Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS): Architecture

Kenneth D Mandl,1,2,3,4 Isaac S Kohane,1,2,3,4 Douglas McFadden,2,3 Griffin M Weber,2,5 Marc Natter,1,4 Joshua Mandel,1,4 Sebastian Schneeweiss,6 Sarah Weiler,7 Jeffrey G Klann,7 Jonathan Bickel,1,4,8 William G Adams,9,10 Yaorong Ge,11 Xiaobo Zhou,12 James Perkins,13,14 Keith Marsolo,15 Elmer Bernstam,16 John Showalter,17 Alexander Quarshie,18 Elizabeth Ofili,19 George Hripcsak,20 Shawn N Murphy7,21

ABSTRACT
We describe the architecture of the Patient Centered Outcomes Research Institute (PCORI) funded Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS, http://www.SCILHS.org) clinical data research network, which leverages the $48 billion dollar federal investment in health information technology (IT) to enable a queriable semantic data model across 10 health systems covering more than 8 million patients, plugging universally into the point of care, generating evidence and discovery, and thereby enabling clinician and patient participation in research during the patient encounter. Central to the success of SCILHS is development of innovative ‘apps’ to improve PCOR research methods and capacitate point of care functions such as consent, enrollment, randomization, and outreach for patient-reported outcomes. SCILHS adapts and extends an existing national research network formed on an advanced IT infrastructure built with open source, free, modular components.

INTRODUCTION
The Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS, pronounced ‘skills’) is one of 11 clinical data research networks (CDRN) funded by the Patient Centered Outcomes Research Institute (PCORI) in 2014. PCORI, a non-governmental organization created under the Patient Affordable Care Act seeks to identify patients with specific illnesses and clinical characteristics. A recent PCORI survey of all PCORnet sites revealed that 37% of the existing CDRN nodes and 31% of the PPRN nodes already used i2b2. Finally, we built SMART (Substitutable Medical Applications, Reusable Technologies)—a platform to enable any developer to contribute to an ‘App Store for Health and Research’ compatible with i2b2-SHRINE instances or compliant EHRs.

Fifteen years ago, SCILHS informatics leaders began a quest to develop informatics infrastructure and regulatory innovation that would convert the emerging electronic health record (EHR) into a research tool for improving patient outcomes. All of our work and open source toolkits have been supported by grants from the National Institutes of Health, Centers for Disease Control and Prevention, and Office of the National Coordinator of Health Information Technology (ONC). First, we built Indivo,1,2,3 the first personally controlled health record, which gave patients their data, and apps to make those data useful. Then, i2b2 (Informatics for Integrating Biology and the Bedside)4–7 created an open source analytic platform to the EHR, to fuse and analyze data produced by the delivery system, and identify research cohorts. i2b2’s flexible common semantic data model readily accommodates a variety of clinical data. Our next advance was SHRINE (Shared Health Research Information Network).8–10 a tool enabling investigators to query i2b2 nodes in real time across multiple sites for collaborative population research. i2b2 has been successfully implemented at more than 100 sites across the USA, thereby enabling investigators to use delivery system data to identify patients with specific illnesses and clinical characteristics. A recent PCORI survey of all PCORnet sites revealed that 37% of the existing CDRN nodes and 31% of the PPRN nodes already used i2b2. Finally, we built SMART (Substitutable Medical Applications, Reusable Technologies)—a platform to enable any developer to contribute to an ‘App Store for Health and Research’ compatible with i2b2-SHRINE instances or compliant EHRs.

These informatics tools and associated research policy advances have already contributed to transformation in the clinical research enterprise—real-time, collaborative population health research is now enabled across SHRINE member sites distributed nationally—but they have yet to yield substantial improvements in the health of our patients. Now, in establishing PCORnet, PCORI has catalyzed a new national research dialog to answer patient-oriented questions and improve human health. We directly address this challenge via a strategy intended to avoid prior mistakes of large-scale, top-down, costly software infrastructure efforts that
failed to scale (e.g., caBIG\textsuperscript{14}), instead building SCILHS with open source, free, modular components\textsuperscript{5} with vibrant user and software developer communities that have already spread virally to scale across heterogeneous health systems.

Here, we detail the informatics approaches taken by SCILHS to identify large cohorts of patients and engage them for research. Our technology strategy links lockstep to processes for regulatory innovation, development of robust governance constructs and policies, and local adoption by hospital leadership and institutional review boards.

**THE SIDECAR APPROACH**

SCILHS adopts and extends a strategy of establishing a freely accessible health data ‘sidecar’ warehouse to the EHR, effectively leveraging existing data collected by EHRs during routine care while avoiding costly, time-consuming EHR integrations (Figure 1). Developed intensively over the past 5 years at Harvard Medical School, this approach employs vendor agnostic, free, open source, scalable, and interoperable technologies to produce the only research-based, shared repository of EHR data that can be queried in real-time. Of already proven value in the research ecosystem, these components support a cost-effective and sustainable research network of >8 million patients.

We consider the heterogeneity of collaborating institutions to be a key measure of success; via adoption of the sidecar approach, we enable any institution to join our SCILHS network. Specifically, a primary goal is inclusion of diverse populations within our CDRN network, thereby enabling capture of the genetic, genomic, and socioeconomic variation that exists beyond insured populations in managed care settings alone. Further, by freely sharing the processes and software that have been developed and supported by Harvard, we hope to catalyze the formation of many other new networks across heterogeneous health systems and institutions, and involve new

---

**Box 1 Alphabetical list of Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) sites**

Beth Israel Deaconess Medical Center  
Boston Children’s Hospital  
Boston Health Net (Boston Medical Center and Community Health Centers)  
Cincinnati Children’s Hospital Medical Center  
Columbia University Medical Center and New York Presbyterian Hospital  
Morehouse School of Medicine/Grady Memorial Hospital (Research Centers in Minority Institutions)  
Partners HealthCare System (includes Massachusetts General and Brigham & Women’s Hospital)  
University Mississippi Medical Center  
The University of Texas Health Science Center at Houston  
Wake Forest Baptist Medical Center

---

**Figure 1** Each site will install the Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) Sidecar, for identifying and reviewing cohorts and the mySCILHS suite to: (a) manage linkage of contact data to the de-identified Patient Cohort list produced by the multisite Shared Health Research Information Network (SHRINE) query; (b) administer and store consent documents; (c) outreach to patients through web-based survey and telephony; and (d) promote ongoing patient engagement through outgoing messaging, including (in the future) return of research results to patients. The web-based survey will be administered using REDCap and Indivo technologies, and will be accessible either by patients at home, or at the point of care, through tablet/kiosk-based interaction. Once completed, patient-reported data will have subject identifiers encoded; its standardized survey metadata will then be loaded into the corresponding SCILHS sidecar (i2b2 node), enabling semantic data linkage with electronic health record data via SHRINE/i2b2, while preserving subject confidentiality. These software platforms will be provided to sites as self-contained, pre-configured virtual machines, enabling rapid dissemination of these technologies while minimizing administrative and software development overhead at each site.
partners in improving our core components, common data models, and ontologies.

The sidecar infrastructure is composed of the following:

- **i2b2** (Informatics for Integrating Biology and the Bedside). Data analytic platform employed for EHR data analytics and clinical research at >100 academic medical centers worldwide (NIH funded).

- **SHRINE** (Shared Health Research Informatics Network). Federated query and response system that enables investigators to discover EHR data housed in i2b2 nodes across multiple independent institutions (NIH CTSA funded).

- **SMART Platforms**. First described in the *New England Journal of Medicine*, SMART has programmatic interfaces and applications that transform both EHRs and their sidecars into platforms that run substitutable iPhone-like apps. SMART enables a national scale ‘App Store’ for PCOR for rapid cycle innovation of PCOR methods (ONC funded).

- **Indivo**. The original personally controlled health record14 16 17 links patients to clinical and research settings. Used by hundreds of thousands of employees of Dossia’s founding companies (Wal-Mart, Intel, and AT&T), Indivo was also the initial software codebase for Microsoft’s HealthVault platform (NIH, CDC, and ONC funded).

- **REDCap** (Research Electronic Data Capture). Electronic data capture tool18 19 with 757 institutional partners, used to survey patients online (NIH CTSA funded).

### DATA MODELS AND ONTOLOGIES

SCILHS will combine EHR data with payer claims to facilitate interoperability of our platform (table 1). By enabling well-designed, cross-mapped ontologies that support a PCORnet common data model, this approach incorporates otherwise disparate clinical data sources into an easily-queried system that stores data in a flexible format. Data are stored in i2b2 using an entity–attribute–value model.20 21 employing a central ‘fact’ table based upon Kimball’s *Star Schema*22 wherein each row stores a flexibly defined, atomic ‘fact’ or observation for a patient.5 Much of i2b2’s versatility arises from its focus on a semantic definition of patient observations that can represent various existing and newly defined data elements: claims, EHR, genetic and imaging data, as well as patient reported outcomes and demographics. Analogous to a capacious warehouse with adjustable shelves and bins, i2b2 accommodates various nomenclatures for data elements, and supports robust tags of associated modifiers and values. This approach enables database indexing of facts and observations to support high performance execution of expressive queries and filters.

i2b2 employs an ontology-based approach that supports flexible, on-the-fly incorporation of new data elements and coding systems. Terminologies such as ICD, NDC, and LOINC may be pre-loaded as hierarchical concept trees; new or ad-hoc terminologies including patient-reported outcome measures or locally defined data dictionaries readily coexist and may be cross-mapped in i2b2. Concepts may be grouped using simple hierarchies and then optionally re-mapped into other reference coding systems and data models (eg, Observational Medical Outcomes Partnership (OMOP) data model). In this way, i2b2 accommodates diverse real-world coding systems while maintaining a straightforward query interface for its users.

The **SHRINE Adaptor Cell** maps local i2b2 terminologies into a common, standards-based SHRINE ontology. This enables a common shared ontology for federated queries while allowing individual i2b2 instances within institutions to retain local hierarchies and terminologies. The Adaptor transforms a federated SHRINE query into a query that runs on the local i2b2 database. The Adaptor then converts the result of that query back into the common SHRINE message format, using well-maintained standards including RxNorm, ICD9, and LOINC. In addition, SHRINE includes tools for ontology mapping and ontology-based data mining. Simple SHRINE customizations enable use of other query systems, for example the QueryHealth distributed query system (ONC) uses PopMedNet to query i2b2.24 25

### SUCCESS TO DATE

SHRINE and i2b2-based research includes characterization of rare morbidities of common diseases,26 very rare diseases such as peripartum cardiomyopathy (discovered in SHRINE and published in *Nature*25), detections of drug–drug interactions,28 and measures of quality and clinical efficacy across self-organized SHRINE networks in Europe, the University of California healthcare systems, and a just-in-time network to study the prevalence of complication rates of type 1 and type 2 diabetes in hospitals across this country. Others have used SHRINE to characterize and track the rising incidence of colorectal cancer29 and further characterize it, and to identify and optimize practice variation in irritable bowel disease and intervene to change that practice.30 i2b2 and SHRINE have been implemented as the base infrastructure for a variety of enhanced chronic disease registry-based research efforts.31 The Childhood Arthritis and Rheumatology Research Alliance uses the SHRINE/i2b2 registry framework to federate clinical care data and patient-reported data from 62 academic medical centers in the USA and Canada32 33 and is currently piloting consensus treatment protocol trials.34–37 The Harvard Inflammatory Bowel Disease (IBD) Longitudinal Data Repository employs the same infrastructure.35 ImproveCareNow38 utilizes i2b2 as its centralized data warehouse for IBD-related quality improvement development at 50 centers.

### Table 1 Approaches to scalability and interoperability

<table>
<thead>
<tr>
<th>Sidecar approach</th>
<th>‘Community-extensible ontologies’</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHR data are managed in a sidecar, readily available at any institution, regardless of EHR vendor product (Epic, etc)</td>
<td>All schemas and ontologies we produce are open source, free, and already widely adopted</td>
</tr>
<tr>
<td>i2b2 uses a simple data model (Star Schema) greatly simplifying the Extract, Transform, and Load procedure. These ETL procedures are established for all major EHR products</td>
<td>Ontologies can be imposed on the data after the fact, enabling a hospital in our network to readily adapt to any ratified PCORI Common Data Model</td>
</tr>
<tr>
<td>SMART platform specifications enable any app developer to create substitutable PCOR apps without knowing details about the underlying hospital systems</td>
<td>For example, there are existing transpositions between OMOP and i2b2 and PopMedNet can query i2b2</td>
</tr>
</tbody>
</table>

EHR, electronic health record; OMOP, Observational Medical Outcomes Partnership; PCORI, Patient Centered Outcomes Research Institute; SMART, Substitutable Medical Applications, Reusable Technologies.
PATIENT ENGAGEMENT

The health systems that have joined SCILHS reflect the American demographic—an essential requirement for reaching statistically valid, clinically meaningful, and patient-centric conclusions about therapies across the diverse spectrum of all healthcare consumers. In order to achieve the comprehensive, patient-centered outcomes infrastructure called for by PCORI, we introduce a new, patient-centric platform (mySCILHS) based on the Indivo system and incorporating the REDCap electronic data capture tool.

mySCILHS will support the Blue Button REST API for standards-based interactions with PPRNs and other patient-selected tools. This API exposes up-to-date, structured clinical summary data for each participating patient. Via a consumer-friendly workflow based on web standards including OAuth2, patients can authorize third-party apps and services, including PPRNs, to access their clinical data.

ANTICIPATED WORKFLOW

Figure 2 shows the workflow from an initial query through the analytic phase in a comparative effectiveness study. Each node in the network maintains an instance of i2b2 containing claims and de-identified electronic medical record data. SCILHS is a true peer-to-peer network, meaning that any SHRINE-based node can initiate a query, using a common ontology, that aggregates results from all participating sites. After the initial query, the investigator can automatically pass the query to each site where duly authorized local site investigators may review individual subject data for study eligibility using i2b2 SMART apps (figure 3). The final patient list is transmitted to the mySCILHS

Figure 2  The Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS) data workflow. We present here a general workflow. There will be important variations depending on the nature of the study, whether in-person consent is required, and whether patient identifiers are needed. Shared Health Research Information Network (SHRINE) architecture implemented as a modular framework. Using a mapper toolkit, each site exposes a common queryable data model, implemented in the ontology (ONT) cell. The ONT cell manages the vocabulary of the data model and is one of several cells in the i2b2 architecture, including the broadcaster-aggregator cell (AGG, broadcasts the query across all i2b2 nodes in the SHRINE peer-to-peer network and aggregates the results), the Identity Management Cell (IM, used for authentication), the Clinical Research Chart (CRC, manages the clinical data), the Workplace Cell (WORK, manages the workflow), and the Substitutable Medical Applications, Reusable Technologies (SMART) Cell (manages the SMART API). We implement the following workflow. A query from a Patient Centered Outcomes Research Institute (PCORI) approved study is translated to a SHRINE central node query either manually, or by a PCORNet adaptor, the specifications for which are still to be determined. The SHRINE Central Node broadcasts the query across the true peer-to-peer network (ARROW 1). i2b2 nodes containing coded data are queried at each site to identify appropriate patients returning obfuscated, aggregate patient counts (ARROW 2). Patient identifiable data remains at each site where investigators can use SMART Apps to review records prior to aggregation (ARROW 3). Also, see figure 3. The patient list is passed to mySCILHS for outreach to patients via apps, survey, or telephony (ARROW 4). Patient generated data are imported into i2b2 via simple input formats (CSV, for example) and placed into the i2b2 data model in a flexible schema that allows these to become first-class queryable data objects (ARROW 5). The adjudicated patient data (reviewed by investigators using SMART Apps and confirmed as valid) from each site, including patient-reported data can be added (ARROW 6) to a research data mart in one of several analytic data models (including the PCORI Common Data Model) with a level of identifiers appropriate to the level of consent obtained. Additional, outside data such as Centers for Medicare and Medicaid claims can be added in this step.
patient-facing software. The mySCILHS research contact management module links de-identified i2b2 records to patient demographics and contact information. Patients are engaged by web survey, telephony, or SMART apps; patient-reported data are returned to i2b2 and are then transferred into a secure comparative effectiveness (CE) study environment for analyses. In the CE environment, further transformations may occur, supporting many other analytic tools and processes. We anticipate that PCORnet-level queries, which may launch against the full complement of 11 CDRNs and 18 PPRNs, will be initiated at the PCORnet adapter. We anticipate that natural language processing (NLP) of provider notes will play an important role for adding complete longitudinal coded data to the hospital-based record. Early findings demonstrate that NLP of hospital-based EHR notes provides quite complete longitudinal data even when compared with Centers for Medicare and Medicaid Services claims data (personal communication, Katherine Liao, Brigham and Women’s Hospital, 2014). Using NLP on hospital and clinic notes will complement our strategy of concatenating EHR data with external sources such as claims and pharmacy data.

IMPLEMENTING AND SCALING

SCILHS includes 10 legally and financially independent institutions whose CEO or equivalent senior institutional official has committed to active participation in governance, policy development, data sharing, and sustainability planning. Each member has pledged to invest additional personnel and resources to ensure the network meets local patient and clinical stakeholder needs. By harmonizing informatics infrastructure, data models, regulatory processes and policies, and patient participation within and across member institutions, we anticipate that SCILHS will achieve and remain a successful model for inter-institutional PCOR. Utilizing the innovative SCILHS sidecar IT approach to EHR access, we minimize local informatics burden, further enabling a sustainable and adaptable PCOR infrastructure.

Author affiliations

1Children’s Hospital Informatics Program at Harvard–MIT Health Sciences and Technology, Boston, Massachusetts, USA
2Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA
3Harvard Catalyst, Harvard Medical School, Boston, Massachusetts, USA
4Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA
5Biomedical Research Informatics Core, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
6Division of Pharmacoepidemiology, Department of Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, USA
7Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts, USA
8Information Services Department, Boston Children’s Hospital, Boston, Massachusetts, USA
9Boston University School of Medicine/Boston Medical Center, Boston, Massachusetts, USA
10Boston University Clinical and Translational Sciences Institute, Boston, Massachusetts, USA
11College of Computing and Informatics, University of North Carolina at Charlotte, Charlotte, North Carolina, USA
12Department of Radiology, Center for Bioinformatics & Systems Biology, Wake Forest University Health Science, Winston-Salem, North Carolina, USA
13Clark Atlanta University, Atlanta, Georgia, USA
14Research Centers in Minority Institutions Translational Research Network, Data Coordinating Center, Jackson State University, Jackson, Mississippi, USA
15Cincinnati Children’s Hospital Medical Center, Cincinnati, Ohio, USA
16Division of Biomedical Informatics, Biomedical Informatics and Department of Internal Medicine, The University of Texas Health Science Center at Houston, Houston, Texas, USA
17University of Mississippi Medical Center, Jackson, Mississippi, USA
18Department of Internal Medicine, Community Health and Preventive Medicine and Clinical Research Center, Morehouse School of Medicine, Atlanta, Georgia, USA
19Department of Internal Medicine, Clinical Research Center, Morehouse School of Medicine, Atlanta, Georgia, USA
20Department of Biomedical Informatics, Columbia University, New York, New York, USA
21Partners HealthCare Systems, Information Systems, Charlestown, Massachusetts, USA

Acknowledgements We acknowledge the invaluable contributions of the many SCILHS investigators, leaders, and supporters and specifically call out those most involved in designing the network in the early phases: Barbara Bierer, Susan Edgeman-Levitan, Jonathan Finkelstein, Alisa Goldfine, Jennifer Haas, John Halamka, Manny Hernandez, John Hutton, Ann Kilbanski, David Ludwig, Joshua Metlay, Mary Mullen, Lee Marshall Nadler, Andrew Nierenberg, Harry Orf, Patricia O’Rourke, Eric Perakislis, Lee Schwamm, Daniel Solomon, Herman Taylor, Patrick Taylor, Aaron Waxman, Laura Weisel, and James Wilsson.

Collaborators SCILHS Network.

Contributors The authors all: made substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; and drafted the work or revised it critically for important intellectual content; and gave final approval of the version to be published; and agreed to be accountable for all aspects of the work in ensuring that questions related to the

Figure 3  A Substitutable Medical Applications, Reusable Technologies (SMART) Platforms HTML5 App running on i2b2, providing a richly featured electronic health record-like view of the data.
accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding This work was supported by the National Institutes of Health: National Library of Medicine R01LM011185 and U54 LM008748; National Institute of General Medical Sciences R01GM104303; National Center for Advancing Translational Sciences 1KL2TR001100, UL1TR000454; National Institute on Minority Health and Health Disparities U54 MD007588; the Office of the National Coordinator of Health Information Technology SHARP Program Contract 90TR0001; and by Contract CDRN-1306-04608 from the Patient Centered Outcomes Research Institute (PCORI).

Competing interests SS is consultant to WHISCON, LLC and to Aetion, Inc., a software manufacturer in which he also owns shares.

Provenance and peer review Commissioned; internally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

REFERENCES
21 Johnson SB, Friedman C. Integrating data from natural language processing into a clinical information system. Proc AMIA Annu Fall Symp 1996;537–41.