Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature

John H Holmes,[1] Thomas E Elliott,[2] Jeffrey S Brown,[3] Marsha A Raebel,[4] Arthur Davidson,[5] Andrew F Nelson,[6] Annie Chung,[1] Pierre La Chance,[7] John F Steiner[8]

ABSTRACT

Objective To review the published, peer-reviewed literature on clinical research data warehouse governance in distributed research networks (DRNs).

Materials and methods Medline, PubMed, EMBASE, CINAHL, and INSPEC were searched for relevant documents published through July 31, 2013 using a systematic approach. Only documents relating to DRNs in the USA were included. Documents were analyzed using a classification framework consisting of 10 facets to identify themes.

Results 6641 documents were retrieved. After screening for duplicates and relevance, 38 were included in the final review. A peer-reviewed literature on data warehouse governance is emerging, but is still sparse. Peer-reviewed publications on UK research network governance were more prevalent, although not reviewed for this analysis. All 10 classification facets were used, with some documents falling into two or more classifications. No document addressed costs associated with governance.

Discussion Even though DRNs are emerging as vehicles for research and public health surveillance, understanding of DRN data governance policies and procedures is limited. This is expected to change as more DRN projects disseminate their governance approaches as publicly available toolkits and peer-reviewed publications.

Conclusions While peer-reviewed, US-based DRN data warehouse governance publications have increased, DRN developers and administrators are encouraged to publish information about these programs.

BACKGROUND AND SIGNIFICANCE

An enterprise data warehouse presents opportunities to conduct previously impractical studies of rare exposures or outcomes where very large sample sizes are needed, such as population-based surveillance, treatment safety, or comparative effectiveness research.[1] However, even a large healthcare organization may have insufficient subjects to support such studies. Increasingly, researchers are turning to *distributed research networks* (DRNs), which provide access to health-related data from multiple organizations. These data include, but are not limited to, clinical, laboratory, pharmacy, and procedure data and may be collected in outpatient and inpatient settings. In a DRN, the input is a user-generated *query* which may be posed as a natural-language request, a structured request

thorough a web-based form, or program code. The output could be aggregated counts, statistical graphics, or de-identified individual-level data. This approach helps protect patient privacy and confidentiality, and addresses the proprietary concerns of the enterprise itself.

DRNs typically include a virtual repository or warehouse[2] [3] and a distributed communication model. Data from multiple sources reside on local servers and authorized users obtain access using agreed-upon principles through a single, secure portal and query system as though concentrated in a single, unified resource.[1] [4–9] Figure 1 illustrates a generic DRN.

The HMO Research Network's (HMORN's) virtual data warehouse (VDW) is an example of such a resource. We use *VDW* as a generic term here to represent the virtual data repository used by DRNs. In a VDW, data are standardized based on a common data model that enforces uniform data element naming conventions, definitions, and data storage formats.[1] [10–13] Both single-use[14–16] and multi-use[6] [12] [15] [17] networks have been created.

The DRN model imposes many governance challenges.[18] *Data governance* has been defined as 'the high level, corporate, or enterprise policies or strategies that define the purpose for collecting data, and intended use of data'[13] or more specifically, 'the process by which responsibilities of stewardship are conceptualized and carried out,' where such stewardship may include methods for acquiring, storing, aggregating, de-identifying, and releasing data for use.[10] Data governance within DRNs must address regulations and policies established at institution, network, and/or federal levels. Recognizing the need for DRN standards and governance to protect information originating in routine patient care, the federal Query Health Initiative[19] seeks to develop and implement standards for 'distributed population health queries to certified electronic health records.'[20]

OBJECTIVE

We conducted a systematic review of the indexed, peer-reviewed literature on DRN data governance. We were interested in the following questions: How are DRN data made available to researchers? What data standards are used in the DRN? Who can query such data? Who can access query results? What specific policies govern the use, security, and retention of these data and query results? How is
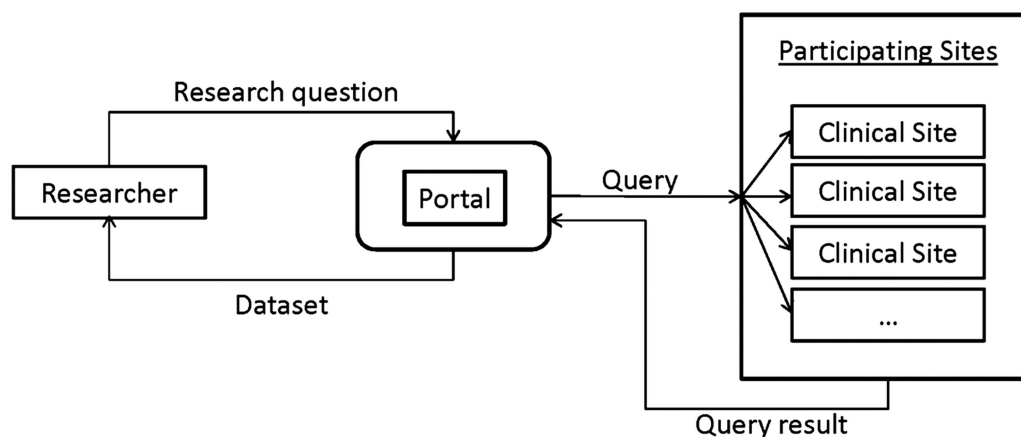
Figure 1 A simple distributed research network. In this schematic, a researcher poses a research question to a portal, typically implemented as a web site with a structured interface that guides the construction of a query. The query is then sent to the participating sites in a predetermined format and language, such as SAS code. The sites run the query and return the result to the portal for use by the researcher, formatted as an aggregated table or de-identified record-level data, depending on the governance policies of the distributed research network.

data governance evaluated? Finally, what procedures have been defined for training users of DRN resources?

MATERIALS AND METHODS
Search strategy
We searched PubMed, PubMed Central, EMBASE, CINAHL, and INSPEC for documents published through July 31, 2013. We included original English-language research articles, reviews, and indexed conference papers and abstracts that described DRN data governance. We excluded documents describing networks outside of the USA due to regulatory differences. With the exception of technical reports, gray (unpublished) literature, was excluded, as were editorials. We used the search terms shown in box 1, expanded as indicated by the truncation ('$') character.

A document was defined as relevant if it contained information about multi-institutional research data, research networks, and governance. Primary documents were examined for additional relevant documents that were also reviewed and added into the analytic corpus.

Analytic strategy
We based our analysis on a faceted classification framework, derived first deductively using the '10 Universal Components of a Data Governance Program' (DGI Data Governance Framework, http://www.datagovernance.com/dgi_framework.pdf) as a high-level taxonomy (see online supplementary appendix 1). Other governance frameworks were not determined to be suitable for our analysis. We then enriched this taxonomy

with concepts that emerged in our corpus. These concepts, or facets, are shown in box 2, with reference to the Data Governance Institute (DGI) framework component(s).

Using the coding tree, two coders (AC and JHH) classified each document. Since we used a faceted classification approach, documents were not restricted to only one category. The two coders compared their classifications and resolved any discrepancies by consensus.

This study was approved by the Kaiser Permanente Colorado (KPCO) Institutional Review Board.

RESULTS
Our search retrieved 6641 documents. After screening for duplicates and relevance, 39 were included in the final review. Figure 2 details the document retrieval process.

Table 1 provides citations for the 39 documents in the final corpus, ordered by first author, with the facets they cover.

Facet 1: Data collation
Data collation refers to an organization's policies and procedures pertaining to assembling data specifically for research purposes.

Data sources include electronic medical records,[16] [21–24] pharmacy and laboratory databases,[23] administrative billing claims,[1] [3–6] [8] [9] and health plan enrollment data.[12] [16] [21] [23] [25]

Box 1 Search terms
'data govern$' 'distributed research network$' 'distributed research' 'distributed network$' 'research network$' 'multi-institutional research' 'data' AND 'govern$' 'data' AND 'research network$' AND 'govern$'

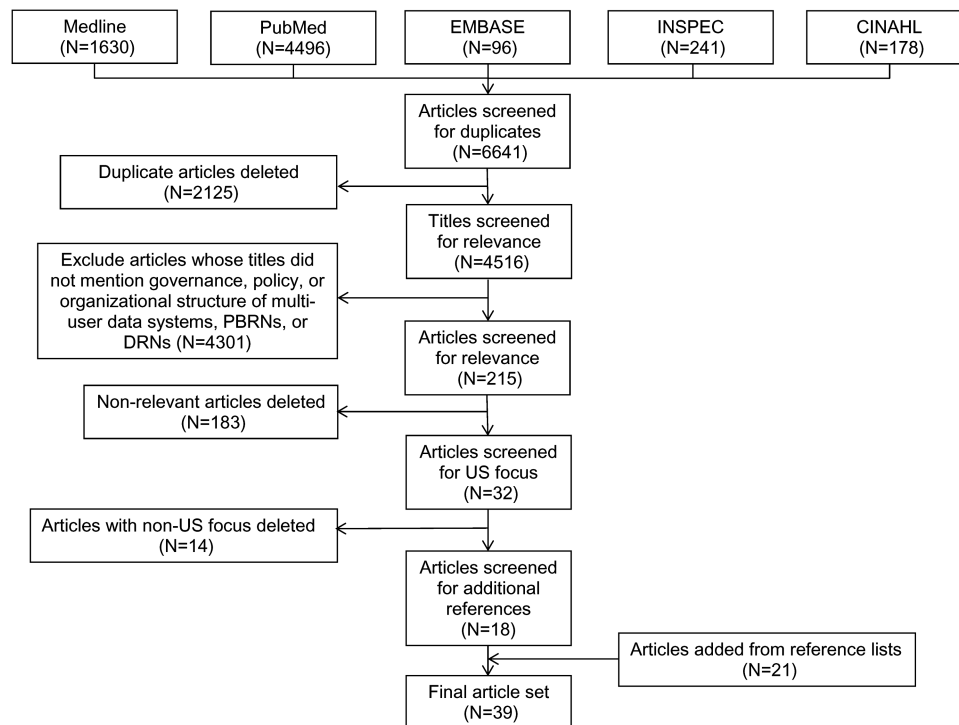| Box 2 Coding tree used to classify documents |
|---|
| Numbers in parentheses refer to the source component of the DGI Framework
Data collation (3)
Data and process standards (1, 3, 5)
Data stewardship (1, 4, 5, 6, 8, 9, 10)
Data privacy (3,6)
Query alignment and approval (3, 5, 9)
Data use (1, 4, 7, 9)
Data security (3, 6, 10)
Data retention (1, 2, 3, 5, 6, 7, 9)
Data audits (2, 3, 4, 10)
User training (7) |

Figure 2 Flow of document retrieval.

The wide variety of sources poses challenges for data collation. Data represent different concept domains (such as drugs, vital signs, or claims), and are syntactically and semantically heterogeneous. For example, body temperature might be represented at one site using Fahrenheit and at another using Celsius. Standards are required for successful data collation. Policies and procedures for addressing these standards were discussed in documents considered in facet 2.

Facet 2: Data and process standards
A *data standard* promotes syntactical and semantic consistency by enforcing a pre-determined set of data representation requirements for each DRN site. A *process standard* refers to the format, language, and content of queries, data models, and processes that affect DRN operation. Both standards are important for interoperability, data capture and accuracy, and analysis. Several articles described these attributes.[2 4 8 13 24] How these standards are created and enforced varies, however. In some cases, a coordinating center develops data standards that all participating sites uphold, while in others data standards are adapted to a common data model that applies to all sites.[26] Some DRNs enforce consistency by providing standards for queries that generate results in a common format and that meet system and resource requirements.[1 3 8 27] The HMORN established a VDW Operational Committee that has a working group which is responsible for overseeing data and process standards.[28]

A paper on the Mini-Sentinel Common Data Model (MSCDM) mentioned that partners were surveyed to determine what data formats should be included.[4] The Cancer Research Network created a single data dictionary that 'guides the assemblage of standardized site-specific databases in each organization.'[3] In the case of the Cardiovascular Research Network,[21] all data are structured into a standardized format in a VDW. This is comprised of: (1) datasets stored behind separate security firewalls at each site including identical variable definitions, labels, coding, and definitions; (2) informatics tools that facilitate data storage, retrieval, processing, and management; and (3) regularly updated documentation of all data elements.

Facet 3: Data stewardship
Data stewardship refers to the way results are curated at local and requesting sites. It involves oversight from legal, auditing, and compliance departments, executive leadership, and institutional review boards. Bloomrosen considered stewardship as central to data governance.[29] In a DRN, where results are transferred outside local institutions, it is often difficult to determine who owns these results. Decisions about data ownership and stewardship affect data accessibility by those outside the contributing organization, even if they are DRN members. The Wisconsin Network for Health Research (WiNHR) established a central authority to govern ownership and stewardship concerns.[26] All institutions in this network are represented on this committee, and have equal participation and authority in promulgating policies and procedures for stewardship.

One key benefit of a DRN is that participating sites retain local control of their data. Most DRNs considered here store their data behind local firewalls and have site-specific data protection, access, and privacy policies.[1 4 9 11 21–23 25] As mentioned by Forrow[5] and by Curtis,[4] the Mini-Sentinel Network complies with the standards imposed by the US Federal Information Security Management Act of 2002 and the HIPAA Security Rule. To this end, Lazarus[22] notes that local information services staff need to check that there are no 'backdoors' that could compromise system security.

Permission to query data in a DRN is governed by the purpose of access and use and by authentication and authorization policies contained in data use agreements. McMurry developed a Distributed Access Control Framework for this purpose.[7] This system records an audit trail of the identities of

Table 1 List of documents contained in the final corpus for review, with their contributions, classified according to the faceted framework used in this review

	First author	Reference number	Data collation	Data and process standards	Data stewardship	Data privacy	Query alignment and approval	Data use	Data security	Data retention	Data audits	User training
1	Baggs	25	X	X	X		X	X	X			
2	Bailey	26	X	X	X		X					
3	Bloomrosen	29			X							
4	Braff	31				X						
5	Bredfeldt	32				X						
6	Brown	11	X									
7	Brown	28	X		X		X	X		X	X	
8	Curtis	4	X	X	X		X	X	X			
9	Fernandes	13	X	X			X	X	X			
10	Forrow	5	X	X	X	X	X	X	X	X		
11	Fullerton	15				X						
12	Gardner	30				X						
13	Go	21	X	X	X		X	X				
14	Godwin	38										X
15	Greene	44										X
16	Holve	42	X	X	X	X						
17	Kim	36				X						
18	Lazarus	22	X		X	X	X	X	X			
19	Lopez	41	X	X	X	X		X	X			
20	Magid	23	X	X	X		X	X				
21	Manion	17				X			X		X	
22	Maro	12	X			X	X		X		X	
23	McGarvey	14	X				X				X	
24	McGraw	6	X	X			X	X	X	X		
25	McMurry	7					X					
26	McMurry	33	X		X	X	X	X		X	X	
27	Ohno-Machado	35						X				
28	Parwani	24	X			X	X		X		X	
29	Patel	2	X			X	X		X		X	
30	Platt	8	X					X				
31	Randhawa	18	X	X								
32	Rosenbaum	10		X	X							
33	Shapiro	16	X		X		X				X	
34	Thompson	39										X
35	Toh	1	X		X		X	X	X			
36	Tucci	27	X	X			X					
37	Velentgas	9	X		X				X			
38	Wagner	3	X				X					
39	Willison	37									X	

the investigator and agency and the time of query. This allows data partners to challenge queries and/or deny access. Shapiro created a real-time system to certify prospective data partners' credentials.[16] Mini-Sentinel policies note that sites may use their own data for any purpose they deem appropriate, but written approval from each participating partner is required for any use of network data for other purposes.[4]

Facet 4: Data privacy

The tension between protecting both patient and organizational privacy and confidentiality, and the need to use clinical and administrative data for research is exacerbated by the HIPAA Privacy Rule. Several DRNs have data access review committees that review proposed secondary uses of data for research.[7 15 17] One group, albeit outside the context of a DRN, has developed a statistical method for releasing secondary data without

compromising patient privacy.[30] The HMORN has adopted a streamlined procedure for institutional review board (IRB) review across the network,[31] as well as a SAS macro that identifies protected health information before data are released to requesters.[32]

Most DRNs require that transmitted data be de-identified. Parwani[24] and Patel[2] use 'honest brokers,' third parties pre-approved by the DRN's responsible IRB, to de-identify medical record information through automated or manual methods. Only the honest brokers have access to the linkage codes between data and identifier. Local pre-processing of protected health information to avoid its transfer is mentioned in several publications but details are lacking.[7 12 22]

IRB oversight is not required for public health surveillance activities. The Privacy Rule permits the disclosure of protected health information if the organization tracks such disclosures.[5]

Facet 5: Query alignment and approval

Data queries should be approved by the data providers to ensure alignment with privacy protections and available resources. This is often accomplished through a portal that restricts queries to a pre-determined set of data elements. The Query Execution Manager is an example of an asynchronous 'pull' approach, which incorporates data providers in the query approval and execution process.[11] In the 'pull' approach, programmer-analysts and/or investigators at participating sites receive and review a new query, and decide whether to run it against their local data. The queries are accessed through a web portal, encrypted email, or similar interface. The encrypted results are uploaded back to the hub or original requestor, usually in delimited (csv) or SAS files.[1 4 7 11 25] Other DRNs that use this approach include the Mini-Sentinel Network[4] and the Nationwide Health Information Network (Healtheway).[7]

Some systems allow researchers within the network to query local data synchronously. Harvard's Shared Health Research Information Network (SHRINE) is one example.[33] In this 'push' query type, the query is directly processed by the remote query sender. In contrast, the HMORN does not permit a researcher external to the organization to directly submit queries to local data, but an external researcher may be sent a study dataset under an IRB-approved protocol. Portal approaches taken by these and others[2 24] also include ensuring that the user is authorized to request the data specified in the query. Several publications elaborated on the tools and format that DRNs used to conduct data queries.[1 4 7 11 25]

Facet 6: Data use

Data use refers to the purposes for which data are requested, accessed, and analyzed. These activities fall into three categories: preparatory to research (PTR), subsequent to obtaining IRB approval of a research protocol such as cohort identification for descriptive and multivariable analyses, and public health surveillance. PTR activities include queries that return aggregated counts or de-identified datasets which contain only aggregated count data, typically to assess the feasibility of a study or to develop sample size calculations.

In contrast, a limited dataset is often required for cohort identification and descriptive statistical analyses. However, in a distributed network, it is difficult to create the single, observation-level dataset required for multivariable analysis. For such analyses, sites may create a pooled analysis dataset, perhaps containing covariance matrices obtained from running separate regression analyses at each site, which are then combined for further regression analysis.[8 34] Methods for accomplishing this more easily are under development.[35 36]

Facet 7: Data security

Several documents described policies or procedures for secure transmission and storage of results through virtual private networks, data encryption, firewalls, and password protection. These networks included the Cancer Research Network,[17] Bioterrorism Syndromic Surveillance Demonstration Program,[22] and Mini-Sentinel Network.[4–6] Each included architectural as well as procedure information. Password protection for access to query software was mentioned only in Patel[2] and Parwani[24]; these two networks (the Pennsylvania Cancer Alliance Bioinformatics Consortium and the Early Detection Research Network colorectal and pancreatic neoplasm virtual biorepository) utilize a centralized database, in contrast to other DRNs.

Facet 8: Data retention

As Willison mentioned, data retention should be a concern among partners in a research network.[37] Only one document in our corpus mentioned procedures for data retention. McGraw states that 'data partners are required to keep the information that has been transformed into MSCDM and used to respond to queries for 3 years.'[6] If additional data are needed in the case of a suspected safety signal, the data partner is 'expressly limited to collecting additional data solely for the purpose of confirming the signal—the data must be destroyed within 3 years according to national standards for data destruction.'[6]

Facet 9: Data audits

Data audits are performed to evaluate information system and data integrity, identify unauthorized system access, and ensure that data are appropriately collected and represented. In any healthcare or health research context, data audits are required under Section 13411 of the HITECH Act. In a DRN, data audits also ensure that data are used within approved research protocols.

Several DRNs in our review have well-defined auditing functions. The Cancer Research Network has a central auditing authority that ensures that each participating institution has the technical support for maintaining security and privacy logs. Auditing cannot be left solely to the local level to address systematic security and privacy issues, but local sites may add auditing procedures.[17] In the Nationwide Health Information Network, the system logs the identity of the requestor, the identity of the agency that certified the investigation, and the time of query. This audit trail allows data providers to identify controversial credentialing and challenge agencies' queries and deny access.[7] The Early Detection Research Network has an audit review system in which 5% of new entries are re-examined by honest brokers, the cancer registrar, and data managers. Findings and recommendations are submitted to the project coordinating committee.[24]

Facet 10: User training

Training new users of any DRN is essential for ensuring adherence to policies, procedures, and standards. Our review of the literature revealed two documents where user training was described. In one, the HMORN analyzed past user experiences to assist with training.[38 39] In the other, drawing on the experiences of the HMORN and practice-based research networks leveraged by the Clinical Translational Science Award, researchers developed an extensive training resource, the Research Toolkit.[40] The Research Toolkit is a large repository of scholarly articles, IRB documents, and proposal development guides. Although not reviewed here, users should know that it contains a substantial amount of information about data governance as it applies to multi-site studies.

DISCUSSION

Our review identified practices of, and challenges posed by, the governance of clinical research DRN data warehouses. A recent review that focused on the growth of health information technology and particularly electronic medical records and their use in comparative effectiveness research, further highlighted these challenges.[41 42]

The literature on DRN data warehouse governance is immature, with only 39 documents retrieved in a broad search of the biomedical and computer and information science literature. Only a few of the 20 Clinical Data Research Networks identified

in a recent technical report[43] have published information about their data governance in the peer-reviewed literature. Of note, many more documents (N=183) describing non-US systems were retrieved, primarily describing DRNs and related systems in the UK. Much is still to be to learned about the challenges posed for data warehouse governance for DRNs in the USA. For example, research is a small component of managed care organizations (MCOs), and research within the MCO is often dominated by day-to-day organizational and financial demands. A research advocate should be involved in organizational decisions to ensure that researchers can take advantage of in-house expertise, such as data specialists, governance experts, and regulatory compliance professionals.

Several additional implications and recommendations emanate from our analysis. First, researchers and public health surveillance experts should develop standard operating procedures for safeguarding data, conduct periodic compliance audits, and provide educational and technical support to facilitate uptake procedures. Following the lead of the PRIMER project,[44] procedures should be documented and published so that they may be evaluated and used by others. Second, codification of DRN data warehouse governance policies and procedures should be a priority as the DRN is designed and implemented and revised periodically as new demands arise. A meta-policy should be in place that provides oversight and approval by representatives across the DRN. Third, an independent oversight function within the DRN should review the data and processes to foster trust among data contributors. Fourth, a shortcoming of the literature is that costs associated with data warehouse governance have not been addressed.

The framework we used in our analytic review of the literature is but one of several. We used the framework we deemed most amenable to modification for the DRN context. However, a new framework or taxonomy could be developed specifically for the DRN community to use in evaluating governance as the DRN area evolves. For example, the Scalable PArtnering Network for Comparative Effectiveness Research (SPAN),[45] a DRN with 11 participating sites, has begun a framework for the DRN community to use that is detailed in 'The SPAN: Purpose, Structure, and Operations' document, posted in the AcademyHealth Repository (http://repository.academyhealth.org/govtoolkit/3/). Although not meeting our corpus inclusion criteria, this document provides the SPAN governance guidelines. Finally, few of the documents in our corpus described policies for complying with HIPAA or IRB requirements, and we recommend that identifying and cataloging these policies should be undertaken in a comprehensive study that includes the gray literature.

Above all, it is important to consider that the DRN data warehouse governance is highly specific to the partnering institutions, the target research domain(s), and the network user community. Furthermore, numerous DRNs were not represented in our corpus because no indexed literature was available for them. The recent compendium of research networks provided by Ohno-Machado et al[43] is an excellent resource for those seeking to understand their function.

LIMITATIONS
Much governance documentation resides in the gray literature, such as web sites and industry white papers. The primary limitation of our review is reliance on the indexed scientific literature. We chose to restrict our document corpus to this literature because it has undergone peer review and includes reports of data warehouse governance specifically in the DRN domain. Few DRNs have published materials about data governance, and

the seeming dominance of the HMORN and Mini-Sentinel Network in our review reflects the fact that they have published relatively extensively. We stress here that this review is intended as a starting point for those working in the area of data governance and DRNs. A more comprehensive review of data governance policies and procedures will require a much larger study involving detailed primary data collection from all types of research data networks.

CONCLUSION
As we develop data resources to support a learning health system,[46] a consistent framework is necessary to govern an increasingly networked environment. Making sure that clinical research DRNs are properly governed will increase public trust and limit risk to, and encourage greater participation by, those holding primary data sources.

Clinical research DRN data warehouse governance policies provide important protections beyond data infrastructure and security. Articulating written governance agreements assists in developing and maintaining a common vision and purpose within the DRN, fosters trust and collaboration across the DRN data providers, and provides a template for addressing issues as they arise. Researchers planning to implement or improve existing data warehouse governance for DRNs need better guidance from the literature. However, as our review suggests, the dearth of DRN governance documents in the peer-reviewed literature indicates that this might not be the appropriate venue for publishing governance policies due to an inability to pass peer review, or be compatible with journal scope or editorial policy. This poses substantial difficulties for the informatics and clinical research communities as we move forward to a more distributed research environment. We thus encourage DRNs to publish information on publicly available websites about their data warehouse governance programs used to support DRNs and to develop and publish metrics that can be used to assess the impact of network governance on the efficiency of research and the protection of patients and participating organizations.

Author affiliations
[1]Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA
[2]University of Minnesota Medical School, HealthPartners Institute for Education and Research, Duluth, Minnesota, USA
[3]Harvard Medical School Department of Population Medicine, Boston, Massachusetts, USA
[4]Kaiser Permanente Colorado Institute for Health Research, Denver, Colorado, USA
[5]Denver Health and Hospital Authority, Denver, Colorado, USA
[6]HealthPartners Institute for Education and Research, Minneapolis, Minnesota, USA
[7]Research Informatics, Center for Health Research, Kaiser Permanente, Portland, Oregon, USA
[8]Kaiser Permanente Colorado, Denver, Colorado, USA

REFERENCES
1 Toh S, Platt R, Steiner JF, et al. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther 2011;90:883–7.

2 Patel AA, Gilbertson JR, Showe LC, *et al*. A novel cross-disciplinary multi-institute approach to translational cancer research: lessons learned from Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC). *Cancer Inform* 2007;3:255–74.

3 Wagner EH, Greene SM, Hart G, *et al*. Building a research consortium of large health systems: the Cancer Research Network. *JNCI Monographs* 2005;2005:3–11.

4 Curtis LH, Weiner MG, Boudreau DM, *et al*. Design considerations, architecture, and use of the Mini-Sentinel distributed data system. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):23–31.

5 Forrow S, Campion DM, Herrinton LJ, *et al*. The organizational structure and governing principles of the Food and Drug Administration's Mini-Sentinel pilot program. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):12–17.

6 McGraw D, Rosati K, Evans B. A policy framework for public health uses of electronic health data. *Pharmacoepidemiol Drug Saf* 2012;21:18–22.

7 McMurry AJ, Gilbert CA, Reis BY, *et al*. A self-scaling, distributed information architecture for public health, research, and clinical care. *J Am Med Inform Assoc* 2007;14:527–33.

8 Platt R, Davis R, Finkelstein J, *et al*. Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics. *Pharmacoepidemiol Drug Saf* 2001;10:373–7.

9 Velentgas P, Bohn RL, Brown JS, *et al*. A distributed research network model for post-marketing safety studies: the Meningococcal Vaccine Study. *Pharmacoepidemiol Drug Saf* 2008;17:1226–34.

10 Rosenbaum S. Data governance and stewardship: designing data stewardship entities and advancing data access. *Health Serv Res* 2010;45(5 Pt 2):1442–55.

11 Brown JS, Holmes JH, Shah K, *et al*. 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.

12 Maro JC, Platt R, Holmes JH, *et al*. Design of a national distributed health data network. *Ann Intern Med* 2009;151:341–4.

13 Fernandes L, O'Connor M. Data governance and data stewardship. Critical issues in the move toward EHRs and HIE. *J AHIMA* 2009;80:36–9.

14 McGarvey PB, Ladwa S, Oberti M, *et al*. Informatics and data quality at collaborative multicenter Breast and Colon Cancer Family Registries. *J Am Med Inform Assoc* 2012;19(e1):e125–8.

15 Fullerton SM, Lee SS. Secondary uses and the governance of de-identified data: lessons from the human genome diversity panel. *BMC Med Ethics* 2011;12:16.

16 Shapiro M, Silva SG, Compton S, *et al*. The Child and Adolescent Psychiatry Trials Network (CAPTN): infrastructure development and lessons learned. *Child Adolesc Psychiatry Ment Health* 2009;3:12.

17 Manion FJ, Robbins RJ, Weems WA, *et al*. Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study. *BMC Med Inform Decis Mak* 2009;9:31.

18 Randhawa GS, Slutsky JR. Building sustainable multi-functional prospective electronic clinical data systems. *Med Care* 2012;50(Suppl):S3–6.

19 Standards & Interoperability (S&I) framework—Query Health. Office of the National Coordinator 2013 [cited 2013 Sep 3]. http://wiki.siframework.org/Query+Health

20 Office of the National Coordinator. Standards and Interoperability Framework—Query Health. Standards and Interoperability (S&I) Framework 2013 [cited 2013 Aug 1]. http://wiki.siframework.org/Query+Health

21 Go AS, Magid DJ, Wells B, *et al*. The Cardiovascular Research Network. *Circulation* 2008;1:138–47.

22 Lazarus R, Yih K, Platt R. Distributed data processing for public health surveillance. *BMC Public Health* 2006;6:235.

23 Magid DJ, Gurwitz JH, Rumsfeld JS, *et al*. Creating a research data network for cardiovascular disease: the CVRN. *Expert Rev Cardiovasc Ther* 2008;6:1043–5.

24 Parwani A, Singh H, Dzubinski L, *et al*. Design and utilization of the colorectal and pancreatic neoplasm virtual biorepository: an early detection research network initiative. *J Pathol Inform* 2010;1:22.

25 Baggs J, Gee J, Lewis E, *et al*. The Vaccine Safety Datalink: a model for monitoring immunization safety. *Pediatrics* 2011;127(Supplement 1):S45–53.

26 Bailey H, Agger W, Baumgardner D, *et al*. The Wisconsin Network for Health Research (WiNHR): a statewide, collaborative, multi-disciplinary, research group. *WMJ* 2009;108:453–8.

27 Tucci DL, Schulz K, Witsell DL. Building a national research network for clinical investigations in otology and neurotology. *Otol Neurotol* 2010;31:190–5.

28 Brown J, McDonald S, Moore K, *et al*. PS1-20: VDW Operational Committee: current activities and future directions. *Clin Med Res* 2010;8:61–2.

29 Bloomrosen M, Detmer D. Advancing the framework: use of health data—a report of a working conference of the American Medical Informatics Association. *J Am Med Inform Assoc* 2008;15:715–22.

30 Gardner J, Xiong L, Xiao Y, *et al*. SHARE: system design and case studies for statistical health information release. *J Am Med Inform Assoc* 2013;20:109–16.

31 Braff J. PS1–03: expanding the functional scope of institutional review board review of multi-site research in the HMO Research Network: from data-only studies, to non-clinical interventions, to clinical trials. *Clin Med Res* 2011;9:175–6.

32 Bredfeldt CE, Butani A, Padmanabhan S, *et al*. Managing protected health information in distributed research network environments: automated review to facilitate collaboration. *BMC Med Inform Decis Mak* 2013;13:39.

33 McMurry AJ, Murphy SN, MacFadden D, *et al*. SHRINE: enabling nationally scalable multi-site disease studies. *PLoS ONE* 2013;8:e55811.

34 Wu Y, Jiang X, Kim J, *et al*. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012;19:758–64.

35 Ohno-Machado L, Bafna V, Boxwala AA, *et al*. iDASH: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc* 2012;19:196–201.

36 Kim K, McGraw D, Mamo L, *et al*. Development of a privacy and security policy framework for a multistate comparative effectiveness research network. *Med Care* 2013;51:S66–72.

37 Willison DJ. Trends in collection, use and disclosure of personal information in contemporary health research: challenges for research governance. *Health Law Rev* 2005;13:107–13.

38 Godwin D, Mahajan B, Silcott M, *et al*. C-A1-04: Challenges faced as new members of the HMO Research Network virtual data warehouse. *Clin Med Res* 2011;9:184.

39 Thompson EE, Greene SM, Larson EB. PS1–15: What has the HMO Research Network done for you, lately? How the network is responding to investigators' needs. *Clin Med Res* 2010;8:46.

40 Research Toolkit. Better Tools for Multi-site Research 2013 [cited 2013 Aug 15]. http://www.researchtoolkit.org

41 Lopez MH, Holve E, Sarkar IN, *et al*. Building the informatics infrastructure for comparative effectiveness research (CER): a review of the literature. *Med Care* 2012;50(Suppl):S38–48.

42 Holve E, Segal C, Hamilton LM. Opportunities and challenges for comparative effectiveness research (CER) with electronic clinical data: a perspective from the EDM forum. *Med Care* 2012;50(Suppl):S11–18.

43 Ohno-Machado L, Alipanah N, Day M, *et al*. *Clinical Data and Patient Research Networks, Patient Registries and Patient Powered Research Networks, Taxonomy and Comprehensive Inventories*. Washington, DC: Patient-Centered Outcomes Research Institute, 2013.

44 Greene S, Thompson E, Baldwin LM, *et al*. PS1-29: Resources to facilitate multi-site collaboration: the PRIMER research toolkit. *Clin Med Res* 2010;8:207.

45 Scalable PArtnering Network for Comparative Effectiveness Research (SPAN). Kaiser Permanente Colorado Institute for Health Research 2013 [cited 2013 Aug 1]. http://www.span-network.org/

46 Smith M, Saunders R, Stuckhardt L, *et al*. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Institute of Medicine, 2012.