Federalist principles for healthcare data networks

Kenneth D Mandl & Isaac S Kohane

Applying federalist principles to networked health record data could facilitate realization of the potential of shared health data.

The $48 billion US investment in electronic health record (EHR) adoption[1] was predicated on a promise—that data stored electronically, rather than on paper, would not only be used for care but would also underpin precision medicine by enabling research on a population scale[2]. Instrumenting the health system for discovery promises the identification of new drug targets, repositioning of medications, partitioning of populations for personalized medicine, quantification of the impact of environment on disease, comparison of the effectiveness of treatments and postmarket surveillance of therapeutics. EHRs have already spawned 'big data' about patients, including diagnoses, notes, laboratory results and medications. In the near future, patient genome sequences, tissue-specific transcriptomic profiles and other high-dimensional data types will enter clinical workflows and support medical decisions. And once appropriate informed consent procedures are in place, patients might enrich these clinical data by linking to additional data from self-reports, pharmacies, health monitoring apps, wearable devices and social media[3].

To achieve their full potential in population research, EHR data must be combined across hospitals and clinics to create large-*N* data sets[4]. Rare-disease research needs millions of patients to match study criteria[5]. Genomic studies seeking to quantify the weak effects of myriad genetic variants require hundreds of thousands[6]. Accountable-care risk calculations need data stratified by demographics and clinical characteristics. Similarly, clinical trials, device development and quality-improvement studies often require multiple sites to provide adequate statistical power.

Recently, the US federal government has allocated hundreds of millions of dollars to centrally managed research and public health networks dependent on data from, and participation by, hundreds of healthcare organizations. Physicians, researchers, department chairs and health system chief executive officers and chief information officers (CIOs) are grappling with a mountain of data access requests from myriad disparate governmental, scientific and commercial constituencies. A hospital CIO already manages dozens of mandated outbound data feeds to federal and state agencies, the Joint Commission and numerous payers. Each requires a different format, and the CIO must invest resources to accurately respond to each distinct request.

Although network participation presents opportunities that could result in improved population health, enabling access to these data is expensive and complex. Because hospital IT departments are the main source of big data for medical discovery, the approach taken to engaging health system participation will dictate the pace, scale and cost of discovery.

How can we put in place mechanisms to ensure that EHR data can be accessed by all those who seek to exploit its potential to reshape health research? In the following article, we argue that network organizers who understand both the health system perspective and health system needs are best placed to drive biomedical progress most effectively.

Instrumenting healthcare systems for research

There will be countervailing pressures on networked data. The status quo in the research enterprise is to generate data *de novo* and for each particular project. In traditional clinical trials or disease-based registries, the resources for data collection are usually provided by a pharmaceutical industry grant or infrastructure grant from the US National Institutes of Health (NIH; Bethesda, MD). Worryingly, this approach has produced expensive data sets that are rarely reused in follow-up studies. In this context, the most direct and simplest approach to acquiring data, in a prospective multicenter study, has been to define a comprehensive common data model for a specific prospective study and then require that each clinical site hew closely to that model.

The fundamental idea behind using health system data to drive research is that a preponderance of potentially useful data is already collected during the course of routine patient care in clinics and hospitals. There are tremendous efficiencies to be realized in using these existing data rather than collecting all data *de novo* for each study (**Table 1**). An alternative approach—reshaping data acquisition across the diverse and unruly delivery system to collect, at the point of care, all the data we might conceivably want for research—is a harder battle: a battle worth fighting, but one that will not be won quickly.

For now, in forming a data network, each node—a hospital, a health system or a practice—collects data during routine care and then vends it to a variety of internal and external customers. Achieving efficiency in the reuse of these data for research means conforming research use questions to the data that have been collected. Research questions that can

Kenneth D. Mandl & Isaac S. Kohane are at Boston Children's Hospital, Boston and Harvard Medical School, Boston, Massachusetts, USA.
e-mail: kenneth_mandl@harvard.edu

Table 1 Instrumented health system study versus traditional trial or registry

	Traditional clinical trial or registry	Instrumented health system study
Data source	All data generated during and for the trial	Electronic health records, bio-specimen banks, laboratory information systems, payor claims, e-prescribing data, inpatient pharmacy data
Data specifications	Data formats fully specified but traditionally specific to the particular study rather than universal	Highly varied clinical data formats, with federal specification by the CMS and other agencies slowly increasing
Data acquisition	Data meticulously collected by trained personnel according to well-specified standard operating procedures	Data collected during the course of routine care by nonstandardized systems, including the 'free text' dictation of physician notes
Study design	Study design fully specified, including data types acquired	No preexisting nationwide standard of data from laboratory systems, or for annotations such as clinical notes
Study hypotheses	Small number of hypotheses tested—e.g., is drug A superior to drug B; often no secondary analysis is planned	Myriad questions to be asked and hypotheses to be tested in the future, not specified at the time of data acquisition
Cost	High cost for data standardization and collection	Low cost for acquisition, but variable cost for transformation and transmission

only be answered with fully standardized, complete and perfect data of the sort collected during a traditional pharmaceutical trial are not the right ones to ask in a study dependent on a contemporary instrumented health system.

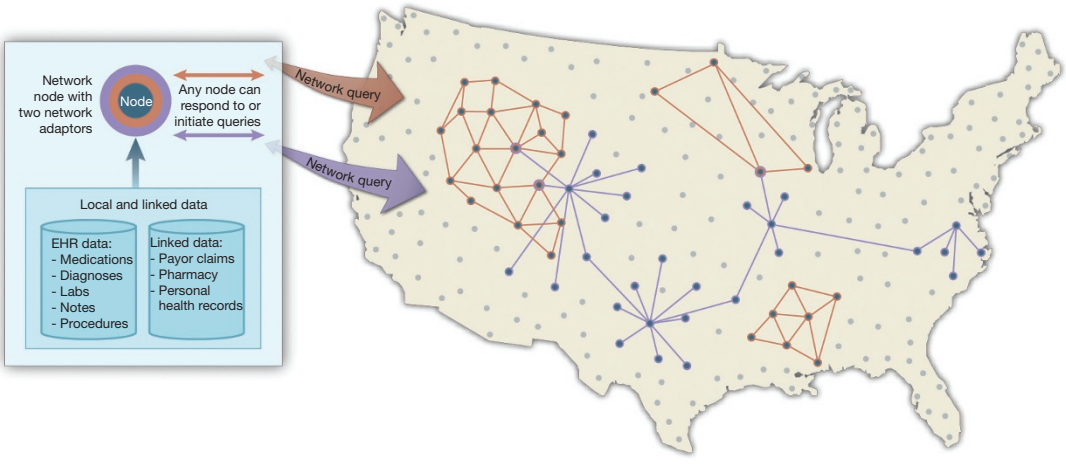
So while a coordinating organization designing a health-system study might initially assume that it looks far simpler and less risky to fully specify all aspects of the data transactions according to the particular study, the problem is that any data model that the network articulates that doesn't closely adhere to the data formats as they were initially collected incurs a substantial effort for 'harmonizing' these data to the study data model and study terminology. For example, a typical hospital EHR will have dozens of blood glucose measurement types (such as fasting blood sugar, 30-minute post–oral glucose challenge and 10-minute post–growth hormone challenge). Whether or not a particular mapping of these glucose data types to the glucose measurements of a study will make sense will depend on the study and the questions being asked. Furthermore, each such effort required will reduce the likelihood of a healthcare institution participating in the study and reduce the efficiencies gained from using health system data in something close to its native format.

The healthcare organization perspective

By what principles should a healthcare organization choose to share data, by what tenets should those efforts be governed, and which technologies are most cost effective? With each request, organizations weigh concerns regarding privacy, leakage of business intelligence and cost against a local benefit to the organization or a public good. The value of participation by healthcare organizations in data research networks must be skillfully framed if the healthcare system is to nimbly use clinically generated data to learn, discover and improve.

Data federation and network design. To network data collected during routine care requires a priori agreement on standards for data exchange, a process called federation. Federated data enables the identification of cohorts suitable for hypothesis testing—for example, a list of all patients in the network with ulcerative colitis who have been prescribed infliximab. If the network captures robust longitudinal trajectories, outcome and epidemiology studies are possible—for example, measuring the incidence of lymphoma in children on infliximab. At the commencement of a clinical trial or creation of a disease registry, leaders agree upon standard formats for data collection. Ideally, data collected for routine care by EHRs could be used in the native format. But in practice, every installation of every brand of EHR generally stores data in a unique, proprietary format, and those data need to be extracted from the EHR for meaningful analysis in another software system.

The most important decision by a federated research network is whether to *combine* data or keep data separate. This decision is ultimately controlled by those responsible for the care of the patients described by these data. This can



Rachel Eastwood

Figure 1 A self-organizing federated data research network. Each institution becomes a node able to join diverse networks by extracting data from the EHR and, as allowable under consent, linking to external data sources. Because different networks may require different data sets and formats, each node may attach more than one 'network adaptor' (here shown in blue and brown), a structure that enables on-the-fly data transformation into the network format. This network design enables a health system to (left) invest in a single informatics resource to serve potentially dozens of 'customers' of its data and (right) use an informatics resource that serves local data and workflow needs (such as patient contact, consent, trial matching and analysis) beyond responding to a query for any given data requestor.

Box 1 Principles for stakeholder engagement in federated networks

Several principles are essential for the creation of a successful federated, multi-institutional network. We outline these below:

- **Transparency.** Local institutions need to know who is accessing their data and for what purposes.
- **Representation.** Participating institutions must take part in the design, selection and approval of studies using networked data.
- **Local benefit.** Networks should support participating organizations with data and analytic tools to advance local research agendas and clinical improvements. Local needs can be distinct: one size does not fit all populations or healthcare systems, so the tools must provide flexibility in data access and a range of analytic capabilities. An informatics capability for easy exploration—interactive queries with real-time response—particularly empowers local users, as does the capacity to implement local clinical research workflows such as data analysis, consenting, patient contact and trial matching.
- **Right to reassortment.** The mesh of participating healthcare institutions should be able to organize and reorganize into opportunistic and productive networks, including for short-lived projects. Each network's infrastructure should support participating institutions in readily joining multiple networks with low overhead (see **Fig. 1**).
- **Cost neutrality.** Participation should be cost neutral, either because participating brings monetary value (for example, gaining insight toward lower-cost care pathways) or because payment is commensurate for work required.
- **Access.** Investigators at participating institutions should be able to use the network as peers. Experts gaining access across the federated network will advance science, clinical care and public health, generating wholly different and transformative sets of questions that no single committee can achieve.
- **Parsimony of data storage standards.** Emerging federated networks should engage health systems by understanding their needs and then helping to simplify the handling of outbound data. Networks should avoid requiring expensive, one-off data transformations. Health systems have no choice but to comply with requirements of the US Centers for Medicare and Medicaid Services (CMS; Washington, DC). Because the CMS requires nearly all healthcare organizations to transform EHR data into standardized formats for health information exchange, divergent costly transformations should be avoided. Instead, if a network requires a specific format, a simple transformation from the CMS-required format, which can be performed on the fly—as needed for the query—should be defined. Also, other emerging clinical formats (such as the Blue Button Initiative for data access by patients or SMART Platforms[14] and the emerging Fast Healthcare Interoperability Resources standards (FHIR) for exchange of data with apps) should be investigated as a *lingua franca* for interinstitutional and intersystem data transfer.

be accomplished by keeping data at each site of care in a traditional distributed model. In a less familiar but more powerful network model, the data are stored centrally, but they are not combined and are still controlled by each local institution[7]. In general, when healthcare data are combined centrally, they are de-identified. Another crucial decision is whether the system that stores the local data and answers queries either facilitates local workflows or simply vends data to a central network.

Local data, local benefit. The distributed model, in which each institution's data are maintained separately, not only affords local control over data and participation in studies but also enables member institutions to develop important local applications that use their own data plus networked-derived intelligence. For research studies, local processes include patient contact, patient consent, record review and patient-permissioned linkage to external data sources: for example, outreach to a cohort of patients on infliximab to solicit patient-reported measures of side effects, including shortness of breath or numbness. Other examples could be the enrollment of ulcerative colitis patients in a pragmatic randomized trial testing efficacy of antibiotics during a flare, or the collection of biospecimens to assess interleukin-15 receptor α expression after infliximab treatment in a sample cohort of patients. Clinical uses of network data and intelligence are almost exclusively local. Distributed data networks could have benefits beyond simply providing local control over a database that can be queried. A thoughtfully designed network permits a full end-to-end informatics system for research with federated data[8]. That is, rather than merely informing investigators, local control of a decentralized network enables policy to be effected by those with direct care responsibility for the patients whose data are monitored.

Pros and cons of decentralized healthcare networks

In 1994, when the World Wide Web was only two years old, the Boston-based W3EMRS collaboration that was the first to use web protocols[9] to federate EHR data from disparate systems. The collaboration assembled on-the-fly patient medication and problem lists from five hospitals. In the next decade, we introduced a public health outbreak detection system using a distributed architecture with robust institutional controls[10]. More recently CARRANet, a thriving federated, distributed rheumatology network involving 60 children's hospitals, has begun to blend data for discovery and care improvement, offering full institutional control over participation and data contribution on a project-by-project basis[7].

Early successes with distributed data networks have encouraged major federal programs, including the US Food and Drug Administration's Mini-Sentinel Network and the NIH-funded Shared Health Information Research Network (SHRINE). These efforts and others involving payors, big pharma, the US Patient Centered Outcomes Research Institute (Washington, DC), foundations and public health agencies now vie for clinical data from the clinical care sites.

The Internet should serve as the model for a distributed system—every web server can be a 'peer' and can pull data from other servers. And in the health system, every participating hospital or practice should be able to use the network as a peer. In the distributed healthcare data model, benefit is maximized while costs are minimized through on-the-fly translation of EHR data into established data standards. The data are then securely conveyed across the web to authorized users. The approach mirrors the successful and incremental evolution of the web standards (such as HTML and HTTP) themselves to create an international system of data sharing transcending any single use case.

Despite these advantages, recent experience suggests that implementation may not be without challenges. There have been several high-profile failures of large-scale, costly federally funded federated healthcare networks that ignored the founding principles of the internet and world wide web. The US National Cancer Institute's (Bethesda, MD) $300 million caBIG implemented a top-down model, developing tools that found only limited adoption by

healthcare enterprises. The one-size-fits-all approach failed; the most elaborate tools developed centrally were the least likely to be adopted. Notably, the standards defined had not been vetted in the marketplace before being required of the health care systems[11]. Likewise, the Centers for Disease Control and Prevention's (CDC; Atlanta, GA) Biosense, an emergency-department biosurveillance network, conducted analyses centrally and returned limited value to participants while competing with existing efforts and losing engagement of network members.

To foster participation and sustainability of distributed networks, what incentives should be offered to organizations, and how should the path to sharing be facilitated? We make recommendations based on a history of successful low-marginal-cost multi-institutional data sharing systems in **Box 1**, outlining several principles that are essential for creating a successful self-organizing federated network (see **Fig. 1**).

Conclusions

The US founding fathers debated federalist principles—the balancing of federal authority and states' rights. For healthcare to become a data-driven enterprise that learns from itself[12] and drives towards the practice of precision medicine[13], health systems, hospitals and even provider practices must become instrumented for discovery research and cost-effective federation of data. How a research network will nimbly balance essential centralized functions with local participation and authority must be an important and essential national discussion.

The investment in instrumenting healthcare with information technology to better understand expenses, reimbursement, quality, disease evolution, public health processes and the genetic basis of disease could underpin an unprecedentedly fertile ecosystem of studies with multiple layers of validation and reproducibility testing. But there is risk of overwhelming the already overextended information technology staff of healthcare institutions. Engaging the health system means not treating it as a square peg being forced into the round hole of traditional prospective trials design. To collectively achieve sustainable national-scale data federation, institutions must be able to fluidly join or leave an evolving assortment of networks. The burden of data provision should be minimized by harmonizing research requests with mandated clinical data formats. Participating clinical organizations should be first-class peers in the network, both responding to and issuing queries (**Fig. 1**) and deriving meaningful local benefit.

The opportunity exists to create a flexible, dynamic resource that serves not only both current national and local needs, but also future functions not yet imagined. Instrumenting the healthcare system for research means leveraging existing care delivery processes to create the big data needed for discovery and improvement.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Blumenthal, D. *N. Engl. J. Med.* **362**, 382–385 (2010).
2. Kohane, I.S., Drazen, J.M. & Campion, E.W. *N. Engl. J. Med.* **367**, 2538–2539 (2012).
3. Weber, G.M., Mandl, K.D. & Kohane, I.S. *J. Am. Med. Assoc.* **311**, 2479–2480 (2014).
4. Dolgin, E. *Nat. Med.* **17**, 1525 (2011).
5. Patten, I.S. *et al. Nature* **485**, 333–338 (2012).
6. Berndt, S.I. *et al. Nat. Genet.* **45**, 501–512 (2013).
7. Natter, M.D. *et al. J. Am. Med. Inform. Assoc.* **20**, 172–179 (2013).
8. Mandl, K.D. *et al. J. Am. Med. Inform. Assoc.* **21**, 615–620 (2014).
9. Kohane, I.S., Greenspun, P., Fackler, J., Cimino, C. & Szolovits, P. *J. Am. Med. Inform. Assoc.* **3**, 191–207 (1996).
10. McMurry, A.J. *et al. J. Am. Med. Inform. Assoc.* **14**, 527–533 (2007).
11. Masys, D.R., Harris, P.A., Fearn, P.A. & Kohane, I.S. *Sci. Transl. Med.* **4**, 149fs32 (2012).
12. Friedman, C.P., Wong, A.K. & Blumenthal, D. *Sci. Transl. Med.* **2**, 57cm29 (2010).
13. National Research Council (US) Committee on a Framework for Developing a New Taxonomy of Disease. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease (National Academies Press, Washington, DC, 2011).
14. Mandl, K.D. *et al. J. Am. Med. Inform. Assoc.* **19**, 597–603 (2012).