A Self-scaling, Distributed Information Architecture for Public Health, Research, and Clinical Care

ANDREW J. MCMURRY, CLINT A. GILBERT, BEN Y. REIS, PHD, HENRY C. CHUEH, MD, MS, ISAAC S. KOHANE, MD, PHD, KENNETH D. MANDL, MD, MPH

Abstract
Objective: This study sought to define a scalable architecture to support the National Health Information Network (NHIN). This architecture must concurrently support a wide range of public health, research, and clinical care activities.

Study Design: The architecture fulfills five desiderata: (1) adopt a distributed approach to data storage to protect privacy, (2) enable strong institutional autonomy to engender participation, (3) provide oversight and transparency to ensure patient trust, (4) allow variable levels of access according to investigator needs and institutional policies, (5) define a self-scaling architecture that encourages voluntary regional collaborations that coalesce to form a nationwide network.

Results: Our model has been validated by a large-scale, multi-institution study involving seven medical centers for cancer research. It is the basis of one of four open architectures developed under funding from the Office of the National Coordinator of Health Information Technology, fulfilling the biosurveillance use case defined by the American Health Information Community. The model supports broad applicability for regional and national clinical information exchanges.

Conclusions: This model shows the feasibility of an architecture wherein the requirements of care providers, investigators, and public health authorities are served by a distributed model that grants autonomy, protects privacy, and promotes participation.

Introduction
We describe our self-scaling, distributed architecture for health data exchange that meets the needs of public health, research, and care delivery. The work reported here builds on the Shared Pathology Informatics Network (SPIN)\textsuperscript{1–13} as a model to protect patient privacy, grant institutional autonomy, and exploit legacy systems and data sharing agreements. This approach has been successfully used nationally\textsuperscript{14,15} to share Health Insurance Portability and Accountability Act (HIPAA) de-identified\textsuperscript{16} human specimens.\textsuperscript{17–19} SPIN also influenced key aspects of the Markle Foundation's Connecting for Health Framework. (Shirky C, personal communication, 2005).\textsuperscript{20,21} Recognizing its broad applicability for exchanging clinical information, the SPIN model has been extended to satisfy the biosurveillance use case\textsuperscript{22} as defined by the American Health Information Community (AHIC). Through these examples, we demonstrate SPIN as a prototype architecture for the National Health Information Network (NHIN).\textsuperscript{22–24}

Background
Significance
Motivated by the need to detect infectious disease outbreaks, track influenza, and provide early warnings of bioterrorism, the AHIC has made biosurveillance a top priority for the NHIN.\textsuperscript{22} There is a growing consensus\textsuperscript{25} that a successful NHIN must standardize information storage and messaging formats, address privacy concerns, accurately identify patients, and resolve varying local, state, and federal regulations. These issues are pervasive across the NHIN use cases.\textsuperscript{22–24} For example, the biosurveillance use case requires both national anonymized coverage for routine analysis and provider authorized re-identification during emergency investigations. Importantly, our approach
not only fulfills the biosurveillance requirements but also supports research and routine clinical care on the same network.

**Shared Pathology Informatics Network**

Our NHIN architecture extends SPIN,\textsuperscript{1–13} which was originally funded\textsuperscript{1,26} by the National Cancer Institute to link the vast collections of human specimens that are infrequently shared for cancer research.\textsuperscript{18,19} SPIN sets forth the institutional agreements and distributed database architecture to grant institutional autonomy and protect patient privacy according to HIPAA regulations. SPIN has successfully completed a feasibility study involving seven independent medical centers sharing millions of human specimens.\textsuperscript{14,15}

Using a peer-to-peer architecture, institutions become SPIN members (nodes) by securing institutional review board (IRB) approvals and deploying the SPIN software. At any time, an institution can withdraw from the network without leaving their data behind or disabling the network. SPIN nodes can serve as “peers” or “supernodes” to query local databases or networks of child nodes, respectively.

SPIN allows institutions to expose de-identified pathology reports while keeping corresponding reports containing Protected Health Information (PHI)\textsuperscript{16} disconnected from the Internet. A randomly generated unique identifier is assigned to both the PHI and de-identified reports in a locally controlled “codebook.” The machine storing the codebook is disconnected from the Internet and protected according to each participating site’s policies. The resulting solution is flexible and compliant with HIPAA regulations.

SPIN provides three levels of increasing access commensurate with investigator credentials and IRB approvals.\textsuperscript{13} First, feasibility studies are conducted using a statistical level query that returns only aggregated results. Second, individual de-identified cases are selected by investigators certified by one of the participating institutions. The third level allows requests for specimens and clinical data that must be approved by the institution storing the requested data. Figure 1 illustrates the SPIN software components that enable increasing levels of investigator access.

**Biosurveillance Use Case**

The AHIC use case\textsuperscript{22} calls for a system that can aggregate biosurveillance data from a network of organizations, use existing data-sharing agreements, monitor patient disclosures, credential investigators, and ensure timely access. Implicit within these goals are patient de-identification during routine analysis and patient re-identification during emergency investigation. The AHIC use case is not implemented-specific and allows a wide range of transport methods (push and pull), access policies (HIPAA and institutional agreements), and identification systems (for patients and public health investigators). Many of these challenges were already addressed in whole or in part by the SPIN research effort, prompting us to develop an extension of SPIN to support clinical information exchange.

**Formulation Process**

The design of the SPIN architecture allowed us to adopt the SPIN distributed database, peer group routing subsystem, and query protocol without modification. However, the requirements of the biosurveillance use case are beyond the original intent of the SPIN research network. For instance, real-time surveillance could no longer use the codebook approach because clinical records must be immediately available during public health investigations. Also, we needed to resolve the variations in disclosure policies across institutions, states, and types of investigations (now going beyond research to include public health); disclosures could no longer be authorized by IRB approvals alone.\textsuperscript{27} Furthermore, public health investigators are not members of the hospital nodes, posing additional challenges to their identification and credentialing. The need for immediate access and external authority prompted significant enhancement of the SPIN de-identification, authorization, and auditing frameworks. Because these authorizations are commensurate with the strength of evidence of abnormal disease activity, we also needed a means to re-identify only the patients who signaled a potential public health threat. The resulting SPIN-based biosurveillance architecture is illustrated in Figure 2.

**Model Description**

Our architecture implements the biosurveillance use case and fulfills five desiderata. First, we adopt a distributed database to prevent creation of a monolithic repository, vulnerable to breach or misuse. Second, we enable strong institutional control to engender participation by the care delivery organizations and laboratories that provide data.\textsuperscript{28} Third, we ensure the accountability and oversight necessary to ensure public trust and protect privacy.\textsuperscript{29} Fourth, we facilitate real-time analysis of anonymized clinical records. Re-identification occurs only during a public health investigation, using only the cohort of encounters that signaled potential threat. This process happens under institutional control, according to hospital policies and commensurate with the needs of public health and the certified authority of investigators.\textsuperscript{30} Fifth, we ensure that our architecture is technically and socially scalable to extend participation, allow new data sources and applications, and facilitate voluntary collaborations in line with the goals of the National Health Information Network.\textsuperscript{31}
free text reports, also are removed. Each patient record is during investigations. Autocoding engines, then transformed and anonymized, for example by blurring geocoded home addresses, so that patient identity is protected, but sufficient location information is transmitted to detect clustering of cases. Other identifiers, such as patient names found in free text reports, also are removed. Each patient record is assigned a random link identifier to allow re-identification during investigations. Autocoding engines then transform free text input into a standard medical vocabulary. Finally, the anonymized and coded data are loaded into the peer database.

Institutions exchange digital certificates with approved peers to certify their identity and secure communications, forming “peer groups” (Fig. 4). Peer groups allow a single institution to concurrently participate in multiple public health, research, and clinical information exchanges. Because fully meshed networks require approval from every other institution, hub and spoke models are more commonly used. Hub and spoke models minimize the number of peer relationships using a single entry point (supernode) for each peer group.

Distributed Database
To leverage existing hospital databases and legacy information systems, we provide a 3-step pipeline of extraction, transformation, and loading modules (Fig. 3). First, patient records are extracted from local databases or extensible markup language (XML) files. The extracted records are then anonymized, for example by blurring geocoded home addresses, so that patient identity is protected, but sufficient location information is transmitted to detect clustering of cases. Other identifiers, such as patient names found in free text reports, also are removed. Each patient record is assigned a random link identifier to allow re-identification during investigations. Autocoding engines then transform free text input into a standard medical vocabulary. Finally, the anonymized and coded data are loaded into the peer database.

Institutions exchange digital certificates with approved peers to certify their identity and secure communications, forming “peer groups” (Fig. 4). Peer groups allow a single institution to concurrently participate in multiple public health, research, and clinical information exchanges. Because fully meshed networks require approval from every other institution, hub and spoke models are more commonly used. Hub and spoke models minimize the number of peer relationships using a single entry point (supernode) for each peer group.

**Figure 2.** Real-time biosurveillance architecture. (1) Filter, anonymize, format, and load biosurveillance data from existing systems into a locally connected peer node database. (2) Automated Routine Analysis. (3) Investigate Abnormal Visits. (4) Request Emergency Access. The modular design allows pluggable transformer modules to be added or removed, providing an extensible and reconfigurable system architecture. The SPIN-distributed query interface allows all members of a peer group to be contacted with a single query. Queries are performed by contacting the root supernode of the peer group, which propagates the message to each peer network or subnetwork until all peers are contacted. Results are aggregated asynchronously in reverse order. From the perspective of a client using the query interface (Fig. 5), there is no difference between a SPIN network query and a local query.

**Figure 3.** Three-stage pipeline: extract, transform, and load. The modular design allows pluggable transformer definitions to anonymize patients and process free text input.

**Figure 4.** Three examples of peer group configurations. Individual peer institutions are displayed in black, supernodes are displayed in light blue. From left to right: (1) Subnetwork of Harvard Medical School teaching hospitals. (2) National SPIN network bridging subnetworks. (3) Fully meshed peer-to-peer network.

The SPIN-distributed query interface allows all members of a peer group to be contacted with a single query. Queries are performed by contacting the root supernode of the peer group, which propagates the message to each peer network or subnetwork until all peers are contacted. Results are aggregated asynchronously in reverse order. From the perspective of a client using the query interface (Fig. 5), there is no difference between a SPIN network query and a local query.

**Institutional Autonomy and Distributed Access Controls**
Although secure transmission methods prevent third parties from “listening in,” over-the-wire encryption does not ensure that queries are made in good faith. Methods to identify and authorize public health investigators also are needed. This is challenging because peers cannot be expected to host up-to-date registries of investigators’ identities, yet hospitals must also remain able to authorize all disclosures. To address this, our framework allows trusted agencies to certify the identities and roles of investigators. Each institution specifies what is allowed to be disclosed for each role according to that institution’s policies. This required us to build a Distributed Access Control Framework, as illustrated in Figure 6.

**Investigator Accountability and Patient Trust**
All investigation scenarios record logging statements at each peer that cannot be removed by external parties. These logs contain the certified identity of the investigator, the identity of the trusted agency who certified investigation, and the time of query. Controversial credentialing of investigators may provide immediate disclosures but will always record an indelible audit trail, preventing clandestine investigations. Care providers are able to challenge the reasonableness of agencies’ queries and deny access to agencies if they do not keep patient and provider interests paramount.

Similarly, patients can audit care provider policies and investigator disclosures. This transparency is provided by the SPIN-distributed query, which returns the policies and query logs from all peers. Because all peer group members receive and log the same broadcasted query, a single institution cannot turn off logging or hide disclosures without coming under immediate scrutiny.

**Real-Time Anonymized Analysis and Patient Re-identification**
SPIN enables increasing levels of investigator access with peer-controlled disclosure. In the research case, investiga-
tors first review anonymized reports to initiate feasibility studies. Next they select a handful of cases for IRB-authorized research. We will now provide another validation of this principle with respect to public health surveillance. Using the SPIN approach, our Automated Epidemiologic Geotemporal Integrated Surveillance (AEGIS) biosurveillance system provides aberration detection, incurs minimal risk to patient privacy, and allows timely investigations to occur under emergency conditions.

Modern biosurveillance approaches rely on data mining to search for unusual patterns of disease. Hence, the algorithms require information on all encounters from all care provider locations. This qualitatively and quantitatively shifts the privacy tradeoff balance: disclosure of patient identity is necessary on only the subset of patients that is part of an identified or suspected outbreak, but automated analysis of all patient encounters is necessary to detect clustering.

Most biosurveillance systems look for spatial clustering among cases to signal possible outbreaks. The case locations are often based on patients’ home addresses, which are very identifying even if transmitted as geocoded coordinates or plotted on a map. Therefore, to preserve patient anonymity, the resolution location data for patients often is reduced to the zip code level. Although sharing patients’ zip codes is allowable under the HIPAA limited dataset, the loss in resolution diminishes the effectiveness of cluster detection algorithms. To preserve privacy while retaining cluster-detection power, we have implemented an algorithm that blurs the geocoded coordinates according to the underlying population density. We share the anonymized (blurred) addresses routinely, but share the precise addresses only when a cluster possibly signifying an outbreak is detected and there is agreement between the public health agency and the institutional data source that the data should be shared. SPIN provides a mechanism for increasing levels of investigator access commensurate with public health need and hospital policies (Table 1). First, a routine analysis query returns anonymized details of all patient encounters within a specified time period. When aberrations are detected, the investigator may re-authorize with a higher level of access and re-identify the list of abnormal patient visits.

**Routine Analysis**

The AEGIS biosurveillance system performs automated real-time analysis of anonymized patient encounters from all
participating care providers, running geospatial and temporal detection algorithms. Due to seasonal and other trends in the data, the algorithms perform better when given a long historical baseline to compare with current health care activity. SPIN provides broad, regional access (Table 1) to anonymized data while protecting patient identity and reinforcing institutional control.

**Alarm Investigation**

When aberrations are detected, alarm notifications are sent to public health agencies with an anonymized summary of the patient encounters that prompted investigation. If further investigation is necessary, officials will increase their access level network-wide by certifying the alarm investigation role (Table 1). This role can be used to request more detailed information about the aberrant patient encounters. The hospital will then return more detailed information for only those patients who signaled the alarm. This occurs either in an open loop mode, in which a person at the institution adjudicates each investigation, or in a closed loop mode, in which the institution returns identifying data if its policies allow it and the querying investigator presents an appropriate role and signed security token (Fig. 6).

**Emergency Investigation**

To ensure a rapid public health response under emergency conditions, we created a permanent closed loop mode for public health authorities in which they can exercise broad investigative powers. Individual institutions still are required to authorize this role, and as with all queries, accountability is enforced post hoc with audit trails. This requires an emergency level role (Table 1) and re-identification similar to that performed in the alarm investigation case. For care providers or institutions uncomfortable with disclosing patient PHI under any circumstance, the local emergency contact information (for example, the infection control nurse) can be provided in lieu of patient records. Investigation may continue through manual lookup of patient records at the source institution using the anonymized link identifiers or medical record numbers.

**Self-Scaling Architecture Promoting Timely NHIN Deployment**

The idea that the NHIN will be grown from the bottom up and not top down is gaining acceptance.\(^{25}\) We assert that our model fulfills this self-scaling need in the following respect: this architecture promotes individual participation and collaboration among Regional Health Information Organizations (RHIOs).\(^{41}\) As shown in Figure 4, a RHIO directly corresponds to a SPIN peer group. Autonomous peers form larger peer groups, and peer groups themselves, can be linked to form larger, networked communities. Autonomy is central to this organizational trust, and ensures that care providers remain stewards of patient privacy.

The SPIN model seeks to expedite early NHIN deployment by leveraging legacy information systems and existing institutional policies. For example, the federated identity and distributed access controls allow hospitals to continue using IRB and HIPAA authorizations. Other examples include the submission tools and query interfaces that extract, transform, and share data from existing databases using standard medical vocabularies. We believe the only way to ensure early participation is to make the technical and procedural burden as light as possible.

**Validation Examples**

**Research**

SPIN has demonstrated both national and regional viability for multi-institution cancer research efforts. On a national scale, SPIN investigators have completed a feasibility study involving seven large medical centers sharing a collective library of millions of annotated human specimens.\(^{14,15}\) On a regional scale, an operational version is deployed and in use at the Dana-Farber Harvard Cancer Center.\(^3\)

**Public Health**

This model is an essential component of one of four open architectures developed with funding from the Office of the National Coordinator of Health Information Technology, and this model fulfills the AHIC biosurveillance use case.\(^{22}\) In January 2007, this architecture was presented to the AHIC stakeholders using the live AEGIS system developed for the Massachusetts Department of Public Health.

**Clinical Care**

Clinical applications within the NHIN will require complete patient histories to be available regardless of where a patient receives care. Using the SPIN query interface, patients and physicians could locate records distributed across the network. Queries across the system could return data populating electronic health records or personally controlled health records.\(^{45,47,48}\) Patients also could authorize disclosures, review HIPAA compliant audit trails, and even consent to research for which they stand to benefit.\(^{49}\)

**Discussion**

**Significance**

Many scientists are calling for a closer connection between translational research and routine patient care. Although human specimens and patient histories represent a valuable resource in the postgenomic era, few investigators have authorization across all locations where patients receive care.\(^{27}\) We developed SPIN to link existing databases while building institutional agreement and protecting patient privacy. As a result, SPIN has been deployed across multiple locations and has addressed pervasive issues in sharing patient data. The broad applicability of this approach allowed us to develop a public health infrastructure with minimal effort. Specifically, we leveraged the distributed database to survey health statistics and detect disease aberrations. SPIN also
Table 2  A Single Peer-to-Peer Network to Support Public Health, Research, and Clinical Care

<table>
<thead>
<tr>
<th>Research</th>
<th>Nonidentifying</th>
<th>Identified Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>Feasibility studies</td>
<td>Case selection</td>
</tr>
<tr>
<td>Public health</td>
<td>Health statistics</td>
<td>Aberration detection</td>
</tr>
<tr>
<td>Care delivery</td>
<td>Environmental factors</td>
<td>Quality improvements</td>
</tr>
</tbody>
</table>

Conclusions

Development of an NHIN requires broad participation using the systems and policies already in place. Concerns about patient privacy and institutional control of data are pervasive throughout public health, research, and clinical care. We propose a distributed architecture that grants autonomy, protects privacy, and promotes participation.46,47

References