular Research Network identified increased risk of death and hospitalization associated with chronic kidney disease among adults with heart failure.[34] Several HMORN sites use the VDW as a major source for nearly all quantitative research conducted, even if not part of a multisite network consortium like those listed above. The VDW's widespread use in assessing study feasibility, conducting multisite research, and conducting single site research across a diverse set of subject matters and disciplines is among its greatest strengths.

Discussion

Over the past decade, the HMORN VDW has grown in breadth, depth, quality, and use. The VDW has been used to support the work of hundreds of research investigators and hundreds of publications. The VDW approach to multicenter research has been successfully adopted by several large-scale multisite networks. At the same time, VDW data sources such as EHRs and health plan data systems have grown and matured, expanding data available for research within the VDW.

While public domain research is valued by HMORN member health systems, the designs of internal IT systems are driven by health care operational priorities, not research. For example, despite their significance in care delivery and research, data on patient race and ethnicity were rarely collected in structured EHR data until required by meaningful use incentives put in place by the Centers for Medicare & Medicaid Services. Adherence to the VDW principle of flexibility and extensibility has been paramount to the continuation of the VDW data model in a rapidly changing health information technology environment. The changing health care environment means that the VDW maintenance, quality improvement, and enhancement require ongoing support and vigilance. The VDW governance and support structure that includes the VOC, VIG, workgroups, and Asset Stewardship Committee play a crucial role in the continued value and use of the VDW.

The VDW is not without its limitations. Foremost is that data in the model are limited to information acquired and preserved by sites in the course of business and care delivery. This limitation exists for any research data model sourced by claims and EHR data. Researchers who use this type of data sometimes incorrectly assume that the presence of variables in the data model implies that those variables are populated at every site over all time; users are advised to ensure the availability of the data they need in a study feasibility assessment. The issue of data availability is increasingly important as nonintegrated health care systems adopt the VDW model, and traditional integrated health plan and care delivery models are supplemented with contracts to clinicians who provide care to health system members but do not use the system's EHR. A business decision by a health care system to contract out specialty care services, for example, may limit or even terminate the availability of specialty care data at a site. Health information exchanges, which increase health care provider access to patient electronic medical information, have the potential to mitigate data availability problems, but only to the extent that exchanged information is available to researchers. The limitation of data availability by site stresses the importance of site documentation and project-specific quality assurance analysis, both of which are strongly encouraged by the VOC.

A second limitation is that implementing and maintaining a multicenter data resource requires substantial resources. These costs are not unique to the VDW; all multicenter research projects will face costs associated with data standardization and curation. In the case of the VDW, we have found that this investment pays dividends with each use, but requires a certain level of use to reach a break-even point. Costs of maintenance includes costs associated with ongoing data quality review, source system alterations, VDW specification changes and identification of variation in implementation across sites. Ongoing maintenance and support of a multicenter research warehouse like the VDW requires a stable funding mechanism that should be supported by the participating institutions and the individual users of the resource. Monitoring the use of a data resource like the VDW within a distributed network is challenging and is an important area for improvement; better tracking of VDW usage would improve future usability and enable identification of best practices.

A third limitation is that some differences in data across sites cannot be resolved through a standardized data model. While VDW data reflect substantial efforts to abstract away data and system differences across sites, their use without knowledge of local source systems and health care policies and practices can lead to dubious conclusions. Using and interpreting VDW data correctly requires extensive knowledge of the health care and IT systems from which VDW data are sourced. Including local staff from each data-contributing site on the research team greatly facilitates data quality checks and interpretation of site-specific data patterns. For example, durable medical equipment can be dispensed through the pharmacy or delivered through an outside contract vendor. Chemotherapy infusion treatment may be documented through pharmacy dispensings or through procedural billing codes.[35] These different methods of delivering care and paying for it will be reflected differently in VDW data at each site. Site representation when designing a study, extracting data, and analyzing results prevents erroneous conclusions.

The VDW continues to adapt to a changing health care and health IT environment. The VDW's historic dependence on SAS is waning as sites increasingly use relational databases to store VDW data and developers build non-SAS-dependent query tools that access VDW data. New data subject areas under consideration include patient-reported outcomes, dental care, infusion therapy, genetic sequencing, clinical text, physician ordering and linkages to birth certificate data. Which subject areas are developed will be driven by the needs of research projects making use of the VDW in the coming years.

Next Steps for the Community

The VDW data model can be emulated by health care and health insurance systems outside the HMORN who are seeking to implement a distributed interoperable health care data system or develop collaborations with the HMORN. We welcome requests for technical assistance from external health care delivery systems and health researchers interested in becoming part of this large data cooperative. Moreover, the VDW governance welcomes suggestions for improvements, expansions, and collaborations.

The HMORN VDW is an evolving and extensible data model because of its ability to include new content and develop new definitions of existing data fields. The VDW is sustainable because it increases efficiency and usability of health care informatics

resources by eliminating duplicative data programming and enabling software and tools that make use of data.

References

1. Gainer V, Hackett K, Mendis M, Kuttan R, Pan W, Phillips LC, et al. Using the i2b2 hive for clinical discovery: an example. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2007:959.
2. Klompas M, Lazarus R, Daniel J, Haney G, Hou X, Campion FX, et al. Electronic medical record Support for Public health (ESP): Automated Detection and Reporting of Statutory Notifiable Diseases to Public Health Authorities. Advances in Disease Surveillance. 2007;3(3):1-5.
3. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2007:548-52.
4. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012;50 Suppl:S60-7. Epub 2012/06/22.
5. Federal Coordinating Council for Comparative Effectiveness Research. Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and Congress. In: Services UDoHaH, editor. 2009.
6. DeStefano F, Vaccine Safety Datalink Research G. The Vaccine Safety Datalink project. Pharmacoepidemiology and drug safety. 2001;10(5):403-6.
7. Platt R, Davis R, Finkelstein J, Go AS, Gurwitz JH, Roblin D, et al. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. Pharmacoepidemiology and drug safety. 2001;10(5):373-7.
8. 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.
9. Wagner EH, Greene SM, Hart G, Field TS, Fletcher S, Geiger AM, et al. Building a research consortium of large health systems: the Cancer Research Network. Journal of the National Cancer Institute Monographs. 2005(35):3-11.
10. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, et al. Building a virtual cancer research organization. Journal of the National Cancer Institute Monographs. 2005(35):12-25.
11. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010;48(6 Suppl):S45-51. Epub 2010/05/18.
12. Maro JC, Platt R, Holmes JH, Strom BL, Hennessy S, Lazarus R, et al. Design of a national distributed health data network. Ann Intern Med. 2009;151(5):341-4. Epub 2009/07/30.
13. Arterburn DE, Alexander GL, Calvi J, Coleman LA, Gillman MW, Novotny R, et al. Body mass index measurement and obesity prevalence in ten U.S. health plans. Clinical medicine & research. 2010;8(3-4):126-30.
14. Newton KM, Larson EB. Learning health care systems: leading through research: the 18th Annual HMO Research Network Conference, April 29-May 2, 2012, Seattle, Washington. Clinical medicine & research. 2012;10(3):140-2.
15. Selby JV. Linking automated databases for research in managed care settings. Ann Intern Med. 1997;127(8 Pt 2):719-24.
16. Katon WJ, Lin EH, Von Korff M, Ciechanowski P, Ludman EJ, Young B, et al. Collaborative care for patients with depression and chronic illnesses. The New England journal of medicine. 2010;363(27):2611-20.
17. Matlock DD, Groeneveld PW, Sidney S, Shetterly S, Goodrich G, Glenn K, et al. Geographic variation in cardiovascular procedure use among Medicare fee-for-service vs Medicare Advantage beneficiaries. JAMA : the journal of the American Medical Association. 2013;310(2):155-62.
18. Habel LA, Cooper WO, Sox CM, Chan KA, Fireman BH, Arbogast PG, et al. ADHD medications and risk of serious cardiovascular events in young and middle-aged adults. JAMA : the journal of the American Medical Association. 2011;306(24):2673-83.
19. Green BB, Cook AJ, Ralston JD, Fishman PA, Catz SL, Carlson J, et al. Effectiveness of home blood pressure monitoring, Web communication, and pharmacist care on hypertension control: a randomized controlled trial. JAMA : the journal of the American Medical Association. 2008;299(24):2857-67.
20. Phelan EA, Borson S, Grothaus L, Balch S, Larson EB. Association of incident dementia with hospitalizations. JAMA : the journal of the American Medical Association. 2012;307(2):165-72.
21. Kahn KL, Adams JL, Weeks JC, Chrischilles EA, Schrag D, Ayanian JZ, et al. Adjuvant chemotherapy use and adverse events among older patients with stage III colon cancer. JAMA : the journal of the American Medical Association. 2010;303(11):1037-45.
22. Bredfeldt CE, Butani A, Padmanabhan S, Hitz P, Pardee R. Managing protected health information in distributed research network environments: automated review to facilitate collaboration. BMC medical informatics and decision making. 2013;13:39.
23. 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.
24. Simon G, Rutter C, Peterson D, Oliver M, Whiteside U, Operskalski B, et al. Do PHQ Depression Questionnaires Completed During Outpatient Visits Predict Subsequent Suicide Attempt or Suicide Death? . Psychiatric Services. 2013 forthcoming.
25. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. Journal of general internal

medicine. 2001;16(9):606-13.
26. U.S. National Institutes of Health. Amendment: NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research - October 2001. Release Date: October 9, 2001; Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-02-001.html.
27. HMO Research Network. HMORN VDW Detailed Data Structures-1. HMO Research Network; no date [updated September 23, 2011; cited 2014 March 10]; Available from: http://www.hmoresearchnetwork.org/resources/toolkit/HMORN_VDWDetailedDataStructures.pdf.
28. Bachman D, La Chance P-A, Hornbrook M. PS1-28: Kaiser Permanente Center for Effectiveness and Safety Research. Clinical medicine & research. 2010;December; 8(3-4):207.
29. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60. Epub 2011/11/01.
30. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care. 2012;50 Suppl:S49-59. Epub 2012/08/09.
31. Moody DL, Shanks GG. Improving the quality of data models: empirical validation of a quality management framework. Inf Syst. 2003;28(6):619-50.
32. Miglioretti DL, Johnson E, Williams A, Greenlee RT, Weinmann S, Solberg LI, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. JAMA Pediatr. 2013;167(8):700-7. Epub 2013/06/12.
33. Ahmedani BK, Simon GE, Stewart C, Beck A, Waitzfelder BE, Rossom R, et al. Health Care Contacts in the Year Before Suicide Death. Journal of general internal medicine. 2014. Epub 2014/02/26.
34. Smith DH, Thorp ML, Gurwitz JH, McManus DD, Goldberg RJ, Allen LA, et al. Chronic kidney disease and outcomes in heart failure with preserved versus reduced ejection fraction: the Cardiovascular Research Network PRESERVE Study. Circ Cardiovasc Qual Outcomes. 2013;6(3):333-42. Epub 2013/05/21.
35. Delate T, Bowles EJ, Pardee R, Wellman RD, Habel LA, Yood MU, et al. Validity of eight integrated healthcare delivery organizations' administrative clinical data to capture breast cancer chemotherapy exposure. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012;21(4):673-80.

Appendix A

A list of online databases containing publications from consortia that have used the HMORN Virtual Data Warehouse as a major source of their quantitative data.

http://crn.cancer.gov/publications

http://cvrn.org/projects/publications/index.aspx

http://www.supreme-dm.org/Publications.html

http://span-network.org/Publications.html