Human Pathology (2007) **38**, 1212–1225

ELSEVIER

Human PATHOLOGY

www.elsevier.com/locate/humpath

Original contribution

A system for sharing routine surgical pathology specimens across institutions: the Shared Pathology Informatics Network☆

Thomas A. Drake MD[a,*], Jonathan Braun MD, PhD[a], Alberto Marchevsky MD[b], Isaac S. Kohane MD, PhD[c], Christopher Fletcher MD, FRCPath[d], Henry Chueh MD, MS[e], Bruce Beckwith MD[f], David Berkowicz MD[e], Frank Kuo MD, PhD[d], Qing T. Zeng PhD[g], Ulysses Balis MD[h], Ana Holzbach PhD[e], Andrew McMurry[i], Connie E. Gee PhD[j], Clement J. McDonald MD[k,l], Gunther Schadow MD, PhD[k,l], Mary Davis MD[k,l], Eyas M. Hattab MD[k], Lonnie Blevins[l], John Hook[l], Michael Becich MD, PhD[m], Rebecca S. Crowley MD, MS[n], Sheila E. Taube PhD[o], Jules Berman PhD, MD[o], Members of the Shared Pathology Informatics Network[1]

[a]*Department of Pathology and Laboratory Medicine, UCLA Medical Center, avid Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA*
[b]*Department of Pathology, Cedars Sinai Medical Center, David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA*
[c]*Children's Hospital Informatics Program, and Harvard Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA*
[d]*Department of Pathology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*
[e]*Laboratory of Computer Sciences, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA*
[f]*Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA 02115, USA*
[g]*Decision Systems Group and Department of Radiology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*
[h]*Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02115, USA*
[i]*Children's Hospital Informatics Program, Harvard Medical School, Boston, MA 02115, USA*
[j]*Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA*
[k]*Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[l]*Regenstrief Institute, Indianapolis, IN 46202, USA*
[m]*Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA*
[n]*Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA*
[o]*National Cancer Institute, Bethesda, MD 20892, USA*

Received 26 June 2006; revised 6 January 2007; accepted 11 January 2007

☆ This work was supported by National Institutes of Health/National Cancer Institute #1UO1 CA91429 and U01 CA91343.
* Corresponding author. UCLA Path & Lab Med, BOX 951732, AL-124 CHS, Los Angeles, CA 90095-1732, USA.
E-mail address: tdrake@mednet.ucla.edu (T. A. Drake).
[1] See Acknowledgment.

doi:10.1016/j.humpath.2007.01.007

Keywords:
Pathology;
Informatics;
Internet;
Tissue bank;
Database

Summary This report presents an overview for pathologists of the development and potential applications of a novel Web enabled system allowing indexing and retrieval of pathology specimens across multiple institutions. The system was developed through the National Cancer Institute's Shared Pathology Informatics Network program with the goal of creating a prototype system to find existing pathology specimens derived from routine surgical and autopsy procedures ("paraffin blocks") that may be relevant to cancer research. To reach this goal, a number of challenges needed to be met. A central aspect was the development of an informatics system that supported Web-based searching while retaining local control of data. Additional aspects included the development of an eXtensible Markup Language schema, representation of tissue specimen annotation, methods for deidentifying pathology reports, tools for autocoding critical data from these reports using the Unified Medical Language System, and hierarchies of confidentiality and consent that met or exceeded federal requirements. The prototype system supported Web-based querying of millions of pathology reports from 6 participating institutions across the country in a matter of seconds to minutes and the ability of bona fide researchers to identify and potentially to request specific paraffin blocks from the participating institutions. With the addition of associated clinical and outcome information, this system could vastly expand the pool of annotated tissues available for cancer research as well as other diseases.

1. Introduction

In many areas of biomedical research, there is a recognition that current clinically oriented research initiatives often require access to larger numbers of patients and specimens than can be represented by a single medical center. Multicenter clinical trials are the norm because they typically represent the only realistic way to enroll a sufficient number of subjects in a reasonable time frame. To maintain adherence to study guidelines and ensure uniform data capture and analysis, such trials have strong top-down organizational structure and centralize data collection and analysis. However, many forms of translational research do not require such rigid control of patient management nor is it always desirable or feasible. For many research purposes, samples and data collected in the routine provision of patient care are sufficient. What is needed is access to a large number of samples with adequate supportive data and an efficient means of sharing data and samples. Pathologic tissue samples from cancer patients and associated data represent one such resource that exists in every hospital that provides cancer care.

In 2000, the National Cancer Institute issued a request for applications for the development of a Web-based system that would allow for cross-institutional searching of surgical pathology and autopsy specimens for research purposes, referred to as the Shared Pathology Informatics Network (SPIN). The vision was that routinely obtained tissue specimens that are retained for at least 10 years in pathology department archives represented a tissue resource that is orders of magnitude larger than the set of tissues collected prospectively for research purposes that compose tissue banks in selected academic centers. The scientific rationale for this is compelling from a variety of perspectives. For one, technologies for using such material are in hand and continue to improve. Tissue microarrays prepared from archival paraffin-embedded specimens are now a central tool for biomarker discovery, and more and more categories of malignancies are being studied. The human genome project has enabled a variety of genetic analyses. Cancer cells can be obtained by microdissection techniques, and DNA and RNA can be extracted from these cells for analysis. Gene expression analyses can be performed on both in situ sections as well on amplified cDNA product in the microarray format. Secondly, with a much greater number of tissues available, the questions asked can be more finely posed.

The goals set for SPIN were the development of a prototype system that allowed Web-based searching by researchers across data from multiple institutions, a mechanism for retrieval of desired specimens identified by such searches, and maintenance of compliance with requirements for appropriate patient confidentiality and consent as dictated by Health Insurance Portability and Accountability Act (HIPAA) and local Institutional Review Board (IRB) regulations. Two consortia were selected, and the involved institutions of both consortia worked closely throughout to develop the prototype system described here. The Harvard-UCLA consortium included the Harvard affiliated hospitals of Beth Israel Deaconess Hospital, Brigham and Women's Hospital, Children's Hospital, and Massachusetts General Hospital, all from Boston, and the UCLA Medical Center and Cedars Sinai Medical Center from Los Angeles. The Indianapolis-Pittsburgh consortium included 5 Indianapolis health care systems: Clarian Health Partners, Community, St. Vincent's, St. Francis l, and Wishard—which include a total of 15 different hospitals organized as a Regional Health Information Organization (RHIO) to serve clinical, public health, as well as research purposes [1]. The University of Pittsburgh hospital system (Pittsburgh, PA) was also part of this second consortium but operated as an independent institution in the SPIN

Table 1 Annotated list of SPIN-related publications

General:

Berman JJ. Pathology data integration with eXtensible Markup Language. Hum Pathol 2005;36:139-45. (An overview of the applications of XML for pathologists)

Grannis SJ, Overhage JM, McDonald CJ. Real world performance of approximate string comparators for use in patient matching. Medinfo 2004;2004:43-7. (A study of different name comparison methods for medical record linkage between independent sources)

Holzbach AM, Chueh H, Porter AJ, Kohane IS, Berkowicz D. A query engine for distributed medical databases. Medinfo 2004 2004:1519. (A description of the query tool developed for CHIRPs SPIN nodes)

McDonald C, Dexter P, Schadow G, et al. SPIN query tools for de-identified research on a humongous database. Proceedings of the American Medical Informatics Association Annual Symposium 2005:515-9. (A description of the query tool developed for the Indianapolis/Regenstrief SPIN node)

Namini AH, Berkowicz DA, Kohane IS, Chueh H. A submission model for use in the indexing, searching, and retrieval of distributed pathology case and tissue specimens. Medinfo 2004;11(Pt2):1264-1267. (A description of the CHIRPS model for uploading de-identified pathology case and specimen information to the SPIN network architecture)

Patel AA, Gupta D, Seligson D, et al. Availability and quality of paraffin blocks identified in pathology archives: A multi-institutional study by the Shared Pathology Informatics Network (SPIN). BMC Cancer 2007 (in press). (Results of an assessment of actual sample availability for cases identified using SPIN)

Schadow G, Grannis SJ, McDonald CJ. Privacy-preserving distributed queries for a clinical case research network. Proc. Privacy, Security and Data Mining. In Conferences in Research and Practice in Information Technology, 14. Clifton, C. and Estivill-Castro, V., Eds, ACS.55 (Maebashi City, Japan, 2002).

Tobias J, Chilukuri R, Komatsoulis GA, et al. The CAP cancer protocols—a case study of caCORE based data standards development for the Cancer Biomedical Informatics Grid. BMC Med Info Decision Making 2006;6:25. (Implementation of a computer-based representation of an existing paper standard of pathology data)

Deidentification:

Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12. (Description of a de-identification tool tailored for pathology reports)

Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127:680-86. (Description of a general algorithm that removes identifying or private information from pathology free text)

Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121:176-186. (Evaluation of a tool for removing safe-harbor identifiers and producing readable deidentified text that retains important clinical information)

Table 1 *(continued)*

Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp 2002:777-81. (A automated de-identification method using substitution methods and publicly available data sources)

Autocoding:

Berman JJ. Automatic extraction of candidate nomenclature terms using the doublet method. BMC Med Inform Decis Mak 2005;5:35. (A method for automatically extracting candidate nomenclature terms from virtually any text and any nomenclature)

Berman J. Modern classification of neoplasms: reconciling differences between morphologic and molecular approaches. BMC Cancer 2005;5:100. (A discussion of the importance of biological classification and an examination of different approaches to the problem of tumor classification)

Gilbertson JR, Gupta R, Nie Y, Patel AA, Becich MJ. Automated clinical annotation of tissue bank specimens. Proc MedInfo 2004:607-10. (Description of a system for automated annotation of banked tissue that integrates data from the cancer registry, the pathology LIS and the tissue bank inventory system)

Mitchell KJ, Becich MJ, Berman JJ, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports Proc Med Info 2004:663-67. (Description of a system for automated annotation of surgical pathology reports with UMLS terms that includes a module for handling negated concepts)

Mitchell KJ, Crowley RS, Gupta D, Gilbertson J. A knowledge-based approach to information extraction from surgical pathology reports. AMIA Annu Symp Proc 2003:937. (Description of a prototype system for knowledge-based information extraction from surgical pathology reports, including organ, procedure, and diagnoses)

Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc 2003:584-588. (Presentation of an automated method for extracting structured information about specimens and their related findings from free-text surgical pathology reports)

network. The acronym CHIRPS (for Consented High-performance Index and Retrieval of Pathology Specimens) as used in this report refers to the peer-to-peer (P2P) network and related software tools developed by the Harvard-UCLA consortium. All of the institutions in Indianapolis represented a separate network with its own software tools, and its features are referred to as the Indiana implementation. Together, these formed SPIN as described below.

Various aspects and perspectives of the SPIN project and its implementation are also presented in other publications as referenced in this report and are listed in Table 1, and a Web site can be accessed at http://spin.nci.nih.gov. We describe here the prototype system developed and the challenges met in doing so and discuss the implications

Table 2 Glossary

Architecture	The functional organization of a computer network.
Autocoder	A software program that assigns easily searchable codes to words or phrases of interest in a document in an automated manner.
CHIRPS	Acronym for "*C*onsented *H*igh-performance *I*ndex and *R*etrieval of *P*athology *S*pecimens", the Harvard-UCLA consortium implementation of SPIN.
Codebook	A database that maintains the match of anonymous identifiers for a specimen used in the SPIN node with the actual patient identifiers used by the local institution.
Data model	A plan that defines data elements and their relationships.
Deidentified	The state of having all information removed from a document that could be used to discern the identity of a person.
Firewall	A system of specialized software that limits network access between 2 or more networks.
HIPAA	The United States *H*ealth *I*nsurance *P*ortability and *A*ccountability *A*ct of 1996, enacted to establish standardized mechanisms for electronic data interchange, security, and confidentiality of all health care–related data [3].
HL7	A set of communications rules by which different computer applications exchange data in an orderly way [8].
IRB	Abbreviation for "*I*nstitutional *R*eview *B*oard", a committee empowered by an institution to sanction research that involves human subjects.
JAVA	A general purpose programming language with features that make it well suited for use on the World Wide Web.
Network	An interconnected set of computers residing at separate locations that can exchange information.
Node	An institutional server/database that is part of SPIN or other computer network
Open source	Computer software that is nonproprietary and whose source is available for review [27].
Parse	To breakdown text to component parts of interest.
Peer-to-peer	A P2P architecture allows hardware or software to function on a network without the need for central servers.
PHI	Abbreviation for "*P*rotected *H*ealth *I*nformation" as defined by HIPAA, which encompasses health information data that reasonably could be expected to allow identification of an individual [2].
Query	A user initiated request for information from SPIN about the number of cases available that fit criteria defined by the user.
Query tool	The software and associated user interface that allows a SPIN query to be made.
RHIO	Abbreviation for "*R*egional *H*ealth *I*nformation *O*rganization."
Scrubber	A software program that removes patient identifiable information in an automated manner.
SNOMED	Acronym for "*S*ystematized *N*omenclature *o*f *Med*icine", a comprehensive concept-oriented clinical vocabulary developed by the College of American Pathologists.
SPIN	Acronym for "*S*hared *P*athology *I*nformatics *N*etwork" (see http://spin.nci.nih.gov).
SQL	Acronym for "*S*tructured *Q*uery *L*anguage"; a particular form of programming language for extracting data from a relational database.
UMLS	Abbreviation for *U*nified *M*edical *L*anguage *S*ystem, a program of the National Library of Medicine. Here, referring to the UMLS Metathesaurus, a very large vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them [28].
URL	Abbreviation for "*U*niform *R*esource *L*ocators", which are the addresses used for Web pages.
XML	Abbreviation for e*X*tensible *M*arkup *L*anguage. It is a means of applying "tags" to items in a document that specify meaning [5].
XML schema	An XML document that defines the data structure and meaning for a given XML application.

such a system holds for pathologists. A glossary of abbreviations and terms is provided (Table 2).

1.1. Overview of how SPIN functions

The SPIN was configured as a network of locally controlled databases, each containing pathology specimen–related data from participating institutions (Fig. 1). For the CHIRPS implementations, each institution's pathology database was separate but derived from the local Pathology Information System (IS) and potentially other institutional sources. The specimens and supporting information represented in the database were completely under the control of the local institution. The data in each database were deidentified and contained no Protected Health Information (PHI as defined by HIPAA) [2-4]. For the Indiana implementation, the pathology, tumor registry, and other clinical data from each of the participating institutions were delivered to a central site [1]. There, each institution's data were maintained in separate physical files but appeared as a single node to the SPIN network.

The SPIN network was organized using a P2P architecture, which is described in greater detail elsewhere (Chueh et al, in preparation). (P2P is a particular mode of structuring a network using the Internet for communication that is popularly known as being the model used by music file sharing systems.) Users access a Web site where query forms are displayed that allow the user to define what they wish to search for using either text or codes. Once formulated, a query is sent out to the network. Each local

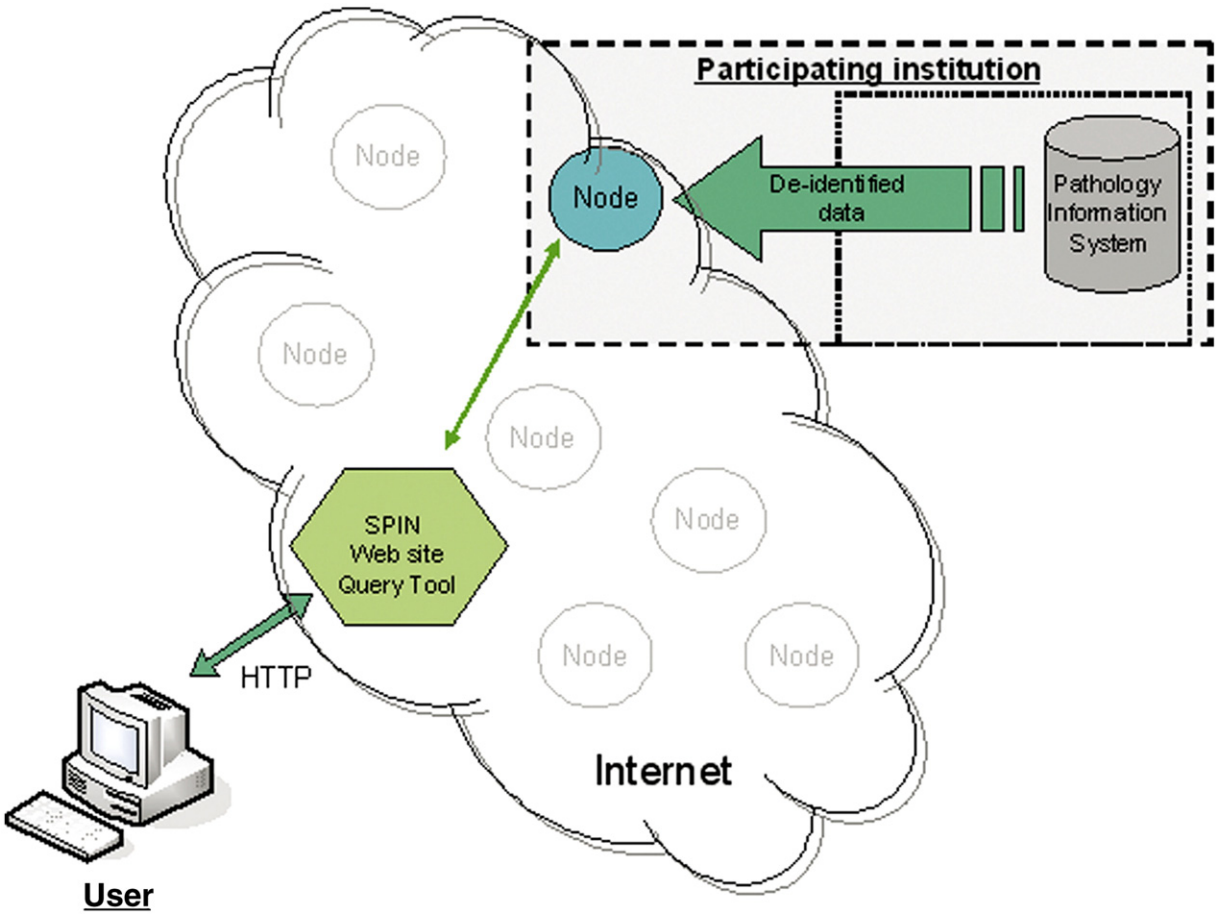


Fig. 1 Schematic of SPIN. Each node represents an institutional SPIN server with the database of deidentified data and associated software necessary for interacting on the network. A user accesses the network by connecting via the Internet to a SPIN Web site with a user interface to compose a query (the query tool). This translates the information request into a message format that is sent on to the participating institutions (nodes in the diagram) for searching their databases. Relevant results are sent back, aggregated, and displayed to the user on the SPIN Web site.

site (sometimes referred to as a "node" in the network) has a server (a computer open to the Internet) that has specific software that allows it to receive the query, interact with the database to find the relevant records, and send out a reply containing these (Fig. 1). Responses from each individual site are aggregated and displayed to the user in graphical and/or tabular format.

The software will allow users to identify a set of specimens according to their specifications and then initiate a request to obtain those specimens from the institutions holding them. The system notifies the sites where the desired specimens reside via the network, and the local site can identify the specimens through the local codebook for retrieval. This capability is being tested within the existing SPIN consortia. Development of procedures and policies for verifying that the user is a bona fide researcher and has IRB approval to use such specimens is also a necessary component. The identity of the patient from whom the specimen was obtained would never be provided to the requestor under any circumstances.

Development of this system involved overcoming a series of challenges that may not be readily apparent from the above description. As one would expect, a series of software tools needed to be developed that supported the query system and the P2P communication across institutions in the network. A database is needed at each site along with the software necessary to link it to the network. Equally important is the need for software tools for transferring relevant data from the local information systems into the local SPIN database. In addition to software to handle the transfer of data into the SPIN database, specific software is needed to remove any PHI (names, accession numbers, etc) as well as software to assign proper codes (eg, Unified Medical Language System [UMLS] Concept unique identifiers or Systematized Nomenclature of Medicine [SNOMED]) for diagnoses and other relevant information when these are not present. These various software tools are described in greater detail below, but in general are written in a common language (JAVA) and use nonproprietary software for common functions. They are

Fig. 2 A portion of data from a colon cancer case showing XML tagged items in the structure of the XML schema.

freely available for use and can be downloaded at https://sourceforge.net/projects/spin-chirps under an open source licensing program. (There is not a generic software tool for the initial steps of data extraction from the local information systems, as these vary with the system and structure at a given institution.)

Apart from software, other issues needed to be addressed. One was developing an agreed upon system for identifying the data in a uniform way (a "data model") so that there would be consistency in the type of information presented across institutions. Another was finding the most appropriate way to obtain IRB approval for the system at each institution that was consistent across institutions. These are issues that are relevant to any system that attempts to share patient-related information among multiple institutions and are discussed further below.

1.2. The nature of a SPIN database

The information one might retrieve would obviously depend, first of all, on the data that are held in each of the local SPIN sites. The local SPIN database holds records of the routinely obtained surgical and autopsy pathology specimens that the institution chooses to make available for searching by users. As discussed below, the records in the node database are derived from the local Pathology IS but represent an extraction of specific relevant data and fields. The records have been stripped of PHI elements and assigned an anonymous code.

For effective retrieval of information across the network, the data residing at individual sites need to be represented in a consistent manner. For this, it was necessary to develop a data model for pathology specimens and related information and for the CHIRPS implementation to represent these using eXtensible Markup Language (XML) [5-7]. This was constructed using the concepts and principles of the HL7 observation messages, a standard widely used in health care for delivering virtually all types of medical information [8,9]. The data model defines the relationship of the pathology specimen with its attributes and other medically relevant data linked to it.

The data model encompasses a large number of potentially available and useful elements. However, many of these may be unavailable or difficult to extract from existing reports. Therefore, a subset was defined as being the minimally necessary data elements for a specimen to be represented in the prototype version of a SPIN database. These included basic demographic information (gender, age, year of procedure), site, procedure, and diagnosis. In the database, these are represented in both primary text and coded formats to allow for searches using either. The primary text is automatically drawn directly from the pathology reports, whereas codes are assigned using an autocoding software tool during the data submission process of populating the database, as discussed further below.

Besides the above basic data elements, the data model includes additional fields for describing the pathology specimen itself, including gross and histologic features such as size and grade. Relevant laboratory and clinical data are supported, including cancer registry coding. In addition, novel supporting data such as those derived from expression microarray or proteomic analyses can be accommodated.

In the CHIRPS implementation, the XML representations of the elements of the data model are defined using an XML "schema." This is used for both populating the database and data retrieval when a query is made, although the node database is a relational SQL database (MySQL). The schema defines the data structure and in the submission process guarantees that the XML elements linked to it follow that structure. Each specimen is defined at a minimum by topological (anatomical), procedural, and diagnostic categories. An example of a portion of data in XML format from a specimen report is shown in Fig. 2.

1.3. Building the local pathology database

The elements that can be brought into the local node database are described above. Here, we discuss issues related to data extraction and tools developed to facilitate this process. The primary data used for populating the database are derived from the local Pathology IS. Pathology IS databases typically are relational databases that have defined fields for patient demographics, specimen identifiers, and descriptors such as tissue source and/or procedure, gross description, microscopic description, and final diagnosis. Except for patient demographics, dates, and accession numbers, most of the other fields are composed of narrative

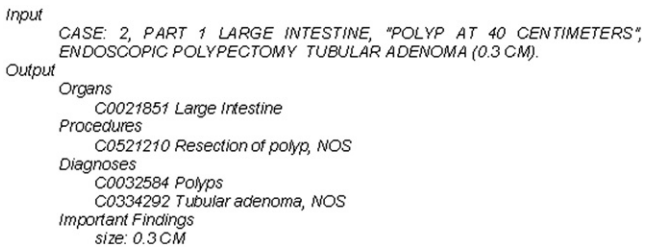


Fig. 3 Example input and output for the CHIRPS autocoder from a simple case.

text entries. This has 3 important implications for data extraction and display in the node database. First, to aid the query process, it is desirable to code tissue sources and diagnoses according to standardized nomenclatures. Second, text fields need to be "scrubbed" to remove any possible patient or institutional identifiers, such as patient or physician names. Third, to pull out certain desired data residing in text blocks, text needs to be "mined" or "parsed" using automated algorithms. The data extraction process for the SPIN node database captures the textual contents of data fields, such as gross description, as well as selected information derived from these using text mining software, as elaborated below.

A complicating factor is that Pathology ISs vary across institutions, and the data collected and available also vary over time. Given the goals of the SPIN project, it is highly desirable to capture specimens and related data that were obtained many years ago, as well as more recently obtained material. In part, this is to maximize retrieval of rare diagnostic categories, but it is also important because historical material provides a longer period of follow-up to determine outcomes. A survey of the archival pathology material in the participating institutions in the Harvard-UCLA consortium indicated that there were all together well over 1,000,000 surgical pathology cases with report data available online and expected to have archived blocks. Estimates of the fraction that were cancer-related varied from 1/5 to 1/3 among institutions. More than 1.5 million additional reports were identified by the Pittsburgh and Indianapolis resources.

As pathologists are painfully aware, the accessibility of historical data is highly dependent on (and varies with) the systems in place at the time for data collection and storage. Currently, most Pathology ISs are client server–based applications with data stored in relational databases. However, before roughly 1975 to 1980, when computer systems began to be more widely used, all data were stored in paper-based format. Most academic institutions that implemented computer systems made an effort to extract essential data from these paper records and include them in an abbreviated format in the computer system. Given this time frame, most pathology departments have also gone through several (often different) computer-based systems, and the extent and format of the archival data imported into the current system typically differ from that stored in the prior system. Even with current systems, much of the data are stored in aggregate text statements composed as preferred by the pathologist, as opposed to fields that are coded or use a standardized terminology. Use of standardized diagnostic coding such as SNOMED is inconsistent across institutions and certainly not retrospectively applied. In most cases where SNOMED codes do appear, they are the 20-year-old SNOP-2 (Systemized Nomenclature of Pathology) codes—not the more recently developed SNOMED reference terminoloy (RT) or SNOMED clinical terminology (CT) codes. The use of synoptic reports is just now being implemented for an as yet limited number of diagnostic categories.

From the above, it is obvious that the structure and extent of available data in current pathology information systems vary across institutions and within an institution across time and is largely still in text format. Extraction of the key elements defined by the data model therefore required the development of text mining tools. The details and validation of the performance characteristics of these are the subject of separate reports, as referenced below and as listed in Table 1.

Use of a scrubber tool is necessary because it is not uncommon for patient, physician, or institutional names and other identifiers to be used within the descriptive text of pathology reports, such as reference to consultation with a colleague or a prior specimen accession number in the diagnosis field. Because all data in the database must be deidentified, these must be removed, and with the processing of tens to hundreds of thousands of records, it is only feasible to do this in an automated fashion [10-13]. The scrubber developed for the Harvard-UCLA consortium uses a 3-pronged approach [10]. First, identifiers known to be associated with a patient, such as name, medical record number, case accession number, are removed. Next, a series of "regular expression clauses" are used that look for predictable patterns likely to represent identifying data, such as dates, accession numbers, addresses, and proper names [14]. The final step is a comparison with a database of names and places to recognize potential identifiers not removed by the first approach. The regular expressions and name lists are easily modified and improved locally by adding institution-specific names (eg, patients and employees). An evaluation showed that it was greater than 98% efficient at removing HIPAA identifiers with a minimal effect on essential text. Other SPIN consortium members have developed scrubber tools as well to serve the same purpose [11-13].

The autocoder tools search through (parse) the text of the report to find keywords or phrases representing the procedure, organ, and diagnosis terms (both assertion and negation of these terms) and assign codes to these based on the UMLS vocabulary [15-22]. Use of codes facilitates searching as discussed below. An example of this is provided in Fig. 3.

In the Harvard-UCLA implementation, the population of the data fields in a local node database was performed through the use of a software tool referred to as the "submission tool" [23]. This requires that the data pulled from the Pathology IS

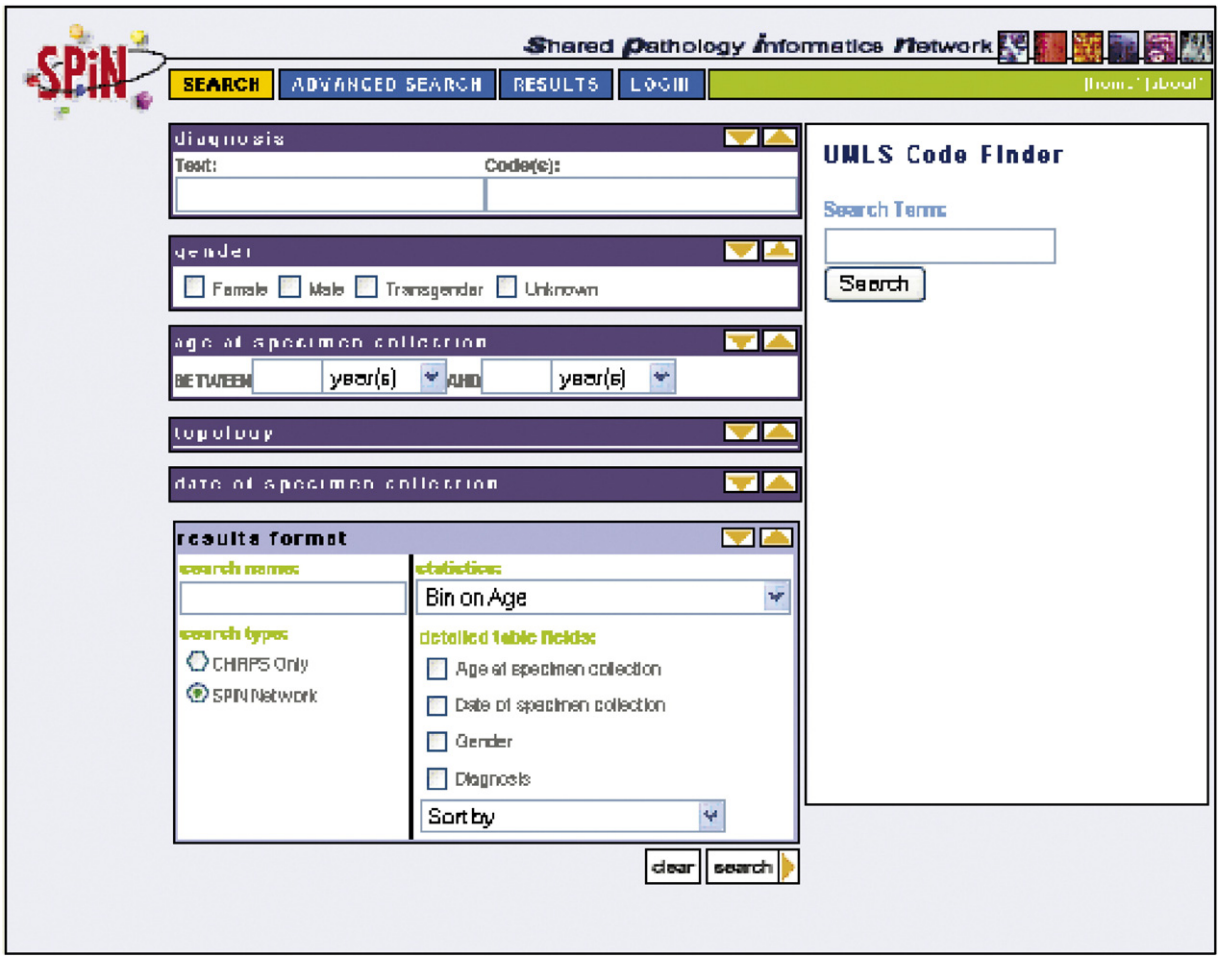


Fig. 4 Screenshot of the CHIRPS query tool search page.

be assigned XML tags according to the XML schema. It is then processed by the submission tool, following a scrubbing step and an autocoding step, and then the processed data are transferred to the appropriate fields in the database. The submission tool also creates an anonymous identifier (a UUID or Universally Unique Identifier) for the record that enters the

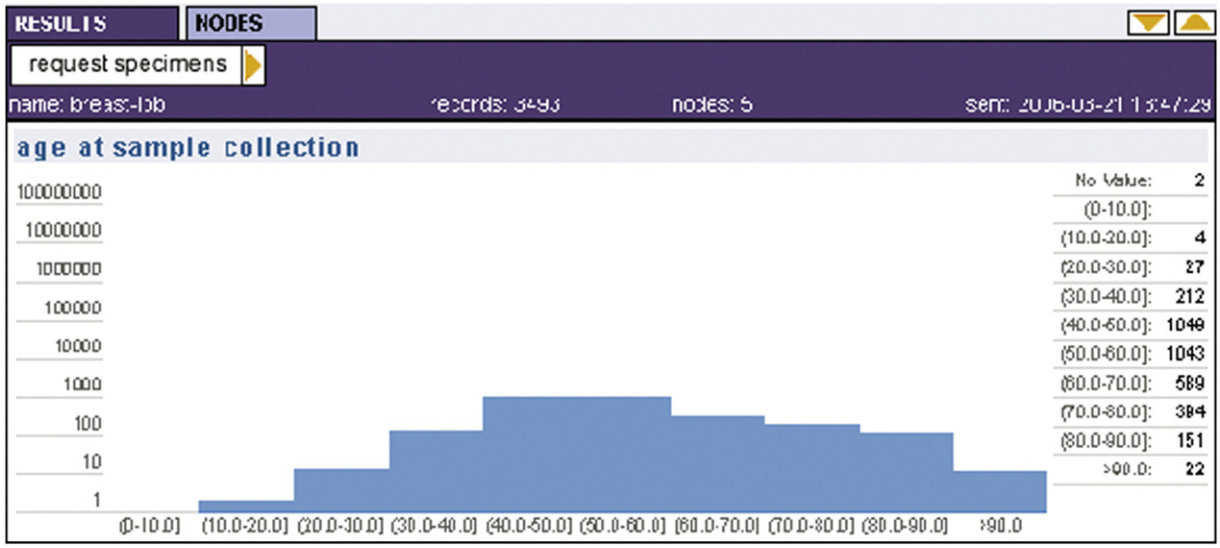


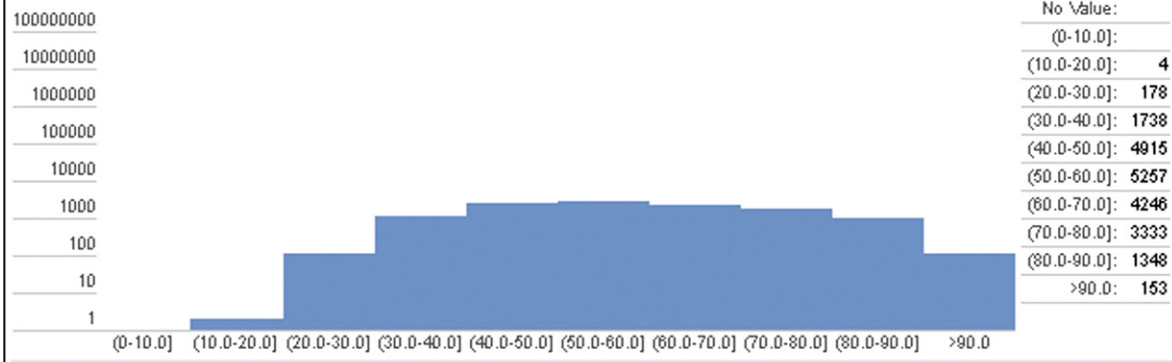
Fig. 5 Screen shot of graphically displayed results from a search for lobular cancer of the breast in females, by decade of age, using the CHIRPS query tool.

RESULTS | NODES

request specimens

name: Breast Ductal CA records: 32079 nodes: 7 sent: 2006-09-15 13:06:44

age at sample collection



No Value:	
(0-10.0]:	
(10.0-20.0]:	4
(20.0-30.0]:	178
(30.0-40.0]:	1738
(40.0-50.0]:	4915
(50.0-60.0]:	5257
(60.0-70.0]:	4246
(70.0-80.0]:	3333
(80.0-90.0]:	1348
>90.0:	153

detail

76	F		RIGHT BREAST, LOWER OUTER, 11 G STEREOTACTIC CORE NEEDLE BIOPSIES FOR CALCIFICATIONS: DUCTAL CARCINOMA IV SITU, comedo type (high nuclear grade) is present. Microcalcifications and necrosis are associated with DCIS. NOTE: There is chronic inflammation adjacent to an involved duct, suggestive of rupture of the duct with a few tumor cells in stroma. This is not definite stromal invasion.
52	F		LEFT BREAST, CAUDAD, WIRE LOCALIZED RE-EXCISION FOR CALCIFICATIONS (WIRE #3): DUCTAL CARCINOMA IV , cribriform and solid types (low to intermediate nuclear grade) is present in 8 of 11 blocks. Necrosis and calcifications are present. DCIS is less than 0.1 cm from the superior, posterior, and lateral margins (multiple foci). DCIS is greater than 0.3 cm from the inferior, anterior (skin), and medial margins. Atypical lobular hyperplasia with pagetoid spread is present. A prior biopsy site is present. DCIS extensively involves areas of sclerosing adenosis. An immunoperoxidase study for p63 and cytokeratin (AE1/AE3) reveals myoepithelial cells in the areas of DCIS. LEFT BREAST, CEPHALAD, WIRE LOCALIZED EXCISION FOR CALCIFICATIONS (WIRES #1 AND 2): DCIS, cribriform and solid types (low to intermediate nuclear grade) with central necrosis and calcifications, is present in 8 of 15 slides. DCIS is less than 0.1 cm from the inferior margin over a broad front and 0.1 cm from the posterior margin focally. DCIS is greater than 0.3 cm from anterior (skin), superior, medial, and lateral margins. Atypical lobular hyperplasia with pagetoid spread is present. Skeletal muscle is present. DCIS extensively involves sclerosing adenosis. Note: This carcinoma would be classified as AJCC Tis IV .
37	F		RIGHT BREAST WIRE LOCALIZATION FOR A MASS AND A CLIP (PART 1), AND AXILLARY NODE DISSECTION (PART 3): INVASIVE CARCINOMA, DUCTAL CARCINOMA IV AND NEGATIVE LYMPH NODES. SYNOPSIS: RESIDUAL INVASIVE CARCINOMA (See NOTE.) Size - 1.0 cm Multifocal - No Histologic Type - Mucinous Histologic Grade (SBR) - II(Moderate) Nuclear Grade - Intermediate EIC - Positive Lymphovascular invasion - Present Dermal lymphatic invasion - N/A Skin invasion - N/A Skeletal muscle invasion - N/A Final Margins (invasive) - Positive (ink and cautery on carcinoma) inferior-deep (block #3) DUCTAL CARCINOMA IV - Present Types - Micropapillary with apocrine features Nuclear grade - Intermediate Calcifications in DCIS - Present Central necrosis in DCIS - Absent Microinvasion (less than0.1cm) - Absent Extent of DCIS Number of blocks/total: 4/4 blocks Final Margins (DCIS) Negative (greater than0.3 cm) anterior, superior Very close (less than0.1 cm) medial (block #2), lateral (block #3) Positive (ink on DCIS) inferior (block #4), deep (block #3) LOBULAR CARCINOMA IV - Absent LYMPH NODES Axillary lymph nodes: Negative Number of positive nodes: 0 Number of nodes examined: 6 TOTAL POSITIVE NODES (axillary) - 0/6 Extranodal extension: Absent OTHER FINDINGS Biopsy site in axillary tissue SPECIAL STUDIES (ER/PR/HER-2) - Pending NOTE: There are pools of stromal mucin, some with floating malignant epithelium consistent with residual invasive mucinous carcinoma, involving a 1.0 cm area. There are multiple foci of residual DCIS and intralymphatic carcinoma in all 4 blocks. RIGHT BREAST, WIRE LOCALIZED EXCISION FOR MICROCALCIFICATIONS (PART 2): Apocrine cysts with calcifications. Florid epithelial hyperplasia. There is no evidence of malignancy.
37	F		LEFT BREAST, FREE HAND CORE NEEDLE BIOPSIES FOR A PALPABLE MASS: INVASIVE DUCTAL CARCINOMA, poorly differentiated (modified Bloom-Richardson grade III/III) measuring at least 1.2 cm in size in this limited specimen. No lymphovascular invasion is seen in this limited specimen. No in situ carcinoma is present in this limited specimen. Fat necrosis, acute and chronic inflammation and myofibroblastic proliferation, consistent with history of prior fine needle aspiration. NOTE: Immunohistochemical analysis for estrogen receptor, progesterone receptor, and HER2/neu will be performed and reported in an addendum.

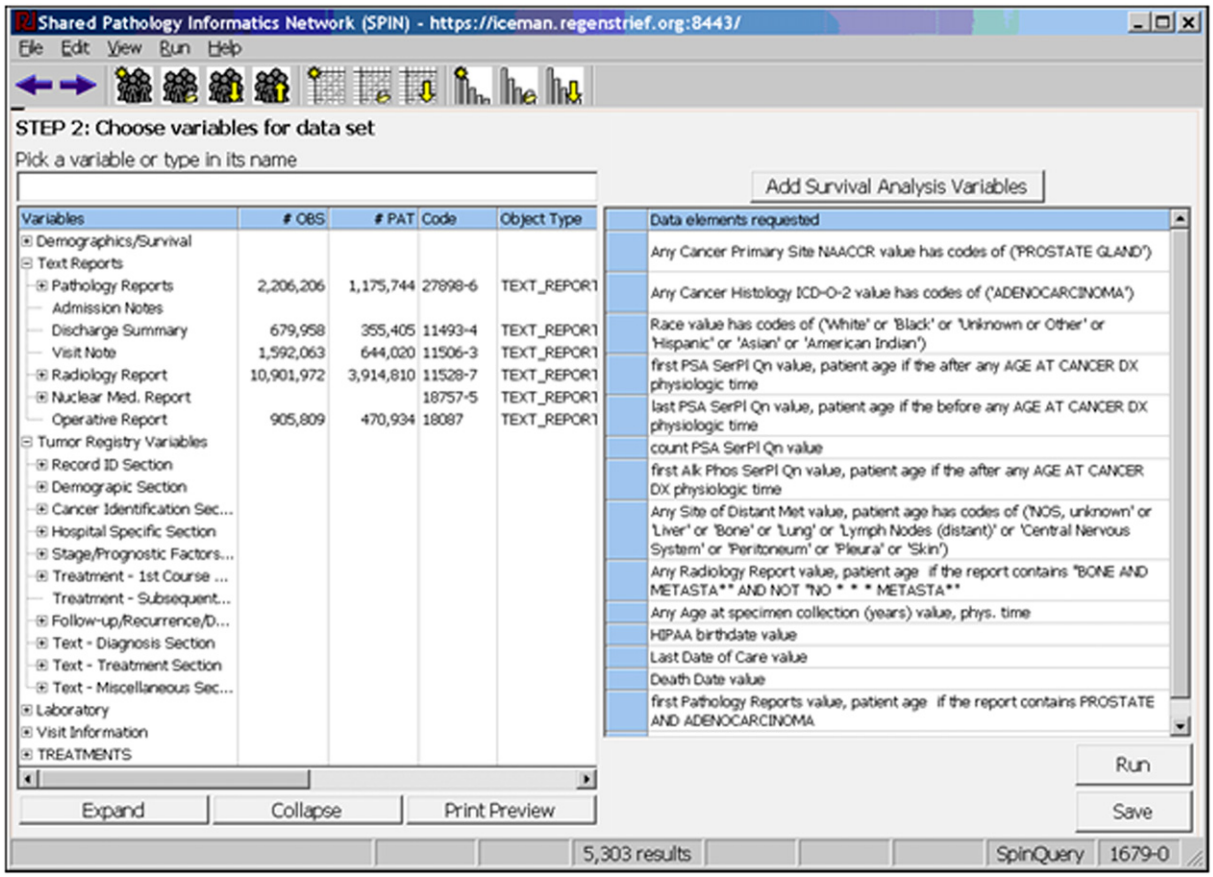


Fig. 7 Screen shot of the Indiana SPIN tool's variable selection step showing variables requested on a cohort of patients with prostate cancer. Note that each variable can yield multiple values; the actual test value, the date performed, the age of the patient when the test was performed, or if it was abnormal. Time restrictions can be placed on a variable, such as the earliest, the latest, and the value in relation to a specific date, or to a specific variable (latest value before "date of diagnosis" as seen in the PSA query). The maximum or average value and number of times the test was performed for that patient can be obtained. Text variables can be searched using logical operators, nearness parameters, and coded formats.

node database and at the same time establishes a corresponding entry in the codebook database to link the two. The 2 associated software tools for autocoding/text parsing and scrubbing can be selectively used or not as desired. In the Indiana implementation, data are fed in real time from participating institutions as HL-7 messages, with autocoding/text parsing and scrubbing steps occurring before population of the database fields. An anonymous identifier is also associated with each case, and there is the equivalent of the above-described codebook to allow for specimen retrieval.

1.4. Searching the SPIN databases

Both the CHIRPS and Indiana groups developed specific software tools ("query tools") with graphical user interfaces to enable searches of the SPIN databases [24,25]. These were designed for initiating a search of the databases that will identify records of pathologic specimens that meet selected criteria and then displaying returned results. These differ significantly as discussed below, but the XML message that represents the query that goes among the nodes in SPIN was the same to everyone. Both are described in greater detail in other publications.

For CHIRPS, the user interface is displayed as a Web page that is hosted by the CHIRPS Web server. It allows the user to choose from a standard set of predefined criteria. The page uses entry fields for desired text or codes and check boxes for specifying selected demographic criteria such as gender and age range (Fig. 4). A simple or "advanced"

Fig. 6 This screen shows the results of a CHIRPS query tool search using the following criteria: search text was "breast ductal carcinoma", any gender, any age at specimen collection, and any collection date. The results included 32,079 cases from 7 active nodes. The top portion gives a graphical and tabular display of the age distribution of the patients. The lower portion gives detailed information on individual cases that were identified. The columns are not labeled, but from the left, they are age at specimen collection, gender, date of specimen collection, and deidentified text of the final diagnosis section of the pathology report. Note that the date of specimen collection field has been redacted in this image.

version of the search page was developed with lesser or greater detail of search criteria. A simple search allows criteria to be set for diagnosis, gender, and age at sample collection. An advanced query allows specification of diagnosis, gender, and age at specimen collection, topology (organ or tissue), and partial date of specimen collection. A user needs only to provide the search criteria, and the query tool turns the query into a formatted message that is sent to the SPIN network. A "code finder" is provided on the page to look up codes to search for.

Results can return in either summarized or itemized formats (or both). Summarized results are presented graphically as histograms of specimen counts, broken down by secondary category such as age or gender (Fig. 5). In the itemized format, a random subset of up to 500 specimens per participating institution per query request has detailed deidentified result sets displayed as one specimen per item (Fig. 6). The full text of the final diagnosis field may be viewed in this format. This format would be available only to registered users with verified Institutional Review Board (IRB) approval.

In contrast to the directed approach of CHIRPS query tool, the Indiana SPIN query tool uses complex logic to take advantage of the additional clinical data available in the Indiana system and to allow statistical analyses to be performed [25]. The query tool is similarly accessed via the Internet. To pose a query, a user first specifies the data elements of interest to define a set of patients to be included in the data set ("cohort definition"). The available variables are displayed for the user to select, which number over 4000, including tumor registry variables and relevant laboratory tests, radiology reports, and other information in addition to the pathology report elements (Fig. 7). For each variable selected, criteria are then defined by the user, which may be numeric, text, or codes. As many variables as desired can be selected, and these can be logically grouped to more narrowly define a desired patient set. A similarly performed second step defines which variables to obtain and display from the patient cohort. The third step allows the user to select data and define statistical tests to be performed. The open-source statistical language R is used for performing the statistical analyses. Similar to the CHIRPS tool, summary data only are returned unless a user has been registered as having IRB approval for their study, in which case the deidentified patient level data are provided (Fig. 8).

Using either tool, the aggregation of specimens from multiple institutions allows for identification of significant numbers of relatively rare neoplasms and a finer level of specification for common neoplasms. For example, if one were interested in studying early versus late onset prostate cancer, there are many hundreds of specimens falling within defined 40 to 50 and 70 to 80 age decades. The availability of paraffin blocks for identified specimens has been assessed in a formal study involving the SPIN consortium institutions, as reported elsewhere [26]. These included both common and rare neoplasms. Overall, usable blocks could be retrieved for approximately 2/3 to 3/4 of specimens identified from SPIN queries.

We are in the process of completing a demonstration study, to be reported elsewhere, in which a large number of lung cancer specimens of different histologic subtypes were evaluated for EGFR mutations that may be indicative of sensitivity to gefitinib therapy. The accumulation of such a comprehensive set of cases for study was greatly facilitated by use of SPIN.

1.5. Protection of patient identity and related considerations

Protection of patient identity, compliance with federal and local regulations, and accommodation of patient consent were primary considerations in the development of the SPIN structure. These considerations affect multiple aspects of the SPIN implementation. Perhaps, the most basic of these is the configuration of the local SPIN database. As described above, each local SPIN database is constituted as a separate database from the local Pathology IS, and the data contained have been stripped of identifying information and assigned anonymous identifiers. The codebook or analogous system that relates the anonymous identifier that SPIN users see with the actual local specimen identification information is "off-line" on a separate computer behind institutional "firewalls." It is accessible only by designated individuals in the local institution and completely inaccessible to any outside person who might be accessing the SPIN node database. In this context, the local pathology department or institution hosting the SPIN database functions in a manner referred to in some IRB-related settings as an "honest broker." It serves as a neutral intermediary between the patient specimens and the researcher requesting samples, maintaining the anonymity of the patient and not having a vested interest in the research.

A second level of protection occurs at the level of the query tools. Two levels of information access have been specified. For unrestricted access queries, the results obtained are only displayed in summarized form (ie, total numbers, without any individual specimen information). Display of itemized results is limited to users who have undergone a certification process with evidence that they come from bona fide investigators. When results are displayed, the individual sites are not specifically identified. Thus, a user may know which institutions participate in the network but will not know where any given specimen is from. In addition, for rare entities, a criterion exists for a minimum number of at least 10 specimen occurrences in any given grouping for results to be displayed.

The nature of the P2P network also provides a layer of protection. The databases are accessible via the Web but only to other SPIN sites. The messages sent to and from sites on the network use a secure messaging standard and encryption protocol, and passwords are required, providing secure transmission and verifying that the request is from a valid user.

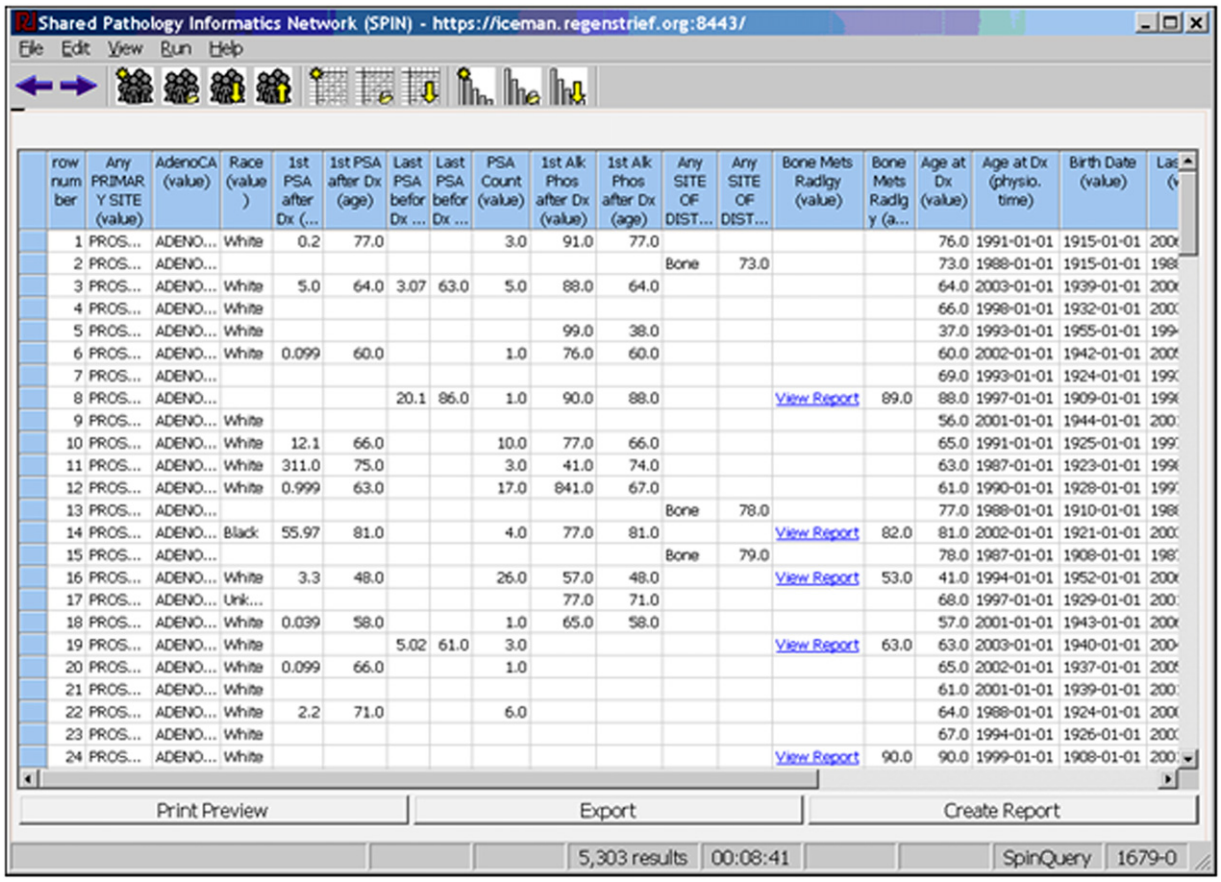


Fig. 8 Screenshot of the Indiana SPIN tool's data file display. To adhere to HIPAA regulations, dates are given only by year, and patient identifiers are removed from text files. The deidentified text files can be viewed from this data file. At this step, individual record summaries can be created or the file can be exported into Excel.

These characteristics that safeguard patient anonymity have made it possible to obtain local IRB approval for each of the participating SPIN institutions. The use of the deidentification software that removes PHI, assigning an anonymous identifier, maintaining the codebook in a safeguarded manner, and providing individual specimen level information only to registered users are key aspects of this. Although IRB approval is a local process, the experience of the SPIN institutions in obtaining approval has established a core set of principles that all the local IRBs have found acceptable. The specimens currently included are archived cases and therefore could be retrieved for research so long as they are not identified as to patient. However, a SPIN implementation could be extended to tissue banks and archives that were obtained under various consent regimens. To be able to respect the various consent regimens, the CHIRPS data model will be augmented to describe a coarse grained hierarchy of patient-specific consent (eg, whether the patient has consented to being recontacted with additional questions). This will require an additional consensus process and a not insignificant change in current workflows (ie, requiring the annotation of the class of consent on specimen acquisition). Although this is an important requirement for further development of SPIN, we note that such annotation of consent is increasingly necessary under the developing consensus around research use of tissue specimens.

1.6. Proposed mechanism for specimen retrieval

As part of the SPIN development work, a proposed mechanism for specimen retrieval has been developed, mentioned in part above. Built into the query tool is a mechanism for users to request specimens identified in their queries. As will be obvious to practicing pathologists, the logistics of actually retrieving a set of requested samples is nontrivial and would require significant support. Issues involved include verification of IRB approvals, retrieval of samples, pathologist verification that the sample is what was desired, obtaining sections, and sending material out. As mentioned above, studies across the participating SPIN institutions have shown availability of paraffin blocks for approximately 70% of cases identified by the query tool.

1.7. Implementing a SPIN network

The SPIN system described here is a prototype network that has not been implemented for public use. Such an

implementation would require further development and resources to support a logistically feasible specimen retrieval mechanism as described above. Implementation of a system to actually retrieve and provide specimens using a query system such as SPIN would probably proceed through a variety of pilot projects. The National Cancer Institute's National Biospecimen Network could include SPIN or some of the SPIN tools under its umbrella. In addition, as discussed below, local networks using the SPIN tools can be implemented. Presently, the software tools developed by the SPIN project are available for use by institutions and groups that would be interested in establishing their own SPIN systems. These are open source, and the hardware requirements are modest [27]. The minimum elements are a server running the query software and the local node databases on servers with associated node software for interacting with the network. Thus, the primary requirement for implementation of a node or a network is having adequate technical expertise. At the present level of development, the setting up of a prototype local SPIN node requires a level of expertise that is likely present at most large academic pathology centers but less likely at smaller centers or community hospitals. However, a knowledgeable individual at one central site can assist other sites in setup, and a local Pathology IS database manager should be sufficiently skilled to maintain an established SPIN database. Perhaps the biggest hurdle is the availability of personnel time, given the relatively lean staffing situation most clinical departments are faced with. For institutions that have already developed a deidentified clinical data repository, the tools developed by the Indiana group could be used to interface a database with a SPIN network.

Although the prototype SPIN structure described in this article was developed for the sharing of data regarding existing routine pathology specimens across independent institutions, in a "public" configuration, the network could be set up to share data among a set of "private" databases (ie, access is limited to specified users). Such a private system has been implemented among 4 hospitals affiliated with the Dana Farber Harvard Cancer Center in Boston. An individual SPIN node can be configured to function as part of both public and private networks simultaneously.

The software developed for SPIN can be expanded or modified for additional or alternative clinical data sets. The involved institutions have entered all specimens with tissue potentially available, and the query tool is relatively generic in its search capabilities. Thus, there is immediate application to other categories of disease besides neoplasia. With modest modifications, the CHIRPS databases and software could be adapted to handle various image types or other types of "cases" besides pathology specimens. Institutions such as the Indiana network of hospitals that represent an RHIO can also participate in the SPIN network via the standard SPIN XML query and response messages and appropriate deidentification techniques.

1.8. Conclusions

Pathologists, as the caretakers of the physical specimens obtained from patients in the course of their medical care, have always had the opportunity to play an important role in the translational research process. There is increasing need for multicenter cooperation in the translational research enterprise, as it relates to studies involving the use of existing patient data and specimens. In this setting, access to existing records across institutions is a critical barrier, and the NCI SPIN program has been a successful pioneer attempt to develop a prototype infrastructure that supports multicenter data and specimen sharing. In doing so, it has emphasized low-cost, decentralization, and local autonomy in the maintenance of the network.

The SPIN program has created a successful prototype high-performance indexing and retrieval system for cancer researchers to identify surgical pathology specimens from conventional clinical archives in participating institutions throughout the United States. The program included development of a scalable and extensible representation of tissue specimens, formulation of a taxonomy for confidentiality and patient consent, and the design and implementation of the necessary software to establish a distributed network architecture for indexing and searching for specimens. This system supports investigator-directed queries with a standard Web browser on the Internet. SPIN uses a P2P distributed architecture for locating specimens that leverages the Internet and promotes local control of data. The program suggests a strategy for annotation of clinical specimens and applications of the system to tissue-based research in other areas of biochemical and molecular medicine.

Acknowledgment

Members of the SPIN include:

Dennis Sunseri, Leslie Ingram-Drake, M-ASCP, Ralph Bowman, David Seligson, MD, and Sarah Dry, MD, from the Department of Pathology and Laboratory Medicine, UCLA Medical Center, David Geffen School of Medicine at UCLA, Los Angeles, CA; William Yong, MD, from the Department of Pathology, Cedars Sinai Medical Center, David Geffen School of Medicine at UCLA, Los Angeles, CA; Robert A. Greenes, PhD, MD, from the Decision Systems Group and Department of Radiology, Brigham & Women's Hospital, Harvard Medical School, Boston, MA; Ahmed Namini, PhD, and Alyssa Porter from the Laboratory of Computer Sciences, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Barrett Goodspeed from the Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; Antonio Perez-Atayde, MD, from the Department of Pathology, Children's Hospital, Harvard Medical School, Boston, MA; Elizabeth Sands, JD, from the Brigham & Women's

Hospital, Harvard Medical School, Boston, MA; Wendy Chapman, PhD, Ghirish Chavan, Rajiv Dhir, MD, John Gilbertson, MD, William Gross, Dilipkumar Gupta, MD, James Harrison, MD, PhD, Kevin Mitchell, MS, Sambit Mohanty, MD PhD, Adita Nemlekar, Yimin Nie, MD, Anil Parwani, MD PhD, Ashokkumar Patel, MD, Melissa Saul, MS, Susan Urda, and Sharon Winters, MS, from the Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA; Thomas Ulbright, MD, from the Indiana University School of Medicine, Indianapolis, IN; Greg Abernathy, MD, and Tamara Dugan from the Regenstrief Institute, Indianapolis, IN; Paul Dexter, MD, is from the Indiana University School of Medicine and from Regenstrief Institute.

References

[1] McDonald CJ, Overhage JM, Barnes M, et al. The Indiana network for patient care: a working local health information infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood) 2005;24:1214-20.

[2] Protecting personal health information in research: understanding the HIPAA privacy rule (NIH publication #03-5388). Bethesda, MD: Department of Health and Human Services, 2003.

[3] Gunn PP, Fremont AM, Bottrell M, Shugarman LR, Galegher J, Bikson T. The Health Insurance Portability and Accountability Act privacy rule: a practical guide for researchers. Med Care 2004;42:321-7.

[4] Berman JJ. Confidentiality issues for medical data miners. Artif Intell Med 2002;26:25-36.

[5] Berman JJ. Pathology data integration with eXtensible Markup Language. HUM PATHOL 2005;36:139-45.

[6] Berman JJ, Bhatia K. Biomedical data integration: using XML to link clinical and research data sets. Expert Rev Mol Diagn 2005;5:329-36.

[7] Mamlin BW, Schadow G, Overhage JM. Open-source toolkit for simple XML annotation. AMIA Annu Symp Proc. 2003. p. 925.

[8] Schadow G, Russler DC, Mead CN, McDonald CJ. Integrating medical information and knowledge in the HL7 RIM. Proc AMIA Symp. 2000. p. 764-8.

[9] McDonald CJ, Schadow G, Suico J, Overhage JM. Data standards in health care. Ann Emerg Med 2001;38:303-11.

[10] Beckwith BA, Mahaadevan R, Balis UJ, Kuo F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 2006;6:12.

[11] Berman JJ. Concept-match medical data scrubbing. How pathology text can be used in research. Arch Pathol Lab Med 2003;127:680-6.

[12] Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121:176-86.

[13] Thomas SM, Mamlin B, Schadow G, McDonald C. A successful technique for removing names in pathology reports using an augmented search and replace method. Proc AMIA Symp. 2002. p. 777-81.

[14] Jeffrey E, Friedl F. Mastering regular expressions. 2nd ed. Sebastopol (CA): O'Reilly Media, Inc; 2002.

[15] Berman JJ. Doublet method for very fast autocoding. BMC Med Inform Decis Mak 2004;4:16.

[16] Berman JJ. Automatic extraction of candidate nomenclature terms using the doublet method. BMC Med Inform Decis Mak 2005;5:35.

[17] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11:392-402 [Epub 2004 Jun 7].

[18] Huang Y, Lowe HJ, Klein D, Cucina RJ. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. J Am Med Inform Assoc 2005;12:275-85 [Epub 2005 Jan 31].

[19] Mitchell KJ, Becich MJ, Berman JJ, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports. Medinfo 2004;11(Pt1):663-7.

[20] Mitchell KJ, Crowley RS, Gupta D, Gilbertson J. A knowledge-based approach to information extraction from surgical pathology reports. AMIA Annu Symp Proc. 2003;937.

[21] Schadow G, McDonald CJ. Extracting structured information from free text pathology reports. AMIA Annu Symp Proc. 2003. p. 584-8.

[22] Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo 2004;11(Pt1):565-72.

[23] Namini AH, Berkowicz DA, Kohane IS, Chueh H. A submission model for use in the indexing, searching, and retrieval of distributed pathology case and tissue specimens. Medinfo 2004;11(Pt2):1264-7.

[24] Holzbach AM, Chueh H, Porter AJ, Kohane IS, Berkowicz D. A query engine for distributed medical databases. Medinfo 2004;1519.

[25] McDonald C, Dexter P, Schadow G, et al. SPIN query tools for de-identified research on a humongous database. Proc AMIAA Symp 2005;515-9.

[26] Patel AA, Gupta D, Seligson D, Hattab EM, Balis UJ, Ulbright TM, et al, and the SPIN. Availability and quality of paraffin blocks identified in pathology archives: a multi-institutional study by the Shared Pathology Informatics Network (SPIN). BMC Cancer 2007 [in press].

[27] McDonald CJ, Schadow G, Barnes M, et al. Open source software in medical informaticswhy, how and what. Int J Med Inform 2003;69:175-84.

[28] Fact sheet: UMLS metathesaurus. Bethesda, MD: National Library of Medicine; 2005.