THE LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS

An Intramural Research Division of the U.S. National Library of Medicine

A Report to the Board of Scientific Counselors April 2011

Moving, Merging, Managing, and Mining Clinical Data for Care and Research

Clement J. McDonald, M.D., Scientific Director
Swapna Abhyankar, M.D., Postdoctoral Clinical Informatics Fellow

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Table of Contents

List of Figures

List of Tables

Moving, Merging, Managing and Mining Clinical Data for Care and Research

1 Collaborations with:

- Academic Model Providing Access to Healthcare (AMPATH) Partnership project (Indiana University/Kenya)
- American Clinical Laboratory Association (38 member laboratories)
- Centers for Disease Control and Prevention (CDC)
- Centers for Medicare and Medicaid Services (CMS)
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)
- Health Resources and Services Administration (HRSA)
- Indiana Health Information Exchange (IHIE)
- Indiana Network for Patient Care (INPC)
- Intermountain Health Care
- Massachusetts Institute of Technology (MIT)
- National Center for Biotechnology Information (NCBI)
- National Center for Research Resources (NCRR)
- National Human Genome Research Institute (NHGRI)
- National Institute of Biomedical Imaging and Bioengineering (NIBIB)
- Partners HealthCare of Boston
- Regenstrief Institute
- RTI International
- Suburban Hospital

2 General background and history

The work I present today represents the continuation of my long-time focus and interest in the computerization of clinical data, which was often described in the past as electronic medical records research. My Regenstrief Institute colleagues and I created one of the first EHRs in 1972. We showed that computer-generated reminders to physicians could improve care process in randomized clinical trials starting in 1976.[1] We built a physician order entry work station in 1986[2] and showed its value in the only randomized trial of such a system.[3] Physicians were actually happy with this system once we tuned it to their needs. Interestingly, what they loved the most was not the computer, but a report generated by the computer that summarized everything about each patient on one page, in small font, that would fit in their white coat pocket.

For years, researchers used the Regenstrief Institute (RI) clinical repository to support various research projects, especially to estimate patient numbers and characteristics available for research studies and as grist for epidemiological studies – one epidemiology study found the first

evidence for a relationship between macrolide (erythromycin) use in newborns and pyloric stenosis.[4] The Indiana University School of Medicine's Dean of Research estimated that 80% of their IRB applications used our clinical database at some stage of their studies. We created the first, and now probably the largest, health information exchange – it now includes 30+ hospitals and carries 3 billion separate clinical results.[5]

In the 1970's and 1980's, much of a patient's clinical data was stored in computers but isolated into separate cubby holes. The laboratory system carried laboratory data, the administrative system carried pharmacy data and diagnoses-related groups (DRGs), and so on – all in separate silo systems. A lot of what we think of as the content for the patient's electronic medical record was "there" in some electronic form, but it was a humpty dumpty – born as separate pieces and not put together. Back then, only insanely difficult and unreliable methods were available for pulling data into medical record systems from ancillary systems. The effort was neither sustainable in our organization nor replicable in others.

So, it was obvious that efficient and reliable mechanisms for moving data from one clinical system to another had to be created before many organizations could create useful clinical data systems. This data had to come in a format that would allow us to merge data from multiple sources into the medical record for one patient.

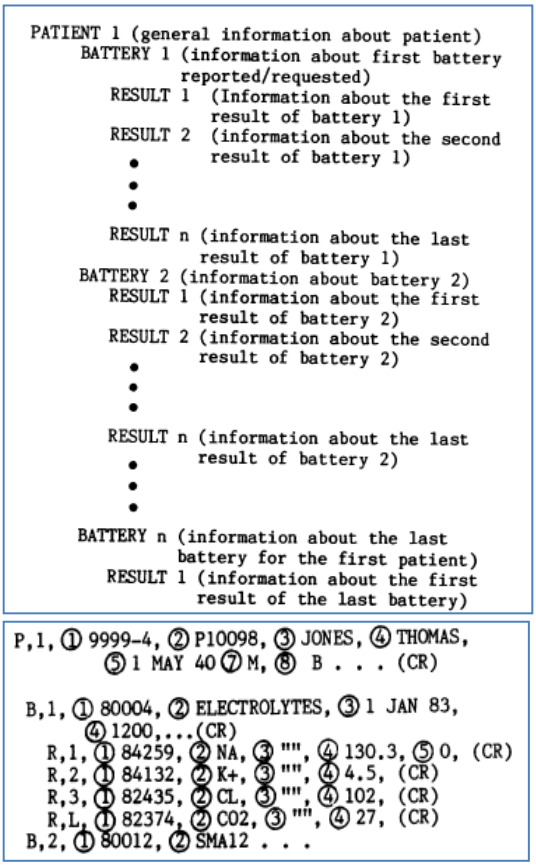


Figure 1. Logical structure (top) and example (bottom) of a laboratory result message, taken verbatim from McDonald CJ, Wiederhold G, Simborg DW, Hammond E, Jelovsek FR, Schneider K. A Discussion of the Draft Proposal for Data Exchange Standards for Clinical Laboratory Results. Proc Annu Symp Comput Appl Med Care 1984; 406-13. PMCID: PMC2578513

In 1984 we organized a meeting at the Symposium on Computer Applications in Medical Care (SCAMC) [6] to discuss this need with the informatics community, and proposed such a universal format (See Figure 1). It was a relatively simple format that included a record that said things about the message, the patient who was the subject of the message, and the order (or report header) for a set of related observations and the individual observations. The proposal defined the fields required by each of these records. The observation record included a field for identifying the observation (question or test), its data type (e.g. whether the value of the

observation would be numeric, text or multiples choice codes), as well as fields for recording the units of measure, reference range, physiologic time of the value, and so on.

That specification came to life as a standard in ASTM 1238-88 in 1988, and as the HL7 V2.1 observation message in that same time frame. By 1993, most hospital and ancillary service system vendors supported it, and today every hospital and large clinic in the U.S. does, as do hospitals, laboratories and clinics in approximately thirty other countries. We estimate that more than 30 billion HL7 observation messages are sent in the U.S. alone each year.

Within hospitals, HL7 messaging was a relatively quick success because the hospital could require that the same set of codes for identifying observations, drugs, and diets be used in all of the systems that processed such data. HL7 has many fields (slots) that require codes that both the sending and receiving systems can understand (Figure 2). Every organization invented their own unique sets of codes for these fields, and no two organizations could understand each other's codes without considerable mapping work. In effect, each organization invented its own coding language and we had Babel all over again.

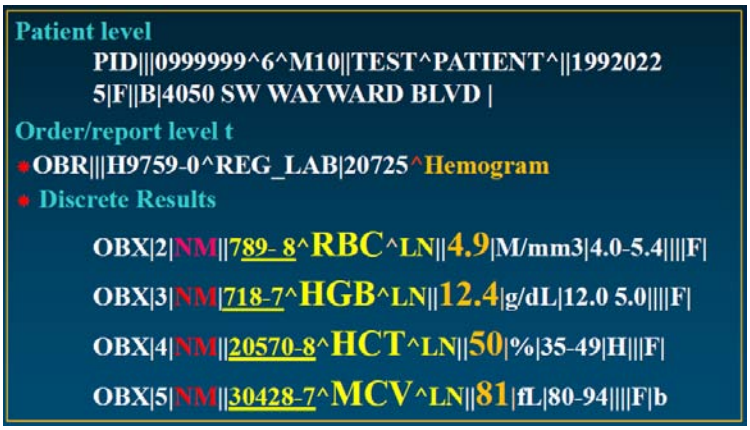


Figure 2. Example HL7 message (color-coded excerpt).

Over time, patients receive care and testing from many independent organizations (practitioners, hospitals, pharmacies, laboratories, public health clinics) and each such organization keeps the data it produces locally. Moving and merging this data is necessary to provide a full picture of the patient for care and/or research. Communication and data sharing among organizations requires that all of the communicants use the same standard codes for observations, drug names, laboratory tests, units of measure, organism names (for culture results), problems and other items. Each of these categories is associated with different challenges, and priorities. I have focused mostly on standards for observation codes because none existed, and without them (or heavy mapping work), outside observations received electronically are "gibberish" that cannot be properly filed, included in flow sheets, or used in clinical decision support or research.

So during my tenure with the Regenstrief Institute, I began an effort to develop a database of standard observation codes and names — called Logical Observation Identifiers Names and Codes (LOINC®)[7] to address this gap. Since coming to serve as the Director of the Lister Hill

National Center for Biomedical Communications at the National Library of Medicine, I have also worked on projects related to other coding standards, including: RxNorm[8,9] (a coding system for medications), SNOMED CT®[10] (Systemized Nomenclature of Medicine—Clinical Terms – a coding system for many kinds of concepts including problem lists) and UCUM (a standard for units of measure). I will describe this work in this report.

The National Library of Medicine (NLM) has had a special role in the development and support of clinical vocabulary standards for more than a decade. It has provided crucial contract, or internal development support for the three keystone vocabulary standards: LOINC, RxNorm, and SNOMED-CT. In 2004, the HHS Secretary designated NLM as the central coordinating body within HHS for Patient Medical Record Information (PMRI) technology standards, based on the recommendation of the National Committee on Vital and Health Statistics (NCVHS) and the federal Consolidated Health Informatics (CHI) Council.[11] These code systems are already widely-used both nationally and internationally, and the first three – LOINC, SNOMED-CT, and RxNorm – are identified as "minimum standard" code sets by federal certification criteria adopted in 2010.[12]

The work that I will describe also supports many federal goals for increasing the level of clinical data automation,[13] and NLM's strategic goals, including Recommendation 3.2 "promote development of Next Generation electronic health records to facilitate patient-centric care, clinical research, and public health" to accomplish Goal 3 – "Integrated Biomedical, Clinical, and Public Health Information Systems that Promote Scientific Discovery and Speed the Translation of Research into Practice."[14] Much of the work is also directly responsive to the recently approved final report of the NLM Board of Regents Working Group on Health Data Standards, including its recommendation to "provide additional tools and services that help vendors and user sites to incorporate standards where they will have a positive impact."[15]

Related Publications

Simonaitis L, McDonald CJ. Using National Drug Codes and Drug Knowledge Bases to Organize Prescription Records from Multiple Sources. Am J Health Syst Pharm. 2009 Oct 1;66(19):1743-53. PMID19767382 : PMC2965522

Fung KW, McDonald CJ, Bray BE. RxTerms - a drug interface terminology derived from RxNorm. AMIA Annu Symp Proc. 2008 Nov 6: 227-31.

3 LOINC database of universal identifiers for observations, reports and orders

(Vreeman D, Huff S, McDonald CJ, and LOINC committee)

3.1 Overview and general progress

LOINC is a database of codes and names for observations, orders, and panels or groups of the observations or orders. Picture a test result or clinical observation such as those shown in Table 1. LOINC's principal interest is assigning codes and names for the variables shown in the first column of Table 1, not the value or result of the observation (in the second column of Table 1). Example LOINC codes are given in Table 2.

Variable	Value	Units of measure
Glucose-serum	120	mg/dL
Glasgow Coma score – eye opening	3-Opens eyes spontaneously	

Table 1. Example test result values and clinical observations.

LOINC code	LOINC long common name
2345-7	Glucose [Mass/volume] in Serum or Plasma
9267-6	Glasgow coma score eye opening
53836-3	ABCD1 gene mutation analysis in Blood or Tissue by Molecular genetics

Table 2. Example LOINC codes.

The Regenstrief Institute and LOINC committee introduced the LOINC database in 1995 with 6,000 terms focused initially on laboratory observations. In 2006 when I arrived at NLM, it carried nearly 47,000 terms. The LOINC database now carries records for more than 60,000 terms including a broad spectrum of laboratory tests, dictated reports, clinical measurements, specialty terms, survey instruments as well as panels (collections) of the aforementioned items.

The LOINC database is not just a listing of LOINC names and codes; it also includes units of measure (for numeric observations), example answer lists (for multiple choice questions), descriptions of the observations, synonyms, classes and hierarchies. LOINC has been adopted by many large institutions, including: Partners HealthCare of Boston, Intermountain Health Care, health information exchanges such as the Indiana Network for Patient Care (INPC), health insurance companies (e.g. United Health Care) and most large clinical laboratories (Quest, LabCorp, ARUP, NMS, Mayo Clinic's Mayo Medical Laboratories. We like Mayo's Web site (Figure 3): https://www.mayomedicallaboratories.com/test-catalog/appendix/loinc-codes.html.

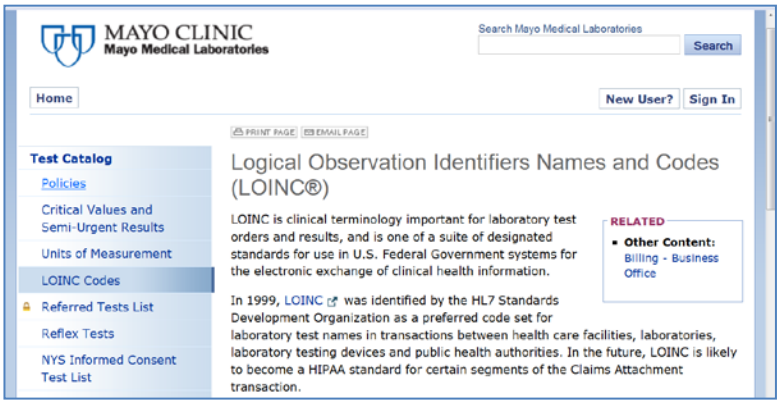


Figure 3. Mayo's LOINC Screen.

Instrument and test kit vendors are now beginning to ask for LOINC codes and to report the codes that correspond to the results they can produce. If they included the LOINC code for each of the tests that are FDA-approved, laboratories would have a much easier time linking LOINC codes to their laboratory tests. The members of an international organization of laboratory instrument vendors – the IVD Industry Connectivity Consortium (IICC) – are planning to map the internal codes of all of their instruments to LOINC codes and then provide them to their customers, perhaps through a common database.

LOINC provides a database, a downloadable desktop browsing and mapping tool called RELMA, and a web site without charge. Anyone can download the LOINC database; users have to register to download the RELMA browser/mapper. In February 2011, LOINC exceeded 10,130 registered users from 140 countries, which are colored green in Figure 4.

Figure 4. LOINC has more than 10,000 registered users in 140 countries (shown in green on the map).

In June 2010, LOINC launched a new web-based search engine (http://search.loinc.org/) which is very accessible, but has fewer features than the desktop version. Both of these systems now use Lucene as their search engine, which is much faster and more capable than the simple database indexing which we used before.

For LOINC to facilitate the transfer of test results and other variables among systems, local systems will initially have to map their local terms to LOINC – a one-time effort. RELMA – the Regenstrief LOINC Mapping Assistant – provides tools to assist this mapping process. Mappers can submit a master file of their tests codes and names along with other attributes about their tests (e.g. the units of measure) to RELMA, and then look up the appropriate LOINC code one at a time. Or they can run a large set (millions) of their outbound HL7 messages to RELMA and it

will create the master file needed for mapping. In these cases, RELMA automatically adds units of measure and example results (taken from the HL7 messages) to the generated master file.

Finally, the RELMA browser program also provides an auto-mapper for laboratory tests. This program uses all of the information about the local test, plus special rules and frequency statistics, to find the best 5-10 mapping candidates for a given local test name. In early studies, the right code was the first choice of the auto-mapper 50% of the time and the right choice was in the 5 highest-ranked choices 75% of the time.

3.2 International use

LOINC has been adopted as the national standard in Canada, Australia, the Netherlands and Germany, adopted widely in China, Brazil, France, Mexico, Spain, Singapore and Switzerland, and adopted less intensely in many other countries. LOINC is also expected to be proposed as the lab result standard in Singapore's national EHR project by the Ministry of Health Holdings.

International groups have produced 12 translations of some or all of the LOINC database into 9 languages (Table 3). LOINC has three French translations (from France, the Canadian government and Switzerland), and three Spanish translations (from Argentina, Spain and Switzerland). The Chinese have published their translations as an 1800 page book that we will show you.

Language	Translations
English	Glucose
Estonian (Estonia)	Glükoos
French (3: Canada, France, Switzerland)	Glucose
German (2: Germany, Switzerland)	Glucose
Greek (Greece)	Γλυκόζη
Italian (2: Italy, Switzerland)	Glucosio
Korean (Republic of Korea)	포도당
Portuguese (Brazil)	Glicose
Simplified Chinese (China)	葡萄糖
Spanish (3: Argentina, Spain, Switzerland)	Glucosa

Table 3. Languages represented in LOINC -- with their translations of glucose (LOINC part code LP14635-4) given as an example.

In the works are translations for Catalan, Dutch, and Russian translation. Regenstrief provides a program to facilitate translation. Users of this program need only translate the unique parts of all of the LOINC terms of interest and the program generates translations of all of the terms that contain these parts.

3.3 Miscellaneous efforts

We are organizing an effort by a third party to map some of the atomic parts of LOINC terms to SNOMED CT, and to map the LOINC laboratory and radiology terms to CPT codes, so that we could use those mappings to ease the work laboratories have to do to map their internal codes to LOINC. We have also been working with the American Clinical Laboratory Association

(ACLA) to clarify the content of many of the most frequently ordered *test panels* and mechanisms to represent them in LOINC. This has had the dual advantage that the major commercial laboratories help us define the standard approach, and then often change their internal systems to conform to it.

LOINC Part Number	LOINC Part Name	Description
LP17806-8	ABO group	The ABO blood group system is the most important blood type system (or blood group system) in human blood transfusion. The associated anti-A antibodies and anti-B antibodies are usually IgM antibodies, which are typically produced in the first years of life by sensitization to environmental substances such as food, bacteria and viruses. ABO blood types are also present in some animals, for example cows and sheep, and apes such as chimpanzees, bonobos, and gorillas. *Source: Wikipedia (edited by Regenstrief Institute)*
LP64576-9	Acid citrate dextrose	Acid Citrate Dextrose Solution (sometimes called Anticoagulant Citrate Dextrose Solution) is a solution of citric acid, sodium citrate and dextrose in water. It is mainly used as an anticoagulant to preserve blood, it is also used during procedures such as plasmapheresis instead of heparin. Two different solutions (Solution A and B) are defined by the United States Pharmacopeia. *Source: Wikipedia (edited by Regenstrief Institute)*
LP18494-2	Activated protein C resistance	Activated protein C resistance (APC) has emerged as the most frequent abnormality in patients with idiopathic thrombosis. Patients found to be heterozygous for APC resistance have a seven fold increased risk for venous thrombosis as compared to the general population. Homozygous individuals have an eight fold increased risk of thrombosis. Familial studies and counseling should be considered for positive patients. The activated form of Factor V (Factor Va) is more slowly degraded by activated protein C. Factor V Leiden mutation (R506Q) is the most common cause of APC resistance. Often measured as the ratio of the aPTT with APC to aPTT without APC. *Source: Regenstrief Institute and National Library of Medicine*
LP14459-9	F2 gene	Coagulation factor II is proteolytically cleaved to form thrombin in the first step of the coagulation cascade which ultimately results in the stemming of blood loss. F2 also plays a role in maintaining vascular integrity during development and postnatal life. Mutations in F2 lead to various forms of thrombosis and dysprothrombinemia. *Source: NLM National Center for CBI Entrez Gene*

Figure 5. Examples of LOINC description text

LOINC includes reference material and text descriptions about much of its content. Some of the test content is created *de novo* by Regenstrief employees, and some is paraphrased from other sources with permission. At present most part components have some description text; however some individual terms have multiple descriptions that are redundant and/or of varying quality. Others may have descriptions that are true, but not relevant to the clinical use of a test for that analyte, e.g. it describes the melting point and molecular weight but not the clinical relevance of lead measured in serum. We have embarked on an effort to edit all of the descriptions for the most commonly reported tests – to eliminate redundancy, and improve the content of skimpy descriptions (Figure 5 shows examples of several descriptions improved by our revision process).

Related Publications

Friedlin J, McDonald CJ. An Evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database. J Am Med Inform Assoc 2010; 17:283-287. PMID:20442145 : PMC2974620.

Lin MC, Vreeman DJ, McDonald CJ, Huff SM. Correctness of Voluntary LOINC Mapping for Laboratory Test in Three Large Institutions. AMIA Annu Symp Proc. 2010; 2010**:** 447–451**.** PMID21347018: PMC3041457

Lin MC, Vreeman DJ, McDonald CJ, Huff SM. A Characterization of Local LOINC Mapping for Laboratory Tests in Three Large Institutions. Methods Inf Med 2010 Aug 20; 46(5). PMID20725694: PMC 3034110.

3.4 LOINC specified as standard vocabulary by federal initiatives and regulations

In 1996, Congress passed the Health Insurance Portability and Accountability Act (HIPAA). HIPAA contains Administrative Simplification provisions that require the establishment of national standards for administrative health data transactions, including the code sets used in such transactions. LOINC has been formally proposed for designation as a HIPAA standard code set for use in claims attachments which will lead to additional expansion of its content. LOINC has been specified as one of the standards for the National Electronic Disease Surveillance System (NEDSS) and Electronic Laboratory Reporting Standards (ELRS) initiatives developed by the Centers for Disease Control and Prevention (CDC).

LOINC has been recommended as a U.S. standard by the National Committee on Vital and Health Statistics and has been adopted as a U.S.-government wide standard by the interagency Consolidated Health Informatics (CHI) Initiative for the electronic exchange of clinical health information. It was also specified as a required standard in U.S. Interoperability Specifications released by the Healthcare Information Technology Standards Panel and recognized by the Secretary of the U.S. Department of Health. It is now part of the HITECH regulations regarding clinical data automation. Most recently, LOINC was identified as a "minimum standard" vocabulary code set for reporting laboratory test results, by federal certification criteria adopted in 2010.[16]

3.5 Standardization of survey instruments and data collection forms

A number of formal and widely used data collection "forms" exist in clinical care, research and administration/management. I will refer to all of these as "survey instruments." The Apgar score, the Glasgow coma score, and the PHQ-9 (a depression score widely used in research and clinical care) are all examples. They all include a series of questions, most of which have pre-defined lists of multiple choice answers which are often associated with "scores" that are tallied to obtain an overall score for the dimension being measured. The PHQ-9 survey instrument attaches a score from 0 to 3 to each answer, and the overall depression score is a sum of the scores from each of the user's responses.

LOINC includes survey instruments in its database, and has created a rich database structure to accommodate them (see Figure 6). This structure includes an association table for nesting question hierarchies and for carrying attributes that are specific to a LOINC variable in the context of a specific form.

It also includes slots for storing the name of the observation (like any other LOINC term), the literal text of the question that the patient must answer, guidance about answering the question, and the validated answer list items with unique identifiers for each. Survey instruments are often copyrighted, so LOINC also provides slots for the copyright and terms of use. With one exception, LOINC has only included survey instruments that permit world-wide use of the terms without charge or special permission. The exception is for the M.D.S 3.0, a survey instrument embedded in a mandated Medicare form that does require permission for use outside of the U.S.

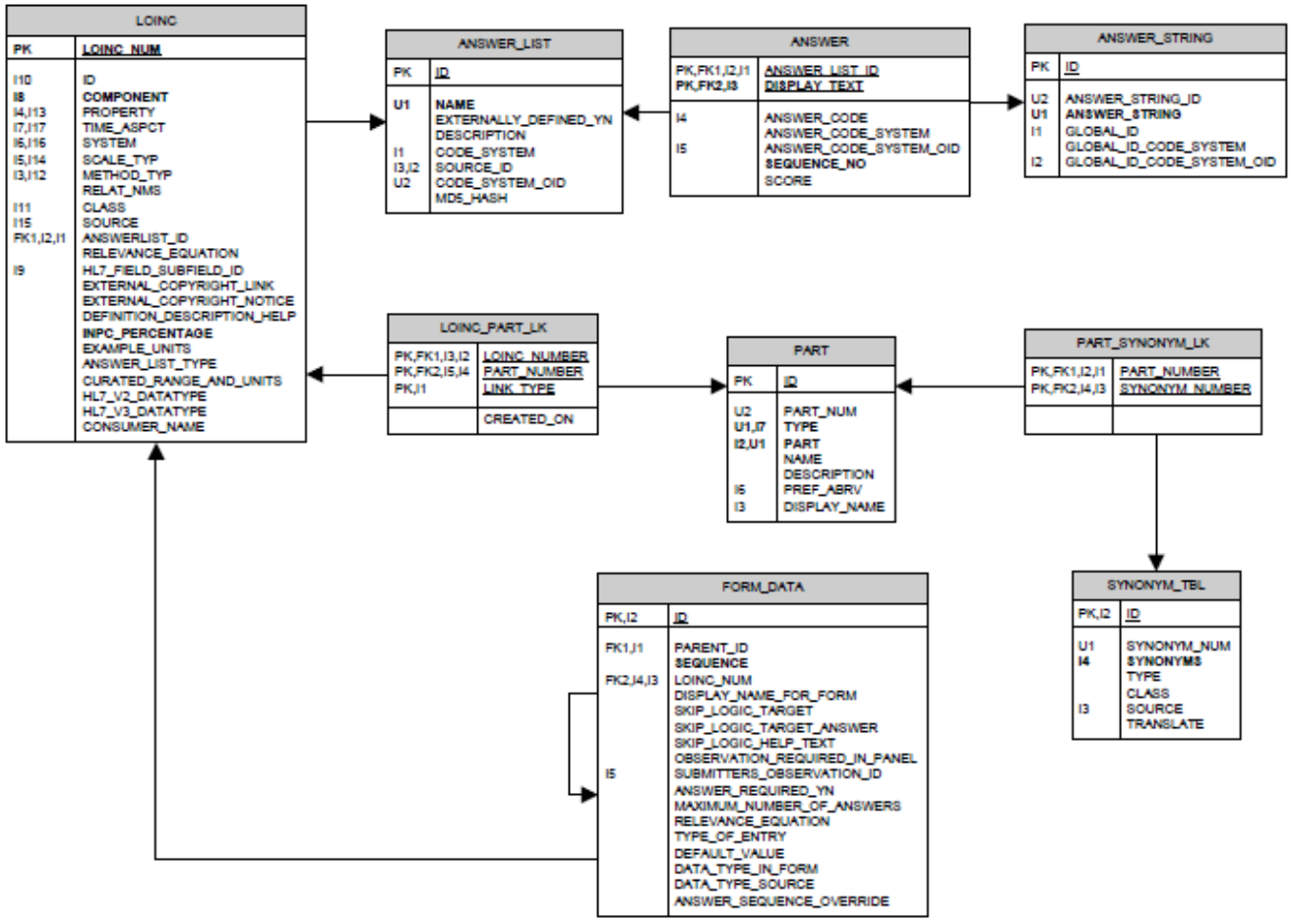


Figure 6. LOINC entity relationship diagram.

During the last four years, LOINC has worked with a number of federal agencies to translate paper and database forms into the standard LOINC structure, for distribution with all of the other

variables in the LOINC database, in order to maximize re-use of variables. The idea is to provide one common catalogue of variables that cross boundaries of clinical care research and administration, to maximize re-use. Thus, if an administrative data form carried a laboratory test result, it would use the same code for that test as the laboratory used; and programs that wanted to generate data input forms would have one common structure with which to work regardless of the kind of data.

3.5.1 Development of LOINC panels for Centers for Medicare and Medicaid Services (CMS) Instruments: CARE, MDS and OASIS

NLM helped CMS implement their Continuity Assessment Record & Evaluation (CARE) project. The CARE instrument was a large survey instrument which included clinical questions, laboratory test results and more traditional survey questions. It was designed for the post acute care payment reform demonstration (PAC-PRD) required under the Deficit Reduction Act of 2005 (S1932.Title V. Sec 5008). Under an MOU with CMS, NLM agreed to help CMS "achieve CHI, LOINC, HL7 and semantics compliance and interoperability." NLM assisted them with design decisions, convinced them to adopt an auto-complete mechanism for data entry and provided the vocabulary for medication (RxTerms) and problem list capture (the latter based on an empiric sample of problem terms from the Regenstrief Institute).

The LOINC group converted all of the items in the CARE survey instrument into a set of more than 330 different LOINC terms organized into six forms targeting the discharge from acute care hospitals and post acute care settings such as nursing homes and home health care. Using this form, and the NLM-contributed vocabularies for drugs and problems, CMS gathered acute care discharge data from roughly 40,000 encounters from multiple hospitals, which collected data with CARE as part of the payment reform demo.[17] We anticipate getting data about the actual strings entered in the problem and medications fields to assess the coverage of the NLM vocabularies.

On request from the HHS Office of the Assistant Secretary for Planning and Evaluation (ASPE), LOINC undertook a similar effort to convert multiple versions of the CMS Outcome Assessment and Information Set (OASIS) form, which is used to evaluate home health care agencies, including version B and five variants of version C. Each of the OASIS forms ranged in size from 100 to more than 150 LOINC terms.[18] LOINC also worked on multiple versions of CMS's Minimum Data Set (M.D.S) for standardized assessment of nursing homes and swing bed providers, including 6 variants of version 2 and the one variant of version 3.[19] All of these forms provide a rich source of long term clinical and functional status data which qualified researchers can now obtain from Medicare.

3.5.2 Large NIH-developed "survey instruments" and LOINC survey packages

3.5.2.1 PhenX – Collaboration with the National Human Genome Research Institute (NHGRI)

The PhenX project funded by the National Human Genome Research Institute (NHGRI) developed a large set of measures across twenty content domains such as "demographics," "anthropometrics," "alcohol tobacco and other substance abuse," "cardiovascular," and "environmental exposures" (https://www.phenxtoolkit.org/). The goal is to provide a set of measures that researchers could incorporate into their data collection forms (rather than inventing their own) and thus enable cross-study comparisons and pooling of data. A PhenX measure consists of narrative with very precise explanations and may include associated figures that illustrate how to take the measure. A PhenX measure is almost never a single question or variable. Most PhenX measures consist of multiple items or questions. Some include one or more full-fledged survey instruments.

This collaboration was a mutually useful one. The LOINC effort to incorporate all of the specific data elements specified on the narrative form revealed variables that were not included in the PhenX variable database and/or were ambiguous. Also, the effort to convert the content of the database records into explicit LOINC terms suggested revisions to the PhenX database to facilitate the conversion, and RTI (NHGRI's contractor) made those revisions.

The effort on the LOINC side required the creation of standardized names for all of the discrete items, the hierarchical relationship among them, the entry of the full PhenX question text, its structured answer list, description text, skip logic, cardinality, citations and when applicable third-party copyrights and terms of use. The work on the domains for "demographics," "anthropometrics," "cardiovascular," and "nutrition and dietary supplement," including all of the accessory content, has been completed. In each domain, different kinds of problems arose and were solved.

One of the new features of the variables in the Nutrition domain was that several of the questions referenced specific graphical images. Because LOINC intends to capture the full meaning of the observation (question), Regenstrief created a mechanism to link a LOINC code to an image file stored on our internet server and modified the RELMA program to display this information when the details for a term are reviewed.

All of the LOINC terms for these five PhenX domains are in the latest public LOINC release (version 2.34 released December 29, 2010). The LOINC release also includes a special export excel spreadsheet format for this content, so users can download PhenX variables from the LOINC web site as one complete package.

3.5.2.2 PROMIS – Collaboration with the PROMIS research team and the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS)

The Patient-Reported Outcomes Measurement Information System (PROMIS®) (http://www.nihpromis.org) is a set of very well validated instruments designed to measure 13 dimensions of patient-reported function such as, anger, depression, fatigue, and pain, all of which were developed by a consortium of experts. The goal was to provide researchers and clinicians a standard way to collect a broad range of functional status measures obtained directly from patients. PROMIS contains a separate set of measures for adults and children written in both English and Spanish.

Each functional dimension can be assessed in two ways. The first method is via computerized adaptive testing (CAT), with results reported on a standard metric. CAT testing is designed to ask the fewest number of questions needed to reach a pre-specified confidence level. Depending on the dimension, the number of questions in the CAT item banks range up to 124. But when the data is collected by CAT, the number of questions asked of any one patient is much smaller. PROMIS also provides a set of short forms for assessing each dimension. The short forms contain a few items – usually 4 to 8 – in a fixed set that can be collected directly on paper forms.

Working closely with the PROMIS team, the LOINC team created LOINC codes for all of the items in the current PROMIS item bank – a total of 660 terms organized into 21 domain item banks. LOINC also included PROMIS's 21 short forms in its database, each as an individual panel. All of the PROMIS items are included in the December 29, 2010, LOINC database release both in the primary distribution format and in the alternate format that packages all of the accessory information in three worksheets.[20]

Related Publication

Vreeman DJ, McDonald CJ, Huff SM. Representing Patient Assessments in LOINC. AMIA Annu Symp Proc. 2010; 2010**:** 832–836. PMID21347095: PMC3041404

3.5.3 Tools and documentation to facilitate the integration of LOINC into clinical settings.

3.5.3.1 Mapper's guide to the top 300 laboratory test orders

This value set was developed as a collaboration between the Lister Hill Center, the Regenstrief Institute and participating organizations that are listed below. It was cited by the Healthcare Information Technology Standards Panel (HITSP) C80 Clinical Document and Messaging Terminology Construct in Table 2-97 "Laboratory Order Value Set" as the minimum starter set of LOINC codes for laboratory test ordering.[21]

We produced the initial list from data about test ordering frequency provided by a number of sources who mapped their test orders and results to LOINC. Specifically, we obtained frequency distributions from 1) a sample of 10 million test orders from the Indiana Network for Patient Care (INPC), 2) a sample of 30 million test *results* from United Health Care (UHC), 3) 1.5 million test orders from five N.E. U.S. hospitals, 4) 200 thousand tests orders from a S.E. U.S. hospital, and 5) a list of 200 tests gathered by survey from an internal medicine project run by Stasia Kahn, M.D. (Chicago). The first four of these data sets had LOINC codes attached; we hand mapped LOINC codes to the tests suggested in Dr. Kahn's list. The INPC, N.E. U.S. and S.E. U.S. lists were weighted toward hospital laboratories (including inpatient and outpatient). The UHC results come from a national database and include only outpatient *results* produced at commercial laboratories. We used the UHC result file to supplement the LOINC *order* codes from the other sources by including entries that we knew from clinical experience were ordered as separate tests and were not already within the high frequency test orders from INPC. We started with the Indiana source and only added terms from other sources to the 99-percentile set of codes from INPC if they were not already included. The original set comprised about 300 tests and covered close to 99% of each of the sources' laboratory volumes.

We presented an early version of the common orders list at the HITSP face to face meeting in Silver Spring, Maryland for HITSP review on November 5th 2009, after which it was sent out for public comment. We received a number of useful comments about this list. To resolve these comments, we had a number of very productive and useful meetings with the American Clinical Laboratory Association (ACLA) regarding the usage and definitions of some of these tests within the industry. As a result of those meetings and their expertise, we have replaced some items that were misconceptualized, reviewed and clarified the definitions of many panels, and added LOINC order codes that were frequent among the ACLA labs, but were not among the common tests from the initial sources. We have removed a handful of tests that were incorrectly included as orders. Some of the tests that were held back still need to be reviewed by ACLA. The set that was released on April 7, 2010 includes 288 codes.[22] This list can be obtained from: http://loinc.org/usage.

3.5.3.2 Mapper's guide to the top 2000 laboratory test results/observations

In collaboration with the Regenstrief Institute, we created this list for laboratories, practices, researchers, and others who wish to map their laboratory test codes to universal LOINC codes. The "Mapper's Guide" provides a starter, target set of LOINC codes against which to map local test codes, as well as guidance about which LOINC codes to choose for which purpose. This list contains approximately 2,000 LOINC codes that represent more than 98% of the test volume carried by three large organizations that mapped all of their tests to the LOINC codes.

Each row in the spreadsheet carries information about one laboratory test observation including its LOINC code and name, example units of measure expressed in UCUM[23] units (http://unitsofmeasure.org/), its class when applicable (e.g. chemistry, hematology, etc.), its relative frequency, and in some cases, guidance about when to choose that test code.

You will soon be able to obtain the Mapper's Guide in two formats from the LOINC web site (http://loinc.org/usage) as:

1) A PDF format for manual review, and
2) An Excel spreadsheet, which can be exported into a CSV file.

The Mapper's Guide shows the most frequent LOINC codes, sorted in descending order by adjusted class, specimen name and long common name. It may include mapping guidance at the level of:

1) The adjusted class (e.g. Figure 7),
2) The specimen,
3) Groups of related tests (e.g. Figure 8), and
4) The individual test.

Three large organizations provided data for this effort: the Indiana Network for Patient Care (INPC),[24, 25] Partners Healthcare System of Boston,[26] and United Healthcare (UHC).[27] Each of these sources maintains a large clinical repository and applies LOINC codes to their laboratory test observations. Two of these sources apply LOINC codes to all of their non laboratory observations as well. Each source provided the LOINC names, the counts and percent that each test represented of their total test volume. We received no patient level data of any kind. The tests included in the guide represent more than 99.5% of the test volume from each of the three sources. However, most of them included one or more terms which were variations on a non-informative name, such as "Miscellaneous send out." Because we want to discourage such naming practices, we did not include them in this mapper's guide; therefore the coverage of the terms in the mapper's guide covers only about 98% of the test volumes from these sources.

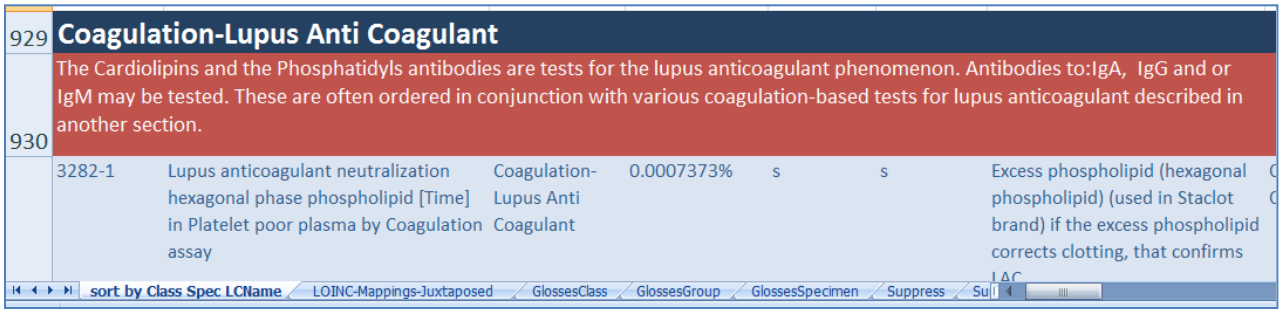
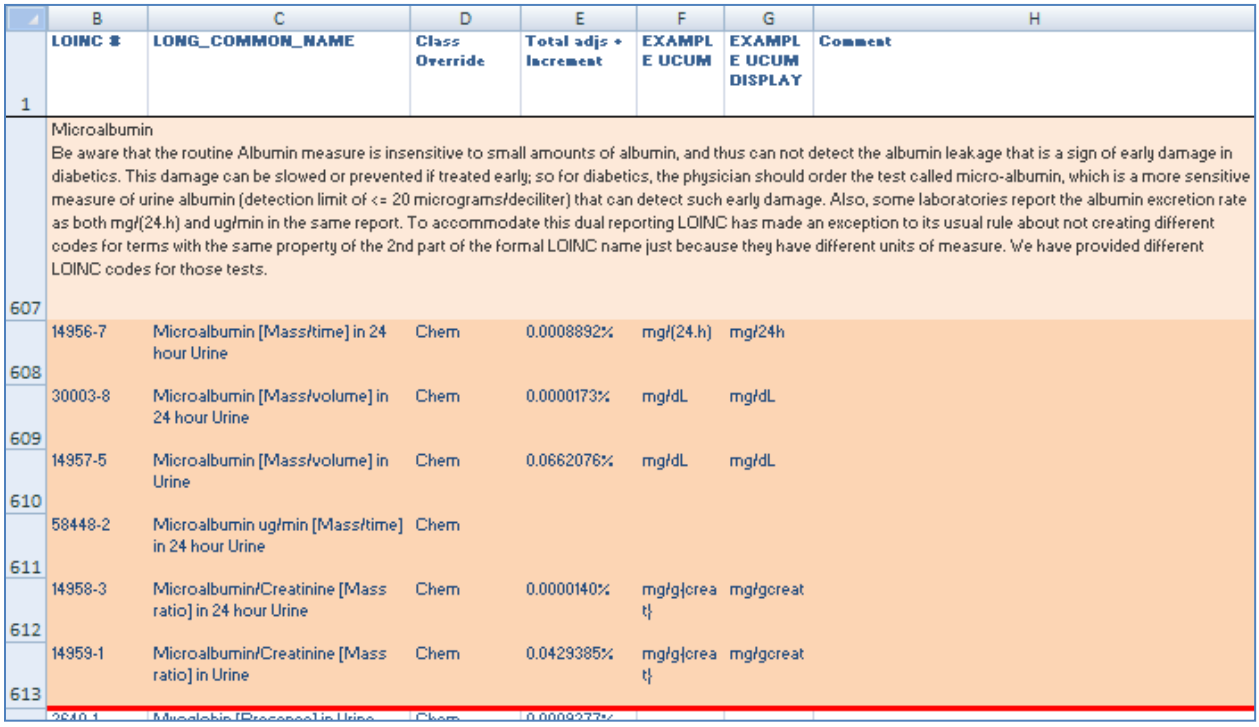


Figure 7. Example of guidance regarding a given class of LOINC terms.

Collectively the three sources cover a broad national spectrum of hospital, clinic, and office practice testing, and all three of the sources represented collections of data taken from many different laboratories. The statistic we report is the un-weighted average of the statistics provided by the three sources, which together represent a sample of more than 600 million test results.



	B	C	D	E	F	G	H
1	LOINC #	LONG_COMMON_NAME	Class Override	Total adjs + Increment	EXAMPLE UCUM	EXAMPLE UCUM DISPLAY	Comment
607	Microalbumin Be aware that the routine Albumin measure is insensitive to small amounts of albumin, and thus can not detect the albumin leakage that is a sign of early damage in diabetics. This damage can be slowed or prevented if treated early; so for diabetics, the physician should order the test called micro-albumin, which is a more sensitive measure of urine albumin (detection limit of <= 20 micrograms/deciliter) that can detect such early damage. Also, some laboratories report the albumin excretion rate as both mg/(24.h) and ug/min in the same report. To accommodate this dual reporting LOINC has made an exception to its usual rule about not creating different codes for terms with the same property of the 2nd part of the formal LOINC name just because they have different units of measure. We have provided different LOINC codes for those tests.						
608	14956-7	Microalbumin [Mass/time] in 24 hour Urine	Chem	0.0008892%	mg/(24.h)	mg/24h	
609	30003-8	Microalbumin [Mass/volume] in 24 hour Urine	Chem	0.0000173%	mg/dL	mg/dL	
610	14957-5	Microalbumin [Mass/volume] in Urine	Chem	0.0662076%	mg/dL	mg/dL	
611	58448-2	Microalbumin ug/min [Mass/time] in 24 hour Urine	Chem				
612	14958-3	Microalbumin/Creatinine [Mass ratio] in 24 hour Urine	Chem	0.0000140%	mg/g{creat}	mg/gcreat	
613	14959-1	Microalbumin/Creatinine [Mass ratio] in Urine	Chem	0.0429385%	mg/g{creat}	mg/gcreat	

Figure 8. Example of guidance at the group level.

3.5.3.3 Newborn Screening Coding and Terminology Guide: Collaboration with HRSA, CDC and NICHD

(Abhyankar S, Zuckerman AE, Lloyd-Puryear M, Goodwin RM, McDonald CJ)

State law requires that all newborns be tested for a set of conditions whose effects can be prevented or reduced if identified early. The exact set of conditions that are tested vary a bit from state to state but almost all states test for the 30 core conditions recommended by the HHS Secretary's Advisory Committee on Heritable Disorders in Newborns and Children.[28]

Newborn screening results (NBS) are typically reported as narrative text, and delivered by mail, fax or telephone. Hence reports can be delayed or lost, putting the infant at risk for dangerous delays in needed treatment. Furthermore, though most of the test measurements are quantitative, laboratories usually report them qualitatively as normal or abnormal, and often omit the cut off values for deciding normal versus abnormal. Therefore, little data is available for understanding the wide variation in false positive rate across states, or for performing quality improvement.

Furthermore, without standardized quantitative data, researchers cannot pool cases from across many states to get a large enough sample size to assess outcomes or the effectiveness of treatment for the many rare conditions that NBS seeks to identify.

The American Health Information Community (AHIC) Personalized Health Care Work Group was formed in 2006, and it "prioritized information exchange for newborn screening test results for standards harmonization and development of interoperability specifications."[29] The goal was to develop a strict electronic NBS report standard that would obviate these problems. The committee, which included members from NLM, CDC, the HHS Office of the Office of the Secretary Personalized Health Care Initiative, the Mayo Clinic, the NIH Eunice Kennedy Shriver National Institute of Child Health and Human Development and private sector stakeholders, developed a list of all of the measures including ratios and other formulaic variations that might be reported by any NBS laboratory and designated NLM to be the lead on the development and maintenance of code sets for these measures. NLM developed guidelines for producing structured HL7 ORU messages to carry NBS results and a complete set of LOINC codes for the variables, based on input gathered at many large meetings and conference calls. NLM also developed the codes for the variables usually recorded on the blood spot card by the birth site, and a novel proposal for recording hemoglobinopathy results.

The HRSA/NLM Newborn Screening Results Messaging Guidance[30] includes an annotated example HL7 message and the newborn screening LOINC AHIC panel – a comprehensive table of variables for the lab tests used for screening and the birth information used for interpreting test results, with their hierarchy of relationships, codes, answer lists and units of measures. Each state can choose the subset that fits their needs – but if they report a specific variable it must use the codes and approach designated for that variable. The guidance uses LOINC for the test codes, SNOMED CT for values of most categorical variables, and UCUM for the units of measure where they apply.

This NLM/HRSA specification has been successful. All 3 major NBS lab information system vendors can now demonstrate fully compliant HL7 messages. The state of Kentucky is about to go live with the delivery of all of its NBS reports to the Kentucky Health Information Exchange (KHIE) , and 75% of Kentucky doctors have signed up to use KHIE. Perkin Elmer, which provides laboratory report delivery services to Kentucky, as well as to a large percent of other NBS laboratories, has implemented delivery mechanisms on all four of its NBS reporting platforms and believes the NLM/HRSA message will work for all of their customers. The Oregon NBS laboratory – which also serves 5 other states (Hawaii, Alaska, Idaho, Nevada, and New Mexico) is far along in their testing of the NLM/HRSA message. Pennsylvania has adopted the specification for both of its labs as has at least one of the laboratories in California. Finally a large hospital in New York City is exchanging test HL7 order and result messages with the New York state NBS lab.

Related Publications

Abhyankar S, Lloyd-Puryear MA, Goodwin RM, Copeland S, Eichwald J, Therrell B, Zuckerman AE, Dowing G, McDonald CJ. Standardizing Newborn Screening Results for Health Information Exchange. AMIA Annu Symp Proc. 2010; 2010: 1–5. PMID21346929: PMC3041276. Available at: http://www.lhncbc.nlm.nih.gov/lhc/docs/published/2010/pub2010037.pdf.

Downs SM, van Dyck PC, Rinaldo P, McDonald C, Howell RR, Zuckerman A, Downing G. Improving Newborn Screening Laboratory Test Ordering and Results Reporting Using Health Information Exchange. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):13-8. PMID20064796 : PMC2995628

3.5.3.4 UCUM: An empiric study showing the poor state of units of measure in HL7 messages and a proposed solution

(Taft L, Schadow G, Wolf P, McDonald CJ)

We have obtained tables of LOINC codes, test names and the local units of measure used to report results for these tests, from 23 organizations – with a total of 110, 000 test names and codes. We have been analyzing the data to assess the adequacy of current unit naming practices and have found that they are quite irregular. For some units of measure, more than 25 string variants exist. The same unit string may mean quite different things, and some laboratories truncate the numerator or the denominator of the unit strings. No automatic interpretation of units is possible under the current circumstances. These data, and the growing interest in electronic transmission of laboratory data and other measurements – such as from laboratories to electronic health records – and the use of such measurements for clinical decision support, all underscore a need for a standard, computer-understandable unit string, and the Unified Code for Units of Measure (UCUM©) is the answer. UCUM (http://unitsofmeasure.org/) is a code system intended to include all units of measure used in international science, engineering, and business. It provides a syntax for creating any valid metric unit and most non metric units as well, and tools for converting between any two dimensionally equivalent measures. The purpose is to facilitate unambiguous electronic communication of quantities together with their units. The focus is on electronic communication, as opposed to communication between humans. A typical application of UCUM is in electronic data interchange (EDI) protocols.

3.5.3.5 Standardization of clinical genetic sequencing reporting and molecular cytogenetics reporting

(Ullman-Cullere M, Heras Y, Huff S, Wood G, Shabo A, McDonald CJ)

In collaboration with Intermountain Health Care, Harvard Medical School - Partners Healthcare Center for Genetics and Genomics, LOINC and the HL7 Genomics Special Interest Group and

Clinical Genomics Work Group, we have developed approaches for electronic reporting of: 1) genetics sequencing and comparable gene chip data, and 2) molecular cytogenetics data. Both of these build on HL7 version 2.x messaging technology (which is widely available) and both take advantage of the LOINC panel structure, and the LOINC answer list and description structure. The approach to reporting genetic sequencing data was formulated as a formal HL7 implementation guide, was submitted to ballot, and accepted as a standard. The approach to reporting molecular cytogenetics data has just been formulated as an HL7 implementation guide, and will be submitted to balloting in the next HL7 balloting cycle.

The contents of these molecular genetics reports are defined in terms of a set of LOINC panels. For example, the report defined for sending sequencing data has one panel for reporting the overall results of the study and a repeating panel – the DNA marker results panel (See Table 4) – that reports the genetic variation at a particular point in the genome. In this report structure, only differences between the reference sequence and the studied sequence are reported, and an instance of the DNA marker panel is included in the report for every such difference (genetic variation) along with the length of the study sequence.

51960-3	**DNA marker results panel in Blood or Tissue by Molecular genetics method**	**R/O/C***	**Data type**
53044-4	DNA marker identified panel in Blood or Tissue by Molecular genetics method	R	
48018-6	Gene [Identifier] in Blood or Tissue by Molecular genetics method	O	CWE
48013-7	**Genomic reference sequence [Identifier] in Blood or Tissue by Molecular genetics method**	**C**	CWE
51958-7	Transcript reference sequence [Identifier] in Blood or Tissue by Molecular genetics method	C	CWE
53045-1	Reference sequence alteration [Identifier] in Blood or Tissue by Molecular genetics method	O	
48004-6	**DNA sequence variation in Blood or Tissue by Molecular genetics method**	C	CWE
48019-4	DNA sequence variation type in Blood or Tissue by Molecular genetics method	O	CWE
48003-8	DNA sequence variation identifier [Identifier] in Blood or Tissue by Molecular genetics method	O	CWE
48005-3	Amino acid change in Blood or Tissue by Molecular genetics method	C	CWE
48006-1	**Amino acid change type in Blood or Tissue by Molecular genetics method**	O	CWE
47999-8	DNA region name [Identifier] in Blood or Tissue by Molecular genetics method	O	CWE
53034-5	Allelic state in Blood or Tissue by Molecular genetics method	O	CWE
48002-0	Genomic source class [Type] in Blood or Tissue by Molecular genetics method	O	CWE
47998-0	DNA sequence variation display name [Text] in Blood or Tissue by Molecular genetics method Narrative	O	ST
53037-8	Genetic disease sequence variation interpretation [interpretation] in Blood or Tissue by Molecular genetics method	C	CWE
53040-2	Drug metabolism sequence variation interpretation [interpretation] in Blood or Tissue by Molecular genetics method	C	CWE
51961-1	Drug efficacy sequence variation interpretation [interpretation] in Blood or Tissue Qualitative by Molecular genetics method	C	CWE

R = Required; O = Optional; C = Conditional

Table 4. DNA marker results LOINC panel

The DNA marker results panel contains another panel, but that panel rarely repeats, so it is convenient to think of the DNA marker results panel as one flat panel with 16 variables, of which only a few variables are absolutely required to convey the results. For results reported as changes

at the nucleotide levels, the following two variables would be the minimum: 1) DNA sequence variation (LOINC #48004-6), and 2) the Genomic reference sequence Identifier (LOINC #48013-7).

Individual LOINC variables carry additional defining information. They all carry text definitions. Categorical variables have a formal answer list; quantitative variables that are not dimensionless have units of measure.

Some examples of important variables:

The DNA sequence variation (LOINC #48004-6) is defined in terms of HGVS syntax as specified in the LOINC term definition:

> Human Genome Variation Society (HGVS) nomenclature for a single DNA marker. The use of the nomenclature must be extended to describe non-variations (aka "wild types").

The genome reference sequence variable (LOINC #48013-7) is defined in terms of specific databases of reference sequences:

> The genomic reference sequence is a contiguous stretch of chromosome DNA that spans all of the exons of the gene and includes transcribed and non-transcribed stretches. For this ID use either the NCBI genomic nucleotide RefSeq IDs with their version number (see: http://ncbi.nlm.nih.gov/RefSeq) or use the LRG identifiers, without transcript (t or p) extensions -- when they become available. (See: Report sponsored by GEN2PHEN at the European Bioinformatics Institute at Hinxton, U.K. April 24-25, 2008)[31].
>
> The NCI RefSeq genomic IDs are distinguished by a prefix of"NG" for genes from the nuclear chromosomes and prefix of "NC" for genes from mitochondria. The LRG Identifiers have a prefix of "LRG_" Mitochondrial genes are not in the scope of LRG

Amino acid change type variable (LOINC #48006-1) is categorical so it includes a definition AND an answer list (Table 5):

Text answer	Answer ID
Wild type	LA9658-1
Deletion	LA6692-3
Duplication	LA6686-5
Frameshift	LA6694-9
Initiating Methionine	LA6695-6
Insertion	LA6687-3
Insertion and Deletion	LA9659-9
Missense	LA6698-0
Nonsense	LA6699-8
Silent	LA6700-4
Stop Codon Mutation	LA6701-2

Table 5. LOINC answer list for Amino acid change type (LOINC #48006-1).

> Codified type for associated Amino Acid Marker. Amino Acid Marker's use the HGVS notation which implies the Amino Acid Marker Type, but the concurrent use of this code will allow a standard and explicit type for technical and display convenience.

The cytogenetics HL7 model follows the same general approach except focus on reporting the details of G-banding, FISH and array-CGH types of chromosome analysis.

Related Publications

Heras Y, Brothman AR, Williams MS, Mitchell JA, McDonald CJ, Huff S. Development of LOINC for integrating constitutional cytogenetic test result reports into electronic health records. J Am Med Inform Assoc. 2011. (*Submitted*).

Heras Y, McDonald CJ, Wood G, Brothman A, Shabo A, Ullman-Cullere M, Huff S. HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Cytogenetics Model, Release 1. ORU^R01 HL7 Version 2.5.1. Ann Arbor(MI): Health Level Seven (HL7); 2011 May.

Huff S, Wood G, McDonald CJ, Heras Y, Joshi V, Babb L, Clark E, Shabo A, Ullman-Cullere M, Pochon P. HL7 Informative Document: HL7 Version 2 Implementation Guide: Clinical Genomics; Fully LOINC-Qualified Genetic Variation Model, Release 1 (2nd Informative Ballot) ORU^R01 HL7 Version 2.5.1. Ann Arbor(MI): Health Level Seven (HL7); 2009 Aug.

4 Comparison of patient-provided medication histories with pharmacy fill records – Bethesda Hospitals Emergency Preparedness Partnership (BHEPP) study at Suburban Hospital

(Fung KW, Kayaalp M, McDonald CJ)

The Suburban Hospital pharmacy's first step in filling a prescription is to enter the prescription into their local computer. If the patient has insurance coverage for medications, that prescription information (at least the drug name, drug ID, and duration) is sent to a computer at a pharmacy benefit manager (PBM). PBMs are national organizations that contract with insurance payers to adjudicate prescription payments – including patient co-pays. RxHub was originally organized as a consortium of 8 PBMs to deliver theses prescription records to care providers (hospitals and/or clinicians) who have a documented care relationship with the patient. Their goal was to assist patient care and medication reconciliation. They can usually provide a list of all of the medications dispensed (new prescriptions and refills) for the last year. Collectively these 8 PBMs process 2.5 billion prescriptions per year accounting for about 65% of the prescriptions covered by insurance. Surescripts merged with RxHub and the number of organizations that contribute prescription records to the consortium has grown substantially.

The medication history is an important part of the emergency department (ED) assessment. However, manually-acquired medication histories are prone to gaps and consume significant nurse/pharmacist time, ranging from 9 to 34 minutes in reported studies. Therefore, in theory, information about the drugs *dispensed* to a patient should be a great help to the capture of medication history – especially in disaster circumstances where there may be not enough personnel time to take adequate medication histories.

NLM established a connection between Suburban Hospital and Surescripts to assess the value of real-time pharmacy dispensing information in the ED. The project was partially-funded by the Bethesda Hospitals Emergency Preparedness Partnership (National Naval Medical Center, NIH Clinical Center, and Suburban Hospital), based on the belief that such a service could save both time and lives in disaster circumstances. We compared the medication records obtained from Surescripts with the ED history to determine the degree to which the electronic records could supplement or supplant the manual ED history.

At Suburban Hospital, nurses gather medication histories during ED triage; and for about half of the ED shifts, pharmacists review the nurse-collected history with the patients and correct it. For three months, Suburban requested Surescripts data, via HL7 messages, on every patient, in order to validate and quality-test the data from Surescripts. The nurse enters four identifying items (Full name, birth date, gender, zip code) as a mini registration when the patient checks in at the triage desk. This registration process triggers an HL7 message in the hospital information system (HIS) that is delivered to Surescripts. Then in response, Surescripts sends an HL7 message back to the hospital that indicates whether or not the patient is registered in Surescripts, and includes records of any filled prescriptions it has for that patient.

During the 3-month trial period, the ED staff did not receive the Surescripts information, and triage nurses collected medications histories as they always had. We automatically compared the histories collected by triage nurses with the prescription records delivered by Surescripts, drug-by-drug, during that 3-month period to ascertain the degree to which dispensing records would add to the manually-collected patient history, and assess the degree to which Surescripts records could *replace* the manual history – especially in the time of a disaster.

Before making the comparison, we standardized the medication names listed in the ED medication history and those in the Surescripts dispensing records (which included prescription initial fills and refills) based on the generic ingredients of each listed item. We produced these standardized names by converting the original names into RxNorm ingredient names. When a drug item contained more than one ingredient, the standardized name was represented as the concatenated names of each ingredient, listed in alphabetic order. We did not take brand names, strength or dosage form into consideration, to avoid trivial distinctions in the comparisons. However, the two lists were not directly commensurate. The ED medication list is just a straight list of unique medications that the patient reports as current. Surescripts provides up to a year's worth of serial dispensings by date and medication, and for each dispensing, a statement of the duration as well as the provider's name. We cannot count all of these medications as current, so we first have to find a cut-off date that would shrink the Surescripts data down to a list more comparable to the ED history. In principle, we can calculate the precise end of supply (EOS) date for each medication, and use that date to decide which medications are current. However, for many reasons these calculations are not precise. Further, if we apply this approach strictly, and

only count medications which have not reached their EOS date, we will discard medications that the patient describes as current. This occurs because of hording, availability of alternate supplies, supply-stretching due to low compliance, and variation in patient interpretation of what "current" means. For example, a patient may consider their small prednisone supply, used only every 2-3 months for severe asthma bouts, as current. We know from published data that patients do list medications whose EOS date passed months ago as current medications,[32] even in countries where the prescription records are very complete. So the choice of cut-off dates for counting medications in the dispensed medications list presents a dilemma. If we pick an EOS close to the present date the method will pick medications from the dispensed list that are current – but it will also discard many medications that the patient considers current. If we pick an EOS date a long time before the present date, the method will be more sensitive, and we will find more of the drugs that the patient considers current, but it will also include many drugs that the patient is no longer taking. To avoid this dilemma, we chose two different cut-off dates — one that is strict and specific, and another that is lax and sensitive.

We used 7 days as the strict cut-off based on Lau's study,[33] and included only dispensed medications whose supply would last until at least 7 days before the ED visit as "current" medications. For the lax cut off, we used one year — again, we selected the cut-off based on Lau's study, which required extending the cut-off to one year to find 98% of the patients reported medications in a complete set of dispensing records.

4.1 Results

For the three month trial period, 9,966 unique patients presented to the Suburban Hospital ED. Roughly 5% of these patients — including all trauma, and urgent, patients — by-passed triage; therefore, we excluded that subset of patients from the primary analysis because they had no ED history for comparison against the Surescripts. We analyzed the data from the ED history compared with the Surescripts data, for the remaining 9,426 patients. These Suburban ED patients were older, much more often enrolled in Medicare, and more often arrived by ambulance (25% vs. 15%), than the national average for patients presenting to an ED.[34] The proportion of blacks was less than the national average; and the proportions of Asian, non-white and non-Black patients were correspondingly greater; the proportion of whites was identical to the national average.

Of the 9,426 triaged patients, 6,431patients (68%) reported one or more current medications to the triage nurse. After adjusting for the age distribution, this percent is comparable to the percent found in a national telephone survey (78%).[35] Of the triaged patients, 6,085 (65%) had a registration record in the Surescripts master file, and 5,040 (82.8 %) of these patients had prescription records in the Surescripts database.

For the 3,721 patient who had medications with EOS following the 7 day cut off, the Surescripts and ED History together provided 6.7 unique medications per patient. Surescripts included 4.3 medications per patient or 64 % of the total, some of which were also present in the ED history. Surescripts delivered an average of 1.5 medications that were not reported in the ED history — a 28% increase over the number of prescriptions delivered by the ED history alone. These improvements are consistent with the results of many smaller studies. We also analyzed the relative importance of the medications contributed by each of the two sources. The Surescripts-only medications tended to be more important than the ED-only medications. When we began delivering the Surescripts reports to the ED staff after our trial period ended, pharmacists reported informally that patients usually confirmed that they were taking the recently-prescribed Surescripts drugs that were not on the original ED history. So Surescripts data does supplement and add value to the manually collected history in this subset of patients. However, Surescripts does not cover all patients, so the 3,721 Surescripts included when we used the strict cutoff, represented only about 58% of the patients taking drugs per the ED history.

The data from the seven day cut-off analyses provides only part of the story. The more lax we make the cut-off time, the more of the ED history drugs we can find in the Surescripts dispensing data. We find 53% of the ED drug history in the Surescripts data when the cut-off is seven days, and that percent increases as we extend the cut-off to look earlier in time, reaching 66% when we set the cut-off at a year. In a country where the pharmacy records are complete, Lau[36] had to move the end of supply (EOS) cut-off date back a full year before he could find 98% of the patient's reported history in their pharmacy records.

With the one year cut-off, the number of patients with dispensing records and the number of unique drugs dispensed both increased—as would be expected. Surescripts delivered 8.1 unique drugs per patient for about 5,100 patients, and a total of 40,000 medications unique to a patient. This represents much more drug information, about a larger group of patients, than the 7 day cut-off. Furthermore, although most of these extra prescription records do not contribute to "current medication" history, they do provide a wealth of useful information, including the degree of care continuity, the names of the prescribing providers, patterns suggesting drug abuse, medication compliance problems and the discontinuation of a maintenance drug the patient may really need. If this information is presented in a flow sheet — as we present it in the current version of reports printed for the ED clinicians and pharmacists (see Figure 9) — the provider does not have to dig through pages of information to glean these patterns.

The one year data give us an idea about the sufficiency of the Surescripts data in disaster circumstances, where the manual histories might not be possible because of ED personnel shortages and/or patient impaired mental status. For patients who have any data reported in Surescripts, the 1 year data will include 73% of the patient's current medications.

Table 6 shows how we derive this figure by taking all of the one-year Surescripts medications that overlap with the ED medication history and adding to that the drugs that were reported *only* in the Surescripts 7 day history. This is certainly an encouraging number for the circumstances in which a manual history could not be taken. Providers could scan the list of the medications, paying more attention to those prescribed or refilled closer to the present and maintenance medications in this list (Figure 9). Moreover, if we repeated the study today, this coverage would likely be even higher, because Surescripts coverage has grown to an estimated 80% of the nations covered lives, from an estimated 65% when we started this study.

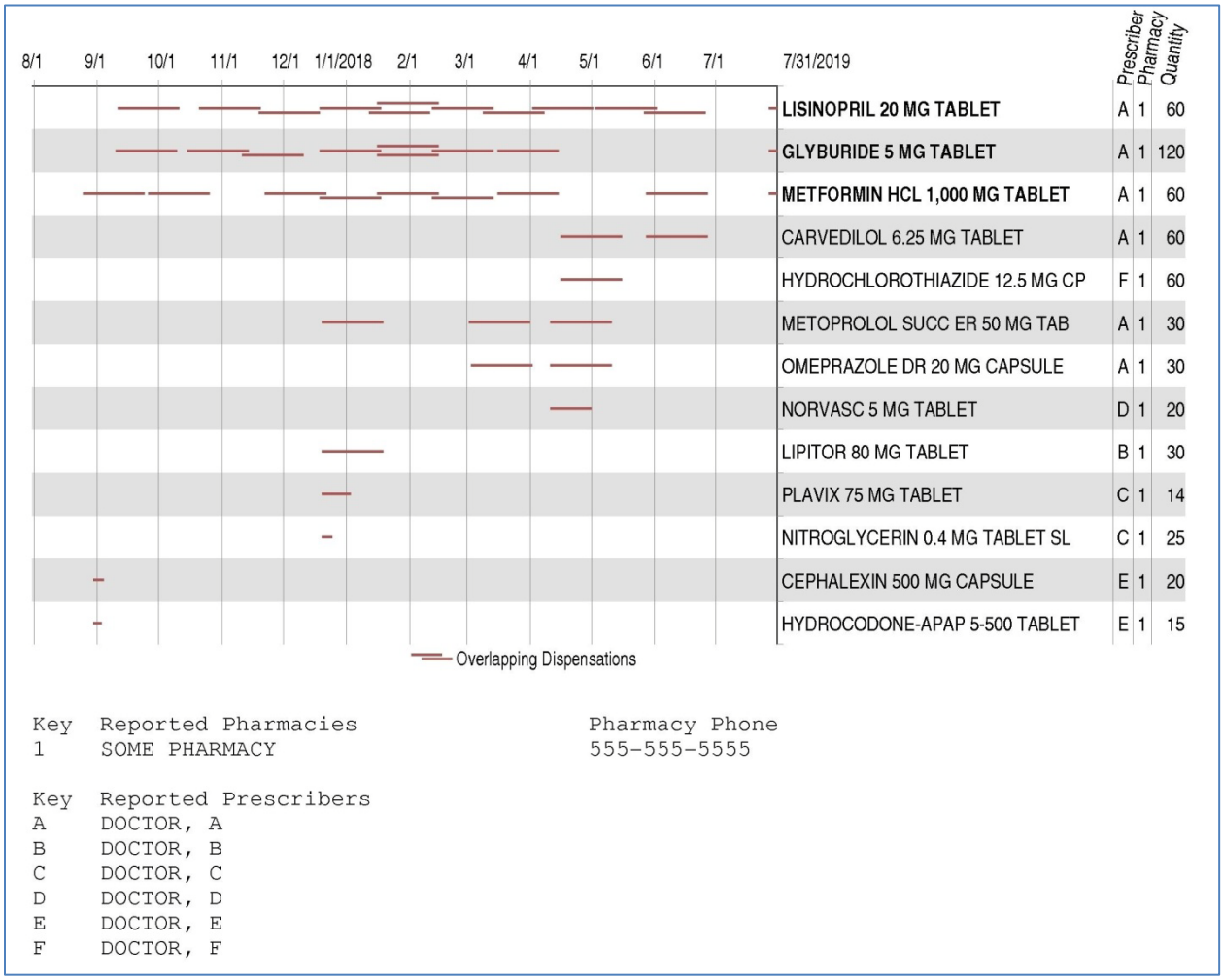


Figure 9. Patient medication history from Surescripts presented on a timeline graph (an excerpt from an example de-identified printed report that NLM designed and generated for clinicians).

Table Row ID	N = 5040 patients	Total # unique drugs for all patients	Average # drug items /patient
	Drugs in Surescripts (SS) at one year (including overlap with ED Hx)	40,976	8.1
	Drugs only in SS HX @ 1 yr cut off (can't count most as current drugs)	27,200	5.4
A	Drugs only in SS HX @ 7 day cut off and –count as current drugs = Minimum added value for SS Hx	5,550	1.1
B	Drugs in both SS and ED HX @ 1yr cut off (the intersection)	13,776	2.7
C	Drugs only in ED HX @ 1 yr cut off	7,140	1.4
D	Current drugs included in Surescripts (A+B)	19,326	3.8
E	Total current meds (A+B+C)	26,466	5.3
	Percent of total "current drugs" contributed by SS (D/E)	**73%**	

Table 6. Analysis of the contribution of a full year of Surescripts data to the total information about current medications.

In conclusion, the delivery of dispensing information contributes an important increment of current medications to the manually collected history, even if one focuses only on medications with an EOS date close to the present. Furthermore, the one year prescribing list which contains more than 70% of the patient current drugs could be a godsend in circumstances where a manual history cannot be obtained.

These results are very positive but they stimulate greater expectations. Just about every prescription in the U.S. is entered into a computer and the vast majority of these computer systems use codes for the medications; so why can't we get dispensing records for all prescriptions belonging to a given patient when they come for emergency care? Care organizations should not have to spend 9 to 34 minutes gathering this data, and even then, still end up with incomplete histories. Today there is also a National Coalition of Pharmaceutical Distributors (NCPD) standard that would support a community of medication consolidators. Policymakers should encourage the many large reservoirs of prescriptions – such as TriCare, many Medicaid systems, large pharmacies' systems, large HMOs that dispense prescriptions directly, and the government prescribing systems such as the VA and DOD – to become part of some consolidating infrastructure so that we could count on complete dispensing records to inform complete medication histories.

5 Mining data: A test bed for clinical database research – MIMIC II from MIT/Harvard

5.1 What is MIMIC II?

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC II) database[37] was developed by MIT under a grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB). MIMIC II contains de-identified data from the intensive care unit

(ICU) at a tertiary care hospital for over 32,000 visits covering more than 26,000 unique patients over an 8-year period. The database includes approximately 19,500 adult patients and 6,500 neonates and over 200 million rows of discrete data, including almost all tests and measurements recorded in the ICU, most of which are numeric, as well as the full text narrative of discharge summaries, nursing notes, and radiology reports. The first version we obtained was version 2.1, which also contained detailed nursing data and measures such as the Glasgow coma score, ventilator settings, and the Simplified Acute Physiology Score[38] – a measure of disease severity at ICU admission. In fall 2010 we obtained MIMIC II version 2.5, which added both in-hospital and post-discharge dates of death. The next update of the database will add Diagnosis Related Group (DRG) codes to the base set of variables. DRG codes use several criteria, including diagnosis, procedures, and complications, to classify patients admitted to a hospital in order to represent resource consumption and therefore guide Medicare reimbursement. To accomplish the de-identification, the MIT researchers replaced names, identifying numbers and other HIPAA identifiers with dummy data. Dates were transformed into random future dates. During this transformation, the temporal relationship among events within a single patient's record was preserved.

We obtained the MIMIC II database for the following four purposes:

1) To learn how to organize and standardize a large, longitudinal clinical database,
2) To understand how best to use de-identified, secondary clinical data for research purposes,
3) As grist for experimenting with innovative search technologies, and
4) As a source of extensive narrative data for testing natural language processing (NLP) techniques.

5.2 Clean up of the MIMIC II database

Work to achieve the first goal is ongoing. First, we converted the MIMIC II database structure into a structure closely corresponding to the fields and segments of HL7 observation messages in order to yield a more general and standardized structure. Then, we had to invest in considerable "clean- up." Many variables in MIMIC II were stored under multiple local codes for the same variable, each with a different name – the entry of tests and measurements was not constrained by a fixed vocabulary; so nursing could add new variables at will. If they misnamed or misspelled an existing variable, the system would create a new duplicate variable with the name they entered.

In addition, the data was not always segmented into its proper field. The MIMIC II database had fields for numeric values and units of measure, but often the units of measure were included with the numeric value in the field intended for the string value instead of being separated into the numeric value and the units of measure fields. For example, you might see "25 mg/dL" as a text value instead of "25" in the numeric field and "mg/dL" in the unit field. In these cases we had to parse the numeric value from the string units so that we could analyze the numeric values. In

another example, blood pressure values were stored with the systolic and diastolic measurements within two fields of a single observation record, and we had to split them into two observation records, one for systolic blood pressure and the other for diastolic. Because nursing hand-entered some laboratory results and the laboratory also delivered many of these results to the system, the same result could have been entered into the database more than once. We identified and isolated such redundant data so as to not include duplicate observation values in our analyses.
We mapped most of the laboratory test variables to LOINC and provided the mappings to the MIMIC II team, who will include these LOINC codes in their next release of the dataset. The LOINC mappings eliminated the many variations in codes and names for the same test as described above. We will be extending the LOINC mapping effort to non-laboratory tests and measurements such as vital signs and ventilator settings in the near term.

Acute physiology scores (APS and APACHE) are widely used to estimate severity of illness in individual patients and to facilitate reliable comparisons of outcomes in patient groups. The full physiologic scores require 34 or more variables, some of which are systematically missing in the MIMIC II dataset and therefore cannot be calculated. The Simplified Acute Physiology Score (SAPS) relies on 14 more readily available, variables (see Table 7) and was specifically developed to evaluate the risk of mortality for patients admitted to the ICU. SAPS scores were retrospectively computed by the MIT research team for MIMIC II version 2.5. In the process of replicating their scores, we found additional data that could be used to compute the scores for almost all of the patients, and we reconciled the calculation discrepancies with the MIT team. After reviewing all outliers, we also determined that some of the extreme values for the variables contributing to SAPS were correct for a given patient, while others were device or data entry errors.

Age
Heart rate
Systolic blood pressure
Body temperature
Respiratory rate or Ventilation status
Urinary output
Hematocrit
White blood cell count
Blood urea nitrogen (BUN)
Serum glucose
Serum potassium
Serum sodium
Serum bicarbonate (HCO3)
Glasgow coma score

Table 7. List of the 14 Simplified Acute Physiology Score (SAPS) variables.

We have also been working on several projects related to the second goal – that of learning *how* to best use this data for research purposes. To start that process we came up with three questions that are currently relevant to ICU care, and the work related to each is described in the following sections. The primary goal of the first analysis is to replicate and possibly extend a prior study. The goal of the second study is to explore a current topic in intensive care and contribute new information to the field. The third study has the goal of developing a prediction rule to identify ICU patients with poor survival.

5.3 Study 1: Are there associations between glucose levels and mortality during and after ICU stay?

(Callaghan F, Abhyankar S, Scariati P, Demner-Fushman D, McDonald CJ)

In 2001, Greet Van den Berghe and her colleagues in Leuven, Belgium reported the results of a randomized, controlled trial in the New England Journal of Medicine demonstrating that subjects who were in a surgical critical care unit for more than 5 days and on mechanical ventilation experienced a 43% reduction in ICU mortality when their glucose levels were maintained between 80 and 110 gm/dL.[39] In 2003, James Krinsley looked at this question using retrospective data collected through an Electronic Health Record (EHR) in a heterogeneous ICU population (medical, surgical, coronary). He found an increasing rate of hospital mortality with increasing mean glucose levels. Subjects with an average glucose between 80 and 99 mg/dL (264 subjects) had a 9.6% death rate, those with an average glucose between 100 and 119 mg/dL (491 subjects) had a 12.2% death rate, and this trend continued with a 42.5% death rate reported in 40 subjects with a mean glucose greater than 300 mg/dL. Krinsley found other significant predictors of death using logistic regression – age, mechanical ventilation, mean glucose, and modified Acute Physiology and Chronic Health Evaluation II (APACHE II) score.[40] Krinsley's follow-up trial

			30-day mortality (N=16237)		**90-day mortality (N=16237)**	
Variable	*Category*	*# subjects*	*Hazard ratio (95% CI)*	*p-value*	*Hazard ratio (95% CI)*	*p-value*
Mean glucose	(for each additional 1 mg/dL)	16237	1.007 (1.006, 1.008)*	<0.0001	1.007 (1.006, 1.008)*	<0.0001
Number of glucose measurements	(for each additional value)	16237	0.989 (0.986, 0.993)*	<0.0001	0.998 (0.996, 1.000)	0.107
Age	<45	2438	Reference	<0.0001 for overall age	Reference	<0.0001 for overall age
	45-<65	5422	1.903 (1.592, 2.275)*		2.135 (1.812, 2.516)*	
	65-<80	5202	2.762 (2.318, 3.291)*		3.341 (2.847, 3.92)*	
	80+	3175	4.589 (3.851, 5.468)*		5.645 (4.808, 6.629)*	
Gender	Female	6889	Reference		Reference	
	Male	9348	1.078 (0.986, 1.177)	0.098	1.170 (1.084, 1.264)*	<0.0001
First Service[+]	CSRU	3788	Reference	<0.0001 for overall ICU	Reference	<0.0001 for overall ICU
	CCU	2783	4.960 (4.071, 6.043)*		4.174 (3.543, 4.919)*	
	MICU	5437	7.027 (5.866, 8.417)*		6.069 (5.230, 7.042)*	
	SICU	4229	5.848 (4.876, 7.015)*		4.923 (4.237, 5.720)*	
Dialysis	No	14949	Not selected by stepwise procedure		Reference	
	Yes	1288			1.099 (0.970, 1.245)	0.139
Modified SAPS	(for each additional point)	16237	1.139 (1.126, 1.151)*	<0.0001	1.113 (1.102, 1.124)*	<0.0001
TPN	No	15172	Reference		Reference	
	Yes	1065	1.138 (0.983, 1.317)	0.084	1.184 (1.045, 1.342)*	0.008
Weight	(for each additional kg)	16237	0.995 (0.992, 0.997)*	<0.0001	0.993 (0.991, 0.995)*	<0.0001
% glucose > 110	(for each	16237	1.324 (1.099, 1.596)*	0.003	Not selected by stepwise	

	additional %)				procedure	
% glucose < 65	(for each additional %)	16237	7.913 (4.699, 13.328)*	<0.0001	5.010 (3.206, 7.831)*	<0.0001
Steroids	No	13018	Reference		Reference	
	Yes	3219	1.259 (1.145, 1.383)*	<0.0001	1.286 (1.184, 1.396)*	<0.0001
Creatinine	(for each additional 1 mg/dL)	16237	1.104 (1.077, 1.132)*	<0.0001	1.083 (1.058, 1.109)*	<0.0001
Hemoglobin	(for each additional 1 g/dL)	16237	1.026 (0.994, 1.059)	0.11	0.975 (0.948, 1.002)	0.073
Platelets	(for each additional 1 K/uL)	16237	0.999 (0.998, 0.999)*	<0.0001	0.999 (0.998, 0.999)*	<0.0001
Ventilator	No	6732	Reference		Reference	
	Yes	9505	1.336 (1.185, 1.506)*	<0.0001	1.127 (1.018, 1.248)*	0.022
Diabetes	No	13316	Reference		Reference	
	Yes	2921	0.551 (0.484, 0.628)*	<0.0001	0.614 (0.551, 0.685)*	<0.0001
Sepsis	No	15064	Reference		Reference	
	Yes	1173	1.472 (1.304, 1.661)*	<0.0001	1.410 (1.268, 1.568)*	<0.0001

Table 8. Hazard ratios for 30- and 90-day mortality compared to stated reference group for each variable, after adjusting for all other covariates.

* = statistically significant; + CSRU = Cardiac Surgery Recovery Unit, CCU = Coronary Care Unit, MICU = Medical ICU, SICU = Surgical ICU.

published in 2004 showed remarkable results – after an intensive glycemic control protocol was implemented in the ICU there was a 29.3% decrease in hospital mortality, as well as a 10.8% decrease in ICU length of stay.[41]

However, other prospective studies from the same time period have had mixed results. Van den Berghe's prospective trial published in 2006 showed that intensive insulin therapy only reduced in-hospital mortality for patients staying in the ICU for 3 or more days from 52.5% (n = 200) to 43.0% (n=166), but had no significant effect on patients who were in the ICU for less than 3 days or when the two groups were combined.[42] In addition, in 2009 randomized-control trial named NICE-SUGAR (Normoglycemia in Intensive Care Evaluation – Survival Using Glucose Algorithm Regulation) found that tight glucose control actually increased mortality. Those in the intensive control group (n=3054) had insulin therapy to achieve target glucose levels between 81 and 108 mg/dL, while those in the control group (n=3050) had a more conventional target of <180 mg/dL. Those in the tight control group had a 27.5% ICU mortality rate, while those in the control group had 24.9% mortality (p = 0.02), and there was no difference in ICU length of stay between the two groups. Also, episodes of severe hypoglycemia with a serum glucose <40 mg/dL occurred in 6.8% of patients in the intensive therapy group and only 0.5% of those in the control group (p<0.001).[43]

Our first goal was to determine whether our results would support Krinsley's conclusions regarding benefits of tight glycemic control or whether they would be in-line with the NICE-SUGAR results for short-term mortality. We started with the approximately 19,500 unique MIMIC II adult ICU patients. Of those, we excluded almost a thousand because they were

missing a hospital identification number and could not be linked to their hospital data by the MIT MIMIC II team. We excluded an additional 2,000 due to other missing data, such as glucose measures or multiple other covariate values. Our final study population is almost ten times that of the Krinsley retrospective study, and at least two and a half times larger than any study – retrospective or prospective – reported to date. Besides having a much larger sample size (>16,000 subjects) compared to Krinsley's retrospective study (1826 subjects), we also have the advantage of having complete mortality data for at least one year post hospital discharge.

We performed survival analysis using the R statistical package with forwards and backwards Cox proportional hazards stepwise regression. We use mean glucose value as the primary predictor, and include age, gender, weight, type of ICU admission, SAPS score – modified to exclude the glucose and age components of the SAPS score to enable including these covariates separately in the models – (see Table 7 for SAPS components), ventilator status, diabetes, sepsis, TPN (i.e. IV nutrition), oral or IV steroids, and lab values such as creatinine, hemoglobin, and platelets as covariates in the preliminary analysis. For this first analysis, we looked at mortality 30 and 90 days from the date of ICU admission.

The survival analysis shows a "J-shaped" relationship between a patient's mean glucose level and both 30- and 90-day mortality (Figure 10). In general, those with a mean glucose between 80 and 120 had the lowest mortality by both 30 and 90 days, while those with mean glucose less than 80 or greater than 120 had worse outcomes. Our results follow the same pattern as Krinsley's original results (also shown in Figure 10), with the only difference being that his in-hospital mortality rate is higher than our 30-day mortality for most of the mean glucose categories.

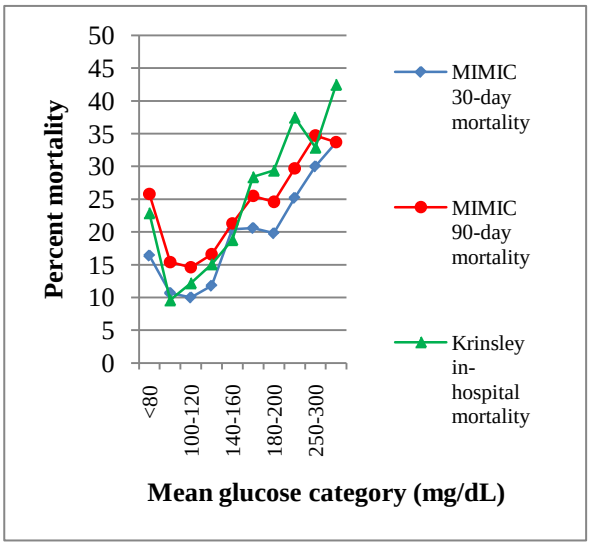


Figure 10. 30-day, 90-day, and Krinsley's in-hospital mortality rate by mean glucose category.

Mean glucose as well as several of the covariates in our model have a significant effect on both 30- and 90-day mortality (Figure 10). When mean glucose is analyzed as a continuous variable, for every 1mg/dL increase the hazard ratio of death increases 1.007 times for both 30 and 90 days. When we looked at the percent of glucose values above 110, by 30 days after admission the hazard ratio was 1.3 for each percent increase in the number of glucose values above 110 (p=0.0036). For each percent increase in the number of glucose values under 65, the hazard ratio for 30 day mortality was a startling 7.8 (p<0.0001). Other interesting covariates in our analysis that were

statistically significant include mechanical ventilation, type of ICU admission, steroid medication, sepsis, and diabetes (all with $p<0.001$).

Our results clearly support Krinsley's original finding that patients with a mean glucose less than 80 mg/dL have a significantly higher mortality than those in the 80-100 group, and that those with mean glucose ≤ 65 do even worse. These results could potentially explain the contradictory results of Van den Berghe and Krinsley compared to NICE-SUGAR. NICE-SUGAR compared patients getting intensive insulin therapy with a goal glucose of 80-108 to a control group with a goal glucose of <180, and found that patients in the intensive therapy group did significantly worse. However, the results were not stratified by the actual mean glucose values in each group, so we do not know if patients in the tight control group with mean glucose values less than 80 or those who had significant hypoglycemic episodes did worse than those with higher values. Given our results, we hypothesize that tight glucose control may, in fact, be protective, as long as it does not result in glucose levels that are excessively low. We are in the process of obtaining several additional interesting covariates to strengthen the model, and since we have complete mortality data for at least one year post hospital discharge, we are also going to extend our analysis to look at survival a year after discharge.

5.4 Study 2: What is the relationship between obesity and survival in the ICU?

(Abhyankar S, Callaghan F, Demner-Fushman D, McDonald CJ)

The prevalence of obesity has been increasing steadily over the past 25 years, and currently in the United States, 1/3 of the adult population is obese and another 1/3 is overweight.[44] Obese individuals are known to have a higher risk of developing chronic diseases such as diabetes, hypertension, osteoarthritis, and coronary artery disease,[45] and conventional wisdom suggests that they would also have worse outcomes during periods of acute critical illness. However, studies examining the relationship between body mass index (BMI) and outcomes after critical illness have had conflicting results, and only one study has looked at mortality after hospital discharge. Some support the hypothesis that obese individuals have worse outcomes during critical illness,[46,47] while others describe a completely different outcome – that obesity either has no relationship on mortality related to critical illness[48,49] or that it may actually be protective.[50,51] These studies have all dealt with small to modest numbers of patients (400-2200). The only exception is one study that only included 48,000 patients from a total population of 88,000. It excluded nearly half due to missing data,[13] undermining their conclusions. Some of the studies that asserted that obesity is protective grouped underweight and normal weight patients together, which could invalidate their conclusions since being underweight is strongly correlated with severe disease and higher mortality compared to normal weight patients. The one study from Australia that examined post-hospital mortality was small (500 patients) but did find a protective effect of higher BMI against mortality up to one year after ICU admission.[14]

MIMIC II has data on approximately 19,500 adult ICU patients. Of those, almost a thousand were excluded because they were missing a hospital identification number and could not be linked to their hospital data by the MIT MIMIC II team. An additional 2,000 were excluded due to other missing data, such as weight or multiple covariate values. Age data was available for all patients younger than 90 at the time of hospital admission. Those subjects 90 and older were grouped together for analysis because the de-identification process truncated specific age values at 89 per the HIPAA Privacy Rule.[52] We obtained weight and height values recorded in the ICU as well as from echocardiogram reports.

We were also able to reverse compute height and subsequently BMI for several hundred additional patients who were missing height data but who had body surface area (BSA) measurements recorded. We assigned height values to the remaining subjects without height data based on median height values for their age and gender. To calculate BMI, we used the standard formula: weight (in kilograms)/[height (in meters)]^2. Patients were grouped according to the BMI categories published by the Centers for Disease Control and Prevention (Table 9).[53]

BMI	Status
<18.5	Underweight
18.5 – 24.9	Normal
25-29.9	Overweight
≥ 30	Obese

Table 9. CDC BMI categories.

			In-hospital death (N=16854)		**Death 365 days after last hospital discharge** (N=16854)	
Variable	*Category*	*# subjects*	*Hazard Ratio (95% CI)*	*p-value*	*Hazard Ratio (95% CI)*	*p-value*
BMI	<18.5	942	1.26 (1.08, 1.48)*	<0.0001 for overall BMI	1.36 (1.22, 1.51)*	<0.0001 for overall BMI
	18.5-24.9	5362	Reference		Reference	
	25-<30	5150	0.81 (0.72, 0.9)*		0.75 (0.7, 0.8)*	
	30+	5400	0.94 (0.85, 1.05)		0.78 (0.73, 0.84)*	
Age	<45	2561	Reference		Reference	
	45-<65	5594	1.3 (1.08, 1.56)*	<0.0001 for overall age	2.04 (1.79, 2.32)*	<0.0001 for overall age
	65-<80	5394	1.54 (1.29, 1.85)*		2.9 (2.55, 3.29)*	
	80+	3305	2.3 (1.92, 2.76)*		4.27 (3.76, 4.87)*	
Gender	Female	7188	Reference		Reference	
	Male	9666	1.02 (0.94, 1.12)	0.622	1.13 (1.06, 1.19)*	<0.0001
First Service$^+$	CSRU	3886	Reference		Reference	
	CCU	2946	4.2 (3.43, 5.14)*	<0.0001 for overall ICU	3.32 (2.96, 3.74)*	<0.0001 for overall ICU
	MICU	5661	5.16 (4.3, 6.18)*		5.03 (4.54, 5.59)*	
	SICU	4361	3.89 (3.23, 4.69)*		3.67 (3.29, 4.09)*	
SAPS	(for each additional point)	16854	1.15 (1.14, 1.16)*	<0.0001	1.1 (1.09, 1.11)*	<0.0001
Ventilator	(for each additional day)	16854	0.95 (0.95, 0.96)*	<0.0001	1.01 (1, 1.01)	0.0016
Diabetes	No	13868	Reference		Reference	
	Yes	2986	0.62 (0.54, 0.71)*	<0.0001	0.8 (0.73, 0.86)*	<0.0001
Sepsis	No	15660	Reference		Reference	
	Yes	1194	1.28 (1.14, 1.44)*	<0.0001	1.45 (1.33, 1.59)*	<0.0001

Table 10. Hazard ratios for death in-hospital and 365 days after last hospital discharge compared to stated reference group for each variable after adjusting for all other covariates.

* = statistically significant; + CSRU = Cardiac Surgery Recovery Unit, CCU = Coronary Care Unit, MICU = Medical ICU, SICU = Surgical ICU.

Our primary goal for this study was to look at the relationship between BMI and both in-hospital and long-term survival among adults admitted to the ICU. We used the Cox proportional hazard model from the R statistical package to do a survival analysis. We were able to calculate the time to death up to 8 years after the subject's initial hospital based on available death data. BMI category is the primary predictor, with age, gender, type of ICU admission, SAPS score (see Table 7 for SAPS components), ventilator status, diabetes, and sepsis as covariates in the initial analysis.

Based on the preliminary analysis, obesity and overweight do appear to have a significant long-term protective effect ($p<0.0001$) for intensive care patients after adjusting for the above-mentioned covariates (Figure 11), and there is also a trend towards a protective effect during the hospital stay. These results confirm the Australian study of 500 patients.

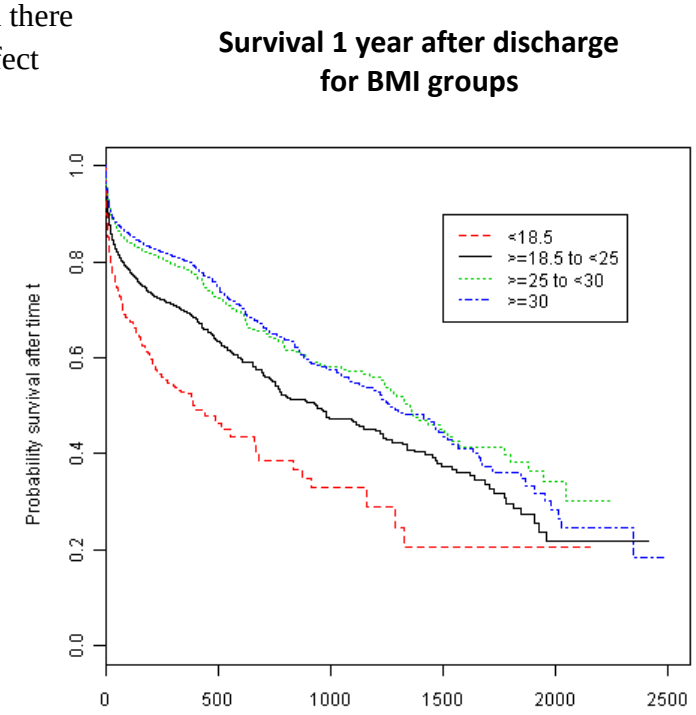


Figure 11. Long-term survival curves by BMI category.

Figure 11 contains the covariates, reference groups, and hazard ratios for both in-hospital and long-term survival. In-hospital survival was significantly better for overweight subjects with a 0.81 times the hazard ratio of death ($p<0.0001$) as compared to subjects with a normal BMI. The obese group had a numerically better hazard ratio of 0.89 but it was not significant by itself (CI 0.85-1.05); however, the overall trend of the BMI categories was also significant ($p<0.0001$). Long-term survival, defined as 365 days after the last hospital admission, was significantly better in both the overweight and obese BMI categories, with hazard ratios of 0.75 and 0.78 ($p<0.0001$), respectively. The underweight group had significantly worse in-hospital and long-term mortality compared with the normal BMI group, with 1.26 and 1.36 times the hazard of death ($p<0.0001$).

Other covariates that were significant predictors of both inpatient and long-term survival were age, type of initial ICU service, initial SAPS score, diabetes, and sepsis (all with $p<0.0001$). As expected, younger patients fared better, as did those with lower initial SAPS scores. Patients in the post- non-emergent cardiac surgery care unit (CSRU) did the best, and those in the medical ICU did the worst, with 5.16 and 5.03 times the hazard of death in the short- and long-term

compared to the post-cardiac surgery patients. Subjects with sepsis fared worse compared to those without sepsis for both time periods. One puzzling finding that we are exploring is the apparent inverse correlation between diabetes and mortality both during the hospital stay as well as after discharge.

There are several potential explanations for the "protective" effect of high BMI on mortality, including the anti-inflammatory effect of agents produced by adipocytes (fat cells),[54,55, 56] that obese patients might be admitted to the ICU for less acute illness due to more intense nursing needs compared to other BMI groups,[57] and that larger nutritional reserves of people who are overweight or obese might provide them with a greater capacity to fight critical illness than the normal or underweight who do not have such reserves.

5.5 Study 3: A prediction model for survival after ICU admission

(Callaghan F, McDonald CJ)

The intensive care usage rates in the U.S. are much higher than in many countries, especially among the elderly and in end of life situations. The contrast with Great Britain is especially striking. In the U.S., people who die in the hospital are five times more likely to spend time in the ICU than in England; and among the elderly, the contrast is even greater. Over the age of 86, ICU use among terminal patients is eight times greater in the U.S. than in Great Britain. ICU care is associated with nearly 50% of deaths in the U.S. but only 10% of deaths in the U.K.[58] Yet people in the U.K. live longer than their U.S. counterparts. Given the great detail in the MIMIC II database we thought it might be worth exploring the predictors of mortality in the ICU to look for information that might influence policies and practice. As a first step, we examined the predictors of mortality at 30 and 90 days and one year in stepwise logistic regression analyses.

5.5.1 The prediction models

In order to identify important predictors of death after ICU admission, we developed three logistic regression models corresponding to the 30-day, 90-day and 365-day outcomes. Table 11 shows the mortality rates and sample sizes associated with each outcome.

	30-day mortality	90-day mortality	365-day mortality
Dead, n (%)	1045 (14.8)	1381 (19.5)	1901 (26.9)
Alive, n (%)	6031 (85.2)	5695 (80.5)	5175 (73.1)
Total, n (%)	7076 (100)	7076 (100)	7076 (100)

Table 11. Mortality rates and counts for death within 30-, 90- and 365-days of ICU admission

Each model was based on the same random sample of half of the patients; we reserved the other half for model validation. We used a stepwise approach search algorithm to find significant predictors of death from the larger database. Predictors that were selected by the stepwise algorithm that were common to all 3 models were: age, service unit, body weight, prothrombin time, code status (e.g. do not resuscitate (DNR) or full code), TPN (total parenteral nutrition, i.e. IV nutrition), serum creatinine, mean glucose, % glucose below 65, insulin,

diabetes, platelet count, Braden Scale score (a predictor of pressure ulcers for patients with impaired mobility), steroid use and modified SAPS score (SAPS was modified to remove age and glucose, which are already present in the model). Lab values (glucose, creatinine, etc.) were taken as averages over the whole ICU stay. For variables such as DNR order, weight and service unit, the first recorded value from that ICU was used. SAPS was calculated from values in the first 24 hours of admission. SAPS is taken from values in the first 24 hours of admission. The rest of the variables are static and as such not time-dependent, e.g. gender. We used averages over the whole ICU stay for the lab variables, and the binary variables are taken over the whole ICU stay. The exceptions were: Weight (first weight recorded) and DNR (first DNR record taken).

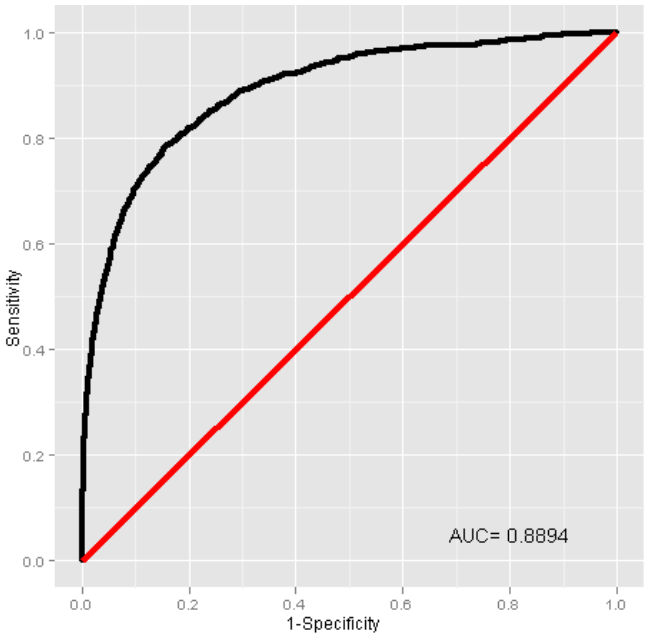


Figure 12. ROC for the 30-day mortality.

In addition, dialysis and % glucose measures that were above 110 selected for the 30-day model; gender was selected for the 90-day model; and gender, dialysis, hemoglobin level and use of ventilation were selected for the 365-day model. Across all of the models the factors associated with higher mortality are age (older patients at higher risk of death), service unit (patients admitted to units other than cardiac surgery had a higher risk of death), longer prothrombin time, DNR order, elevated serum creatinine, elevated glucose, % glucose readings above 110 and below 65, steroid use, and higher SAPS score. In addition, being male increased the risk of death for the 90- and 365-day models, and dialysis increased the risk of death in the 365-day mortality model.

In preliminary results, the receiver operating curve (ROC) – a measure of overall predictive ability of the model – showed excellent discrimination with the area under the curve (AUC) statistic equal to 0.89 for death within 30 days (see Figure 12). This is almost identical to the AUC of the Apache III model,[59] which predicts in-hospital mortality, was based on a huge sample and demands additional data collection.

Similar results were found for death within 90 days with an AUC=0.87, and death within 365 days, AUC= 0.85 (Figures not shown). Values of the AUC over 0.8 are considered as showing "excellent" discrimination, while values over 0.9 are considered to be evidence for "outstanding" discrimination.[60] The models can be used to predict the probability of death for each subject.

Predicted probability of death (%)	Number in each predicted category, N	Percent in each predicted category, %	Observed number of deaths, N	Observed Percent death, %
	30-day mortality			
91-100	85	1.2	83	97.6
81-90	112	1.6	95	84.8
71-80	146	2.1	118	80.8
61-70	146	2.1	93	63.7
51-60	164	2.3	94	57.3
41-50	213	3	95	44.6
31-40	312	4.4	115	36.9
21-30	474	6.7	102	21.5
11-20	892	12.6	112	12.6
0-10	4532	64	138	3
Total:	7076	100	1045	14.8

Table 12. Predicted and observed mortality percentages for death within 30 days.

Predicted probability of death (%)	Number in each predicted category, N	Percent in each predicted category, %	Observed number of deaths, N	Observed Percent death, %
	365-day mortality			
91-100	162	2.3	153	94.4
81-90	249	3.5	213	85.5
71-80	296	4.2	214	72.3
61-70	328	4.6	208	63.4
51-60	348	4.9	212	60.9
41-50	473	6.7	218	46.1
31-40	579	8.2	192	33.2
21-30	762	10.8	188	24.7
11-20	1276	18	163	12.8
0-10	2603	36.8	140	5.4
Total:	7076	100	1901	26.9

Table 13. Predicted and observed mortality percentages for death within 365 days.

This predicted mortality can be categorized into deciles (predicted probability of death 90-100%, predicted probability of death 80-90%, etc.) and the observed mortality rates can be compared for each of these predicted categories. Table 12 and Table 13 show the observed mortality rates for the 30-day and the 365-day models, respectively. In the 30-day model, there are 85 subjects that have a predicted probability of death of 90-100%, and of these 85 subjects we observed 83 deaths, which gave an observed mortality rate for that category of 97.6%, suggesting the admission was futile for most of the patients in this small population.

Of particular interest are the subjects who come in to the ICU with a "Do not resuscitate" (DNR) order, as opposed to a "Full Code." They comprise 299 patients or 4.2% of our development sample. Table 14 shows some of the covariate and outcome distributions for this group. By 365 days, 195 of the 299 (65%) DNR patients are dead, compared to 1706 out of 6777 (25%) of the 'no DNR' patients. In addition, DNR patients tend to be older (median age of 82 versus 66 for non-DNR patients), more likely to be admitted to the MICU (205 of the 299 DNR patients or 69%) compared to non-DNR (2087 out of 6777 or 31%),

	DNR, N (%)	No DNR, N (%)	Total
Total:	**299 (4.2)**	**6777 (95.8)**	**7076 (100)**
Outcomes:			
Death within 30 days	124 (1.8)	921 (13)	1045 (14.8)
Alive within 30 days	175 (2.5)	5856 (82.8)	6031 (85.3)
Death within 90 days	157 (2.2)	1224 (17.3)	1381 (19.5)
Alive within 90 days	142 (2)	5553 (78.5)	5695 (80.5)
Death within 365 days	195 (2.8)	1706 (24.1)	1901 (26.9)
Alive within 365 days	104 (1.5)	5071 (71.7)	5175 (73.2)

Table 14. The predictors and outcomes for DNR vs. non-DNR.
*indicates the standard deviation cannot be calculated for age due to de-identification of patients over 90 years old.

and less likely to be receiving ventilation (111 of the 299 DNR patients, or 37%) compared to non-DNR patients (4263 out of 6777, or 63%). These are patients whose first recorded DNR status is DNR or similar, where a patient's DNR status was typically recorded every 1-2 hours from time of admission.

We were surprised by the number of DNR patients admitted to the ICU. It does raise the question about why such patients with guidance not to resuscitate are admitted to a unit designed to provide immediate resuscitation. A large portion of these patients were discharged alive from the ICU: one half were alive at 90 days, and a third at a year. Further exploration will be needed to determine whether there were special circumstances that justified the admission of so many DNR patients to the ICU.

5.6 Developing and testing natural language processing techniques

(Demner-Fushman D, Abhyankar S, Callaghan F, McDonald CJ)

In addition to structured data such as laboratory test results and vital signs, which have explicit labels for each discrete variable and specified fields for storing the value of that variable, the MIMIC II database also contains large amounts of free text narrative such as is found in a discharge summary or radiology reports. These narrative reports are rich with clinical data and are the only source for many kinds of information. For example we wanted to ascertain smoking status, source of admission and discharge destination for some of the clinical studies described above. We wanted smoking status because it is an important outcome predictor in many circumstances, and we wanted to obtain the admission source and discharge destination to use nursing home and other chronic facilities to assess the patient's level of independence.
Only discharge summaries contain this information in the MIMIC database. LHC has substantial experience with the use of natural language processing (NLP) techniques to extract such information from free text narrative. In this section we describe the approach we used to extract the discharge destination from free text narrative to illustrate the process and indicate the success we had.

The version of the MIMIC II database that we used contained 25,056 non-empty discharge summaries. We developed a lexico-syntactic approach to variable extraction based on the discourse structure (e.g. section headings and organization), of the discharge summaries.
We first reviewed a sample of reports by hand to identify the sections of the discharge summary that would likely contain information about the variables of interest. For example there is a section in the discharge summary called "Discharge Disposition" which carries information about the discharge destination in more than half of the MIMIC II discharge summaries. The location from which the patient was admitted usually appears in the first few sentences of the

summary but it can be found anywhere in the report. For example, our system found the source of admission (bolded) in the following note:

> "Impression/Plan: 71 yo F with MMP h/o PVD s/p right AKA [**2017-10-13**], DM, CHF, CAD s/p stents, chronic atrial fibrillation, and multiple other medical problems, presents with hypotension and tachycardia.
>
> 1. Hypotension and Tachycardia - **patient was admitted to the FICU from rehab** with symptoms of hypotension, elevated WBC and tachycardia."

Smoking status was usually recorded under the sections called social history or history of present illness. Table 15 describes the section headings on which we focused our NLP searches for each of these variables.

From the discharge section, we extracted about 1,200 distinct text strings that carried information about the discharge destination from the sections described in Table 15, further processed these strings (mostly sentences) and then used the verbs and the location names that occurred in these processed strings to create a dictionary of verbs and another dictionary of location names for use in extracting admission sources and discharge destinations from any part of the report. Specifically we programmed the computer to search the text for verbs (discharged, sent, transferred, go, brought etc.) and locations (home, hospice, floor, etc.). If both term types are found in a sentence, we parsed the sentence with the Stanford parser. If there is a "to" dependency between the verb and location, the program assigns the discharge code of the location. If there is a "from" dependency an admission code (Table 16) is assigned.

Smoking status	Discharge status	Source of admission
Social history* History of present illness	Discharge disposition* Discharge status Discharge plans	History of present illness* History Chief complaint

Table 15. Example sections containing information for smoking status, discharge status, and source of admission.
* indicates sections with most information

Source of admission	Code	Patient count
Home	1	6366
Hospital	4	3428
Surgery		3821
Skilled nursing facility	5	532
Unknown	9	10226
Rehabilitation		683

Table 16. Source of admission.

In Table 17, we provide details about the extraction of discharge destinations. We first derived about 1,200 distinct discharge disposition values from the "Discharge Disposition" section and then reduced those to 64 distinct patterns that fall in to eight Medicare "Patient status" codes. We used coded information about patients' death in the hospital to evaluate accuracy of the discharge destination extraction for code 20. All but four of the 1794 patients assigned to the expired category (Code =20) were correct. These results are quite respectable. Interestingly, all four incorrect assignments of the code are due to the summaries that state "Discharge Disposition:

Expired," but then indicate a "discharge condition" as "good," "fair," or "stable" *and* include instructions for follow up care. Even to a human such contradictory information is confusing, and we are left not being sure if these patients lived or died.

Destination	Code	Patient count
Home	01	6634
Hospital	02	1944
Skilled nursing facility	03	4472
Home with care	06	3819
Unknown	07	4790
Expired	20	1794
Hospice	50	69
Rehabilitation	62	1534

Table 17. Presents the results of the extraction of discharge destinations.

Smoking status patterns were derived based on our world knowledge. We searched the "Social history" for sentences containing all derivational and inflectional variants and abbreviations of the word smoke (both as noun and verb), tobacco, cigarette, and pack (noun). Our goal was to classify each patient's smoking status as "ever smoked," "never smoked" and unknown. We derived 82 positive smoking status patterns. For example, "active smoker," "keeps smoking," "ex-smoker." Because smoking status is often indicated by verbs, we could not use NegEx[61] a program we usually use for eliminating negative statements. Instead, we derived 20 negative smoking patterns (for example, "does not smoke," "never smoked"). There are also two pseudo-positive patterns: "smoker in the household" and "smoking crack." Automatically searching for these patterns, we found that 6,994 patients have positive smoking status, 8,120 negative and the remaining 9,942 discharge summaries do not discuss patient's smoking status.

These methods do require some manual review (annotation) of the chart content to get an idea of the spectrum of relevant verbs and nouns used to describe a fact of interest. However, most of the detailed text parsing and searching is done by the computer, so these approaches are feasible for modest to possibly large numbers of variables that might be of interest to a research project.

Related Publication

Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-72. PMID19683066. PMC2757540.

5.7 An experiment with an exotic approach to searching medical databases using Solr instead of a relational database

(McDonald CJ, Kanduru A, Navarro A, Demner-Fushman D)

Today, most discrete data is stored in relational databases such as Oracle, DB2 or PostgresSQL. Relational databases have a strong theoretic foundation and a search language, SQL, based on Codd's[62] relational calculus. Relational databases require formal definitions (Schema) that

specify the content of the fields in each of the database tables and the linkages between records in different tables. The technologies for relational databases have been honed and optimized over the last 20+ years, have accumulated a large and rich experience base, and have displaced all of the preceding approaches to database construction. Recently, a set of different approaches – often referred to collectively as NoSQL ("not only SQL") databases[63] – have emerged that either do not require a pre-defined schema, or allow it to be changed and expanded at will. These databases use text documents instead of formal records as their building blocks and are designed to accommodate galactic size data collections. Google's Bigtable[64] is an example.

This project was stimulated by our experience with Lucene, a very fast and flexible indexing system from the open source Apache group. We have had very good experience with Lucene as the search engine in many projects including the NLM's L-PHR. It is also the search engine behind many large-scale NIH projects, including NCI's caBIG vocabulary server and NLM's new UMLS vocabulary server. Solr is another Apache project that adds many capabilities to the basic Lucene system, including more sophisticated search and aggregation capability, the ability to ingest many kinds of source documents, and to highlight the search words found in a document. It is Lucene on steroids. Solr is also described by some as a NoSQL system. [65]

We became intrigued by Solr's capabilities while using it to construct a number of LHC content sites. Though Lucene's original purpose was to find words and word patterns in text documents, under Solr it can find, and understand, dates and numeric values, provide counts and simple statistical information, and aggregate data across groups. So we wondered whether Solr could serve as an effective search engine for collections of clinical data that typically include both free text documents and structured observations. When we began discussing this idea, the betting was that it would be slow and clumsy when dealing with structured data that contained numbers and dates.

To test this idea we created a Solr document collection out of the MIMIC II data stored in Oracle. The MIMIC II database in Oracle has a structure that corresponds roughly to the general structure of an HL7 observation message. There is a record for the patient (PID), for the encounter (PV1), for the report header/order (OBR), and for the individual observation (OBX). In the MIMIC Oracle database, an individual text document, such as an x-ray report, is stored as an observation record in the OBX table just like a laboratory test result. The name/ID of the study/report is stored in one field of an OBX record, and the full text report in another.

A Solr collection is a collection of documents. When we imported the MIMIC II observation content into the Solr collection, we treated each OBX record as a separate document, but supplemented it with demographic fields (e.g. birthdate, gender), from the patient's registry record, so that these fields are readily available to Solr queries. Think of the document as an

OBX record that inherits fields from related tables in the database. We are adding more fields from other sources to the base OBX record in our next version of the collection.

Solr can use field tags in its searches. You decide which individual or multiple key fields to index and also which fields to put in the "ALL" field – which typically contains the whole document – and present what is searched as the default.

We have been able to run queries such as: counts of the number of observations by Observation ID, and statistics (average count, min and max, and standard deviation) for each numeric variable by observation within a single patient. In Table 18, we show examples of the time to run and the number of records retrieved for each of such queries. From our early results, we now know that Solr can successfully search structured data – including numeric data. Even better, it appears to be 3 to 100 times faster than the Oracle running the same queries. For interested parties, we will have a hand out at the BSC meeting showing the SQL and the Solr statements for each of the queries. However, the results in Table 18 should be interpreted cautiously, because the Oracle times might be improved by optimization we plan to do.

QUERY	Solr time (sec)	Oracle Time (sec.)	# retrieved (Oracle = Solr)
Total Count of all observations	1.1	198.9**	1
Count of Observations by Observation ID (or name) –ignoring patients	1.4	406.4**	11,964
Count of number of patients who have at least one observation for each observation ID	214.9	601.5	11,964
Statistics by observation ID for a selected patient with 494 distinct kinds of numeric observations	0.009	1.1	494
Statistics on Observation values for a specific patient with lots of observations	0.009	0.5	356
Count of distinct observations for each distinct patient.	278.7	724.7	26,639
Statistics for each observation within each patient	To be delivered	486.5	7,479,297
Table Size			208,071,836

Table 18. Comparison of Solr and Oracle query times.

Furthermore, there are some complicated queries that Solr cannot do, at least not with the capabilities available in the current release (Solr 4.0). However, it *can* execute many of the queries of interest to researchers for pulling specific observations from longitudinal or medical record databases. We are fairly sure that Solr can pull statistics (mean, sum, max, min, standard deviation) by variable within patients (the query we showed for a single patient) across the whole database; find the first or last value within a given hospital or ICU stay, across all, or a specified set, of variables and patients; find patients who met specified criteria for multiple variables including mixtures of variables with text and numeric values. We hope to demonstrate some of these queries at the BSC meeting.

In the final analysis, Solr may or may not be faster than a relational database for highly structured data. But at worst, its speeds will be competitive. It will provide an opportunity for fast and sophisticated text searching blended with structured searches, and will provide considerable flexibility as well as almost out-of-the-box data browsing opportunities. While it will never serve the role of a transaction processer, data can be added to the Solr content on the fly – but lazily – with delays of minutes not milliseconds.

6 De-Identification: Developing and testing the NLM Scrubber

(Kayaalp M, Brown A, Divita G, Ozturk S, Dodd ZA, McDonald CJ)

De-identification is a process that provides an extra layer of privacy defense when clinical data is used in research. Research on de-identified data can be done with a simple IRB exemption; therefore, de-identification enables research on large bodies of clinical data – such as those that exist in administrative databases and electronic medical record systems – without the delay or heavy time investment required for a full IRB review, and encourages the exploration of long-shot hypotheses that would not warrant a large time investment.

The requirements of de-identification are specified in the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996, which dictates the removal of 18 individually identifiable health information elements that could be used to identify the individual, or the individual's relatives, employers, or household members. De-identification of structured databases is relatively easy. One simply removes all of the fields (such as patient name and hospital number) that are specified in the HIPAA regulations. However, much of the richest clinical data in electronic medical record systems is recorded as dictation and stored as narrative text. The de-identification of narrative text is a challenge, because it requires finding text strings or numbers that may be identifiers without stripping out so much content that the records are not useful for research.

The purpose of this project is to develop software that can scrub the HIPAA specified identifiers from narrative text reports and thus de-identify them. We call this de-identifying program the NLM Scrubber. This effort had two phases. The first was to develop a set of hand scrubbed and annotated reports that we could use as a gold standard for testing and improving the NLM Scrubber. We used HL7 version 2.x messages obtained with IRB exemption from the NIH Clinical Center as the grist for this effort and designed the NLM scrubber to de-identify HL7 reports. HL7 reports are widely available and most narrative text in medical record systems can be obtained as HL7 messages, but the basic algorithms will work with other text formats as well. To develop the gold standard, we first developed an annotation tool that would find suspected identifiers with a very high sensitivity and let the human reviewer decide whether the suspect

was really an identifier and if so, what kind (e.g. date, identifying number, address, patient name, provider name, etc.).

Nurse contractors and a federal employee used the annotation tool to find and classify all identifiers within a large sample of reports. To date we have annotated the identifying information in nearly 17,000 narrative reports, for close to 7,500 distinct patients. This set includes roughly 6,000 reports from the dictation system (SoftMed), the preponderance of which were history and physicals, clinical summaries and almost 10,000 radiology reports (about half of which were plain films and CT scans). It also included 500 laboratory reports but we did not use them in this study because the numbers to date are not sufficient; we need time to increase these numbers.

To increase the independence of the reports in this sample, we selected only the most recent of each report type per patient. We used the content of the HL7 OBR-4 field (the ordering code) to define the report type. We used 1,140 of these gold standard reports to test and tune the algorithms in the NLM Scrubber before beginning the study described below.

The NLM Scrubber system operates through a sequence of pipelined processes: 1) HL7 message parsing, 2) part of speech tagging, 3) protected personal health information identification, and 4) redaction. We want to conserve as much of the clinically important information as we can, while removing all of the individually identifiable health information.

We have developed tools for finding and removing dates, and general identifiers such as the patient's hospital number and addresses. In this report, we focus on the de-identification of person names because that is the most difficult de-identification task, and our software is most developed for that type of identifier.

Source	Number of distinct source records at time of this study	Comment
Social Security	450 million individuals	Obtained first or last name that occurred more than twice in the data set
Social security death tapes	74 million individuals	These overlap considerably with the above
Medline author list	20 million articles	PubMed does not have full first names for many authors. A single publication can have many authors, and one author can be on many papers. This list did not have population-based information.

Table 19. Sources for the NLM Scrubber person name dictionary.

The NLM Scrubber uses a number of techniques to identify person names, including general person name dictionaries, frequency statistics for names and non-names, text patterns (e.g. words following "Mr." or "Dr." are likely to be names), and heuristic algorithms to identify names.

We built a large person name dictionary containing more than 3.8 million unique names, the vast majority of which also carried population-based frequency statistics. These names came from the

sources in Table 19. The total set included 1.3 million unique first names and 2.9 million unique last names. About 400,000 of the names are both first and last name.

Large numbers of words that have high frequency in narrative text such as "The," "He," "She," "It," "Is," "Can," "May," and "Of" are personal names in some languages (e.g. Chinese, Thai) and found in our person name dictionary. Consequently, we have to use statistical methods to distinguish person names from routine text, and we compare probabilistic distributions of a given word in various data sources including the content for clinical journal articles and the entire collection of Wikipedia articles.

To assess the performance of our de-identifier we selected a random sample of 3,095 clinical dictation records from our set of gold standard reports. These 3,095 reports represented 1,635 distinct patients and contained 1.2 million words. Of these, 22,583 were personal name words and 10.4% of these name words were patient names. None of the records used in this test set had been previously used for testing or development, nor did we take advantage of a custom list of patient name and provider names from the source hospital to assist the de-identification, or use any information about the format of the reports from that institution.

We ran the test set through the NLM Scrubber, and it performed well with a sensitivity of 99.9%. Of the 22,583 personal name words (counting a first name and a last name as separate words), it failed to scrub only 13 name words. Twelve of the name words that the system failed to scrub were parts of provider names. These included 8 instances of failures to find last names (4 unique names) and 4 failures to find two-letter middle initials (like AB as in “John AB Smith”). The scrubber never failed to remove a full name (i.e. a first and last name). The one failure on a patient related name was a failure to identify a nickname for a patient’s husband as a name.

HEENT: The patient was normocephalic with focal alopecia as noted. Sclerae anicteric. Moist oral mucosa with no lesions.

Neck: Supple.

Cardiac: Heart sounds were distinct with a regular rate and rhythm, no murmur, no abnormal heart sounds. Pulses were full and equal.

Chest: Clear to auscultation throughout.

Abdomen: No hepatosplenomegaly. Bowel sounds were auscultated and were normoactive. Abdomen was soft and nontender.

Extremities/Integument: The patient had a fused joint in her right hand second digit joint. Otherwise, no deformity, erythema, rash, edema or tenderness of extremities.

Musculoskeletal: Normal gait was unremarkable. The patient also had normal tandem gait, heel walking, and toe walking. Strength was 5/5 in all extremities.

Figure 13. The word “supple,” highlighted in yellow, was incorrectly classified as a patient name.

The NLM Scrubber’s specificity was 98.7%, which meant it misidentified 1.3% of the non name words as name words (1 out of 75) and removed them. This degree of false positive scrubbing has little effect on the readability of the remaining text (see example in Figure 13). We take this as a reasonable trade off for such high sensitivity.

The sensitivity of the NLM Scrubber is as good as any published figure and better than most. MIT-DeID has one of the best published performances, with a sensitivity of 100% for patient names,[66] but the two studies are difficult to compare. On the one hand, the MIT study took advantage of custom patient and provider name dictionaries, and their performance fell sharply without that support. The NLM study used only general dictionaries and yet achieved similar results. On the other hand, the MIT study dealt with typed-in nursing notes which are much more difficult to scrub, and our study examined only formal dictation whose content is more regular.

We ran four other natural language processing de-identification/person name recognition systems – including Mitre-MIST,[67] MIT-DeID, GATE-Annie,[68] and LingPipe[69,70] – on the same test set as the NLM Scrubber- and the results were very favorable to our Scrubber. The Mitre-MIST scrubber uses a machine learning paradigm and we let it train on uses a machine-learning paradigm. We trained it on 1,140 reports that did not overlap with the test set. However, we did not have a chance to fully optimize the tools for the systems that were designed for medical de-identification, and we will not present that data until we do.

Related Publications

Friedlin FJ, McDonald CJ. A Software Tool for Removing Patient Identifying Information from Clinical Documents. J Am Med Inform Assoc. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):601-10. Epub 2008 Jun 25.PMID18579831 : PMC2528047.

7 Low cost portable chest x-ray system and image analysis in Kenya

(Antani S, Karargyris A, Kohli M, Thoma G, Jenders RA, Goodwin RM, McDonald CJ)

AMPATH is a partnership founded in 1997 between Moi University School of Medicine, Moi Teaching and Referral Hospital (MTRH – Kenya's second national referral hospital in Eldoret Kenya), and a consortium of U.S. medical schools. The consortium is led by Indiana University School of Medicine and includes Brown University School of Medicine, Duke University School of Medicine, Lehigh Valley Hospital, Providence Portland Medical Center, University of Utah School of Medicine, and University of Toronto Faculty of Medicine. AMPATH also collaborates with more than 20 other universities and research organizations. The AMPATH mission is to address and reduce barriers to high quality care in the resource-constrained setting of Eldoret, the fifth largest city in Kenya, located on the Western side of the Rift valley. In partnership with USAID, AMPATH now manages the largest AIDS prevention and treatment program in sub-Saharan Africa, caring for more than 100,000 HIV-infected adults and children, in 48 sites across a 300-mile swath of western Kenya (www.iukenya.org/hiv.aids.html). Nearly one-half of these

patients are on anti-retroviral drugs. Most of these sites can only be accessed by four-wheel drive vehicles over difficult, unpaved roads.

Nearly a quarter of newly-diagnosed HIV positive patients harbor tuberculosis, which is a major cause of morbidity and mortality for these patients and represents a source of contagion to their HIV negative family and neighbors. Chest radiography is important to identify these patients, and to make decisions about their treatment and management. X-ray equipment is not available at most of the remote AMPATH sites, and the cost of installing and operating such equipment at all of the sites is beyond what is feasible in Kenya. The leadership at AMPATH believes that a lightweight, portable radiology system, which could be taken to remote sites by four wheel drive vehicles, would be feasible.

The NLM project's goal is to develop image processing algorithms that could identify the highest risk patients while they are still in the clinic and before the images could be delivered to and read by a Kenyan radiologist, and it will also document problems associated with the delivery and operation of this equipment and any solutions to these problems that could be used in other under-developed regions.

AMPATH has digitally photographed a collection of more than 1000 film-based chest x-rays taken at some x-ray equipped sites and has delivered them physically to Eldoret as JPEG files for the purpose of testing a system for electronically delivering chest x-ray images electronically to Kenyan radiologists for reading. In the first phase of this effort, AMPATH will de-identify its existing images for use by NLM to test algorithms that might provide early indicators about which patients are at modest to high risk of tuberculosis. They will use light weight easily transported radiology equipment purchased under this agreement to collect high resolution DICOM images at remote sites, and will track, record and deliver to NLM data related to equipment function, operation and use. In later stages of the project, NLM will explore ways to capture de-identified versions of the higher resolution images for automatic analysis and classification

This project leverages the image processing, analysis, and communication expertise at the LHNCBC, and aligns with NIH and NLM policy and strategic planning objectives in global and rural health.

8 NLM's Personal Health Record project: Merging, managing and mining data

(Abhyankar S, Lynch P, Wang Y, Jenders RA, Goodwin RM, Mericle L, Tao F, Muju S, McDonald CJ)

Note – this is a brief overview of the L-PHR. We will provide a full demonstration of the L-PHR at the BSC meeting and deliver a copy of a submitted paper that is under review for publication to the members of the BSC and to the attendees of the BSC meeting.

8.1 Background

Personal health record (PHR) systems are electronic health record (EHR) systems that were originally created to allow individuals to manage their own health. The earliest literature regarding the PHR concept is from 1978,[71] but serious development of PHRs did not begin until the late '80s and early 90's. Kaiser Permanente[72] and Beth Israel Deaconess Medical Center[73] both have long-running PHRs that are linked to the organization's EHR and provide a window (portal) into a given patient's existing EHR. Such PHRs are often referred to as tethered PHRs. Many of these tethered PHRs also allow administrative tasks such as refilling prescriptions and scheduling appointments. Stand-alone PHRs have evolved in parallel. Investigators from Harvard Medical School/Children's Hospital Boston introduced the idea of complete patient control of a PHR regardless of the original source of the data. [74] It is now called Indivo™ and is an open-source system designed to give individuals full control of their own health data.[75]

We have developed a PHR (the "L-PHR"), based on nationally accepted vocabulary standards. We call it the L-PHR because it was developed at the Lister Hill Center and at the National Library of Medicine. The intended audience for the L-PHR is individuals who want to keep track of their medical history and/or that of their children or other relatives whose health care they manage, such as elderly parents. The system will serve as a keeper of clinical records, an educator about the drugs and disorders that are recorded, and a prod toward healthy behavior and preventive care.

8.2 PHR Framework and Contents

The L-PHR has data tables for each category of key medical information, including health conditions, medications, immunizations, surgeries, and allergies, questions to ask the provider, and medical contacts, with corresponding data entry tables on the main L-PHR form. The main form also provides sections for recording specific observations that are relevant to preventive care measures such as a mammogram, low density lipoprotein (LDL) measurements, colonoscopy, and a handful of other variables. A second data input form is used to record a wide

range of other observations (e.g. complete blood count, daily exercise log, diabetes tracker), and the L-PHR generates a Web entry form on the fly based on the LOINC panel description for the chosen observation. All observation data is recorded in one file system containing an observation table and a panel table, which correspond roughly to the Health Level Seven (HL7) v2.x OBR (Observation Request) and OBX (Observation) segments.

The L-PHR makes extensive use of vocabulary standards throughout its content. We use RxTerms for drugs, LOINC for observations and observation panels, the Unified Code for Units of Measure (UCUM) for units of measure, the Centers for Disease Control and Prevention (CDC) CVX table for immunizations,[76] and SNOMED-CT for problems and surgeries. Allergy codes are still under development. These code systems are already widely-used both nationally and internationally, and four – LOINC, SNOMED-CT, RxNorm and CVX – are identified as "minimum standard" code sets by federal certification criteria adopted in 2010.[77]

8.3 Features of the L-PHR

The most important feature of the L-PHR is its commitment to standard vocabulary systems for encoding all of the important items in the profile. Not only does this approach offer the only viable option for automatic merging of clinical data from laboratories or hospitals, it also enables decision support, many of the display and data capture features, and the one-click access from a term on the page to educational material.

The second feature that sets L-PHR apart is the use of one large web page instead of multiple smaller ones for capturing an individual's key health data. It allows the user to systematically enter medical conditions, medications, surgeries, allergies, etc. without having to open any new windows or even use the mouse. Our rationale is that a single large page (like a spreadsheet) is a better format for many kinds of data entry and data management. Each page flip breaks user attention and requires re-orientation, with associated user time costs.

The L-PHR provides a variety of options for data entry, and ultimately it is the user's decision to decide which technique is most useful to her. When a field has a short or modest list of options the system presents the full list to the user as soon as she enters the field. The user can then select the desired option by typing its number, name (or part of its name), clicking on it, or by arrowing down to it and then pressing return. For fields that have too many choices to display in one fixed list, such

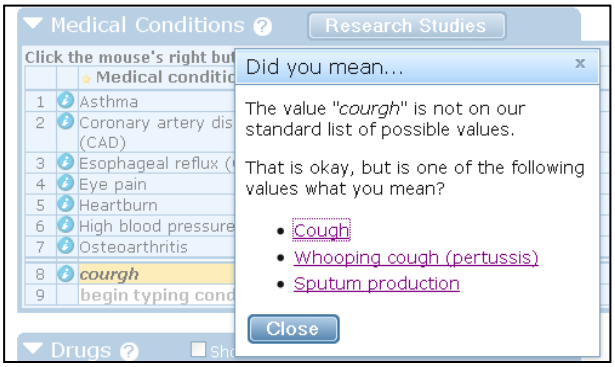


Figure 14. Demonstration of term checking.

as condition and medication, our auto-complete feature is a useful tool: as soon as the user types 3 characters, the L-PHR generates a list of matching items, and with each extra character, the list becomes shorter and more specific. If the user intentionally or unintentionally types in a term that is not in our data tables, the L-PHR does a series of checks to find the closest coded matches, and the user has the option of selecting one of these choices or staying with the original entry (see Figure 14).

Rule-based decision support permeates the L-PHR. We have developed a custom rule-authoring system that allows clinical authors without any programming experience to create reminder rules. There are two types of rules (fetch and value) that retrieve and make calculations on specific pieces of data from an individual's record, and a third type (reminder) that uses the fetch and value rules to generate custom text reminders based on the data that the user has entered. To date, our primary focus has been on preventive care reminders based on the advice of the U.S. Preventive Services Task Force (USPSTF) or the CDC. Currently, the L-PHR contains 54 fetch rules and 45 value rules, which in different combinations comprise 18 high-level reminder rules covering a variety of topics for both adults and children (Table 20). Several more reminders are in the planning stages, and any number of reminders can be added and customized. The L-PHR also supports another class of rules for controlling the field display on the form according to patient characteristics. For example, if the user is an adult age 50 or older, the L-PHR will display colonoscopy data entry fields; and if the user is a woman over 40, she will have the opportunity to enter information regarding mammograms.

Reminder Rules Topic	Target group
Breast cancer screening	Women over 40 (there are different versions for women over 40 and women over 50)
Cervical cancer screening	Women over 18
Colon cancer screening	Adults ≥ 50
Aortic aneurysm screening	Male smokers ages 65-75
Meningococcal vaccine	Adolescents
Tetanus vaccine	Adolescents/Adults
Influenza vaccine	All ages
Varicella (chickenpox) vaccine	All ages (depending on history of chickenpox)
Varicella zoster (shingles) vaccine	Adults ≥ 60
Smoking cessation	All smokers
Anemia screen	Children between the ages of 1 and 3
Pneumococcal vaccine	Adults ≥ 60 or high-risk individuals ≥ 2 years old
Bone density screening	Women ≥ 60
Daily folate	Women of reproductive age
Cholesterol screening	Adults ≥ 20 depending on risk factors
Elevated LDL warning	People with recorded LDL > 130

Table 20. The variety of topics and user groups covered by the reminder rules to date.

Another characteristic unique to the L-PHR is the ability for the user to enter almost any kind of data using on the fly conversion of LOINC panel descriptions into data capture forms. First, users pick a panel category, such as laboratory, radiology, or personal health trackers, which brings up a second menu with a list of panels available in that category. All of the panels and variables come from LOINC, so the panel data from the L-PHR can be exchanged in a fully coded form with any other site that uses LOINC, and any LOINC panel could be used to generate an input form. The PHR contains the knowledge for hundreds of panels contained in LOINC, and at this time there are over 250 active panels or tests for users to choose from. These observations cover

a wide variety of disease and wellness trackers, labs and radiology studies, and information specific to obstetrics and pediatrics.

The content in the PHR is linked to trusted educational resources using the standard codes for that particular content. Conditions and medications are directly linked to specific information within NLM's MedlinePlus,[78] and the conditions section also provides a link to a ClinicalTrials.gov search window. Vaccine information is linked directly to the CDC's Vaccine Information Sheets, and the reminder messages contain links that open new windows linking to specific information on either the USPSTF or CDC website, depending on the source of the preventive recommendations.

The L-PHR also has a rich flow sheet and graphing capability. Users can choose to display some or all of the observations they have recorded over time. The flowsheet includes a small sparkline[79] graph for each variable. Users can see a full graph for any quantitative variable by clicking on the small sparkline graph, and the larger graphs can either be displayed as line graphs or bar graphs, depending on the user's preference (see Figure 15).

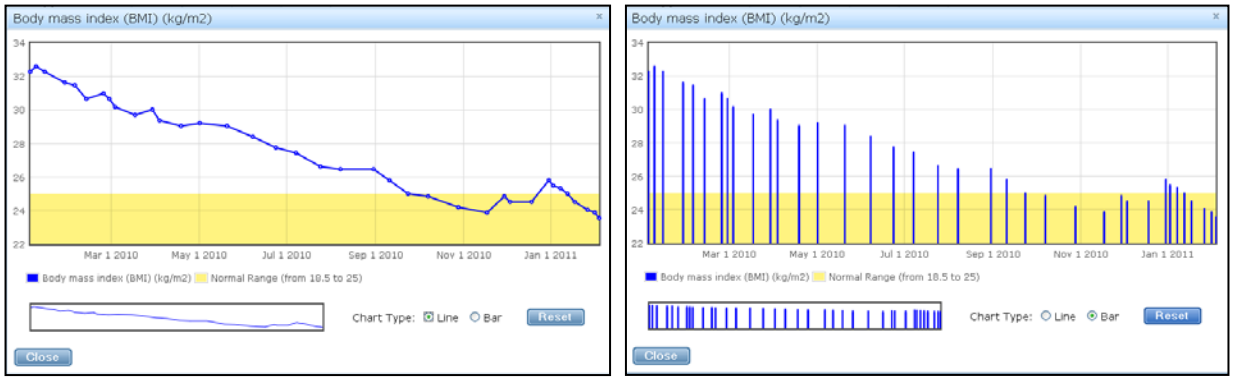


Figure 15. The user can choose to view data as a line graph or a bar graph.

8.4 Future

NLM is working with a community hospital that plans to adopt the software and deploy it as an un-tethered PHR available to patients and members of the community. We will obtain de-identified information from the host in order to assess usability, usage patterns and the adequacy of the vocabularies. Early implementation of the L-PHR will provide valuable data into how and why consumers use such tools, and over time we hope to have a measureable impact on consumer health and well-being.

Related Publication:

Abhyankar S, Lynch P, Wang Y, Goodwin RM, Jenders RA, McDonald CJ. The L-PHR: a Standards-Based Personal Health Record for Your Family. AMIA Annu Symp Proc. 2011. *(Submitted)*

9 References

[1] McDonald J. Protocol-Based Computer Reminders, The Quality of Care and the Nonperfectability of Man. N Engl J Med. 1976 Dec 9;295(24):1351-5.PMID: 988482.

[2] McDonald CJ, Tierney WM, Blevins L. The Benefits of Automated Medical Record Systems for Ambulatory Care. Proc of Comput Applications Med Care 1986; 175-181.

[3] Tierney WM, Miller ME, Overhage JM, McDonald CJ. Physician Inpatient Order-writing on Microcomputer Workstations: Effects on Resource Utilization. JAMA. 1993 Jan 20;269(3):379-83.PMID: 8418345.

[4] Rosenman MB, Mahon BE, Downs SM, Kleiman MB. Oral erythromycin prophylaxis vs watchful waiting in caring for newborns exposed to Chlamydia trachomatis. Arch Pediatr Adolesc Med. 2003 Jun;157(6):565-71. PMID: 12796237.

[5] McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, Mamlin BW. The Indiana Network for Patient Care: A Working Local Health Information Infrastructure (LHII). Health Aff (Millwood). 2005 Sep-Oct;24(5):1214-20. PMID: 16162565.

[6] McDonald CJ, Wiederhold G, Simborg DW, Hammond E, Jelovsek FR, Schneider K. A Discussion of the Draft Proposal for Data Exchange Standards for Clinical Laboratory Results. Proc Annu Symp Comput Appl Med Care 1984; 406-13. PMCID: PMC2578513.

[7] Logical Observation Identifiers Names and Codes [Internet]. Indiana: Regenstrief Institute, Inc.; c1994-2010 [cited 2011 Mar 24]. Available from: http://loinc.org/.

[8] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange, IT Prof;2005 Sept/Oct;7(5): 17-23.

[9] National Library of Medicine (NLM). Unified Medical Language Systems: RxNorm [Internet]. Bethesda, M.D.: NLM; 2004 Mar 22 [updated 2011 Feb 25; cited 2011 Mar 24]. Available from: http://www.nlm.nih.gov/research/umls/rxnorm/.

[10] International Health Terminology Standards Development Organisation (IHTSDO). Systematized Nomenclature of Medicine--Clinical Terms. Denmark: IHTSDO; [cited 2011 Mar 24]. Available from: https://uts.nlm.nih.gov/home.html.

[11] Thompson, Tommy G. (Secretary, U.S. Department of Health and Human Services, Washington, DC). Letter to: John Lumpkin, M.D., M.P.H. (Chairman, National Committee on Vital and Health Statistics (NCVHS), Hyattsville, M.D.). 2004 Sept 22. Located at: http://www.ncvhs.hhs.gov/040922lt.pdf)

[12] Office of the National Coordinator for Health Information Technology. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule, 45 C.F.R. Part 170. (2010). Available from: http://edocket.access.gpo.gov/2010/pdf/2010-17210.pdf

[13] U.S. Department of Health and Human Services Office of the Secretary. HHS Strategic Plan for Fiscal Years 2010-2015 [Internet]. Available from: http://www.hhs.gov/secretary/about/priorities/strategicplan2010-2015.pdf

[14] Charting a course for the 21st century : NLM's long range plan 2006-2016 / NLM Board of Regents. [Bethesda, Md.] : U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, 2006. http://www.nlm.nih.gov/pubs/plan/lrp06/NLM_LRP2006_WEB.pdf

[15] Interoperable Information: enhancing NLM's contribution to the nation's health IT agenda: Final Report to the NLM Board of Regents from its Working Group on Health Data Standards. [Bethesda, Md.] : U.S. Dept. of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, 2009 May. http://www.nlm.nih.gov/od/bor/BORHDSWG-report.pdf.

[16] Office of the National Coordinator for Health Information Technology. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule, 45 C.F.R. Sect 170.207. (2010). Available from: http://edocket.access.gpo.gov/2010/pdf/2010-17210.pdf

[17] Gage B, Reilly K. Overview of the Medicare Payment Reform Initiative – Fall 2009. Waltham(MA): RTI, International; 2009 Sept 15. Available from: http://www.pacdemo.rti.org/UserFiles/File/3_PAC_PRD2_Overview_9.15.09_ND.pdf

[18] Centers for Medicare and Medicaid Services (CMS). OASIS Data Set. Baltimore(M.D.): U.S. Department of Health and Human Services; 2009 Sept. Available from: https://www.cms.gov/HomeHealthQualityInits/12_HHQIOASISDataSet.asp

[19] Saliba D, Buchanan J. Development and validation of a revised nursing home assessment tool: M.D.S 3.0. Prepared for Centers for Medicare and Medicaid Services (CMS) Office of Clinical Standards and Quality, Quality Measurement and Health Assessment Group, 2008 Apr. Available from: http://www.geronet.med.ucla.edu/centers/borun/M.D.S%203.0%20Final%20Report.pdf

[20] Vreeman DJ, McDonald CJ, Huff SM. Representing Patient Assessments in LOINC. AMIA Annu Symp Proc. 2010; 2010: 832–836. PMID21347095: PMC3041404.

[21] C80 Clinical Document and Message Terminology Component. Healthcare Information Technology Standards Panel (HITSP), American National Standards Institute (ANSI) [cited 2011 Mar 24]. Available from: http://www.hitsp.org/ConstructSet_Details.aspx?&PrefixAlpha=4&PrefixNumeric=80

[22] Regenstrief Institute. LOINC News: Common lab orders LOINC value set version 1 now available. Indianapolis(IN): Regenstrief Institute; 2010 Jun 30. Available from: http://loinc.org/news/common-lab-orders-loinc-value-set-version-1-available.html

[23] Schadow G, McDonald CJ et al: Units of Measure in Clinical Information Systems. JAMIA. 6(2); 1999 Mar/Apr; p.151-62. Available from: http://www.jamia.org/cgi/reprint/6/2/151

[24] McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, Mamlin BW. The Indiana Network for Patient Care: A Working Local Health Information Infrastructure (LHII). Health Aff (Millwood). 2005 Sep-Oct;24(5):1214-20.PMID: 16162565.

[25] Indiana Network for Patient Care: Supporting clinical care through health information exchange [Internet] Indianapolis (IN): Regenstrief Institute. [cited 2011 Mar 24]. Available from: http://www.regenstrief.org/medinformatics/inpc

[26] Welcome to Partners HealthCare: founded by Brigham and Women's Hospital and Massachusetts General Hospital [Internet]. Boston(MA); [cited 2011 Mar 28]. Available from: http://www.partners.org/

[27] United Healthcare: Helping People Lead Healthier Lives [Internet]. c.2011 [cited 2011 Mar 28]. Available from: http://www.uhc.com

[28] Recommended Uniform Screening Panel of the Secretary's Advisory Committee on Heritable Disorders in Newborns and Children. [Rockville, M.D.]: U.S. Dept. of Health and Human Services, Health Resources and Services Administration, 2011 Feb. http://www.hrsa.gov/heritabledisorderscommittee/uniformscreeningpanel.htm

[29] Downs SM, van Dyck PC, Rinaldo P, McDonald C, Howell RR, Zuckerman A, Downing G. Improving Newborn Screening Laboratory Test Ordering and Results Reporting Using Health Information Exchange. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):13-8. PMID20064796 : PMC2995628

[30] Constructing Newborn Screening HL7 Messages. [Internet]. Bethesda(M.D.): U.S. National Library of Medicine, [updated 2010 Dec 14; cited 2011 Mar 28]. Available from: http://newbornscreeningcodes.nlm.nih.gov/nb/sc/constructingNBSHL7messages

[31] Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Béroud C, Dobson G, Lehväslaiho H, Taschner PEM, den Dunnen JT, Devereau A, Birney E, Brookes AJ, Maglott DR. Locus reference genomic sequences: an improved basis for describing human DNA variants. Genome Med. 2010; 2(4): 24. PMCID: PMC2873802.

[32] Lau H, Florax C, Porsius AJ, de Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. Br J Clin Pharmacol, 2000 Jun;49(6):597-603. PMID: 10848724. PMCID: PMC2015045.

[33] *Ibid.*

[34] Nawar E, Niska R, Xu J. Advance data from vital and health statistics: National hospital ambulatory medical care survey: 2005 emergency department summary. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Health Statistics; 2007 June 29. Report No. 386.

[35] Mitchell AA, Kaufman DW, Rosenberg L. Patterns of medication use in the United States: A report from the Slone Survey. Boston(MA): Slone Epidemiology Center at Boston University; 2006. Available from: http://www.bu.edu/slone/SloneSurvey/SloneSurvey.htm

[36] Lau H, Florax C, Porsius AJ, de Boer A. The completeness of medication histories in hospital medical records of patients admitted to general internal medicine wards. Br J Clin Pharmacol, 2000 Jun;49(6):597-603. PMID: 10848724. PMCID: PMC2015045.

[37] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark R. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC –II): A public-access intensive care unit database. Crit Care Med. 2011 Jan 28;39(5). PMID: 21283005.

[38] Le Gall JR, Loirat P, Alperovitch A, et al. A simplified acute physiology score for ICU patients. Crit Care Med. 1984 Nov; 12(11): 975-7.

[39] Van den Berghe G, Wouters P, Weekers F, et al. Intensive insulin therapy in the critically ill patients. N Engl J Med. 2001 Nov 8; 345(19): 1359-67.

[40] Krinsley JS. Association between hyperglycemia and increased hospital mortality in a heterogeneous population of critically ill patients. Mayo Clin Proc. 2003 Dec; 78(12): 1471-1478.

[41] Krinsley JS. Effect of an Intensive Glucose Management Protocol on the Mortality of Critically Ill Adult Patients. Mayo Clin Proc. 2004 Aug; 79(8): 992-1000.

[42] Van den Berghe G, Wilmer A, Germans G, et al. Intensive insulin therapy in the medical ICU. N Engl J Med. 2006 Feb 2; 354(5): 449-61.

[43] Finfer S, Chittock DR, Su SY, et al. Intensive versus conventional glucose control in critically ill patients. N Engl J Med. 2009 Mar 26; 360(13) 1283-1297.

[44] Ogden CL, Carroll M.D.. Prevalence of overweight, obesity, and extreme obesity among adults: United States, trends 1976–1980 through 2007–2008. Hyattsville, M.D.: National Center for Health Statistics Health E-stats; 2010 Jun 4 [cited 2011 Mar 22]. Available from: http://www.cdc.gov/nchs/data/hestat/obesity_adult_07_08/obesity_adult_07_08.htm

[45] National Heart, Lung, and Blood Institute; National Institute of Diabetes and Digestive and Kidney Diseases. Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults: the evidence report. Bethesda, M.D.: National Institutes of Health; 1998 Sep. NIH Publication No. 98-4083. Chapter 2, part C: Overweight and obesity: background, Health risks of overweight and obesity; p. 12.

[46] Bercault N, Boulain T, Kuteifan K, et al. Obesity-related excess mortality rate in an adult intensive care unit: A risk-adjusted matched cohort study. Crit Care Med. 2004 Apr; 32(4): 998-1003.

[47] Brown CV, Neville AL, Rhee P, et al. The impact of obesity on the outcomes of 1,153 critically injured blunt trauma patients. J Trauma. 2005 Nov; 59(5): 1048-51.

[48] Aldawood A, Arabi Y, Dabbagh O. Association of obesity with increased mortality in the critically ill patient. Anaesth Intensive Care. 2006 Oct; 34(5): 629-33

[49] Landi F, Onder G, Gambassi G, et al. Body mass index and mortality among hospitalized patients. Arch Intern Med. 2000 Sep 25; 160(17): 2641-4.

[50] Marik PE, Doyle H, Varon J, et al. Is Obesity Protective During Critical Illness? An Analysis of a National ICU Database. Crit Care & Shock. 2003; 6: 156 – 162.

[51] Peake SL, Moran JL, Ghelani DR, et al. The effect of obesity on 12-month survival following admission to intensive care: A prospective study. Critical Care Medicine. 2006 Dec; 34(12): 2929-2939.

[52] Standards for Privacy of Individually Identifiable Health Information Final Rule. Dates: This final rule is effective on October 15, 2002. Entry Type: Rule; Page: 53182-53273 (92 pages); Document Citation: 67 FR 53182

[53] Centers for Disease Control and Prevention. About BMI for adults [Internet]. Atlanta, GA: Centers for Disease Control and Prevention; [updated 2011 Feb 15; cited 2011 Feb 16]. Available from: http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html.

[54] Lord GM, Matarese G, Howard JK**,** et al. Leptin modulates the T-cell immune response and reverses starvation-induced immunosuppression. Nature. 1998 Aug 27; 394(6696): 897-901.

[55] Bornstein SR, Licinio J, Tauchnitz R, et al. Plasma leptin levels are increased in survivors of acute sepsis: associated loss of diurnal rhythm, in cortisol and leptin secretion. J Clin Endocrinol Metab. 1998 Jan; 83(1): 280-3.

[56] Fiorentino DF, Zlotnik A, Mosmann TR, et al. IL-10 inhibits cytokine production by activated macrophages. J Immunol. 1991 Dec 1; 147(11): 3815-22.

[57] Akinnusi ME, Pineda LA, El Solh AA. Effect of obesity on intensive care morbidity and mortality: A meta-analysis. Crit Care Med. 2008 Jan; 36 (1) 151-158.

[58] Wunsch H, Linde-Zwirble WT, Harrison DA, Barnato AE, Rowan KM, Angus DC. Use of intensive care services during terminal hospitalizations in England and the United States. Am J Respir Crit Care Med. 2009 Nov 1;180(9):799-800. PMID: 19713448.

[59] Knaus WA, Wagner DP, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ. Lotring T, Damiano A. Chest. 1991;100;1619-1636. PMID: 1959406.

[60] Hosmer D, Lemeshow S. Applied Logistic Regression. 2nd ed. John Wiley & Sons, Inc.; 2000. 383p.

[61] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301-10.

[62] Codd EF. Relational Database: A Practical Foundation for Productivity: 1981 Turing Award Lecture. Association for Computing Machinery. 1981 Nov 9. Available from: http://awards.acm.org/images/awards/140/articles/2485527.pdf

[63] Ingersoll G. NoSQL, Lucene and Solr [Internet]. San Mateo(CA): Lucid Imagination. 2010 Apr 30 [cited 2011 Mar 28]. http://www.lucidimagination.com/blog/2010/04/30/nosql-lucene-and-solr/

[64] Chang, F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (Google Inc.). Bigtable: A distributed storage system for structured data. 7th Usenix Symposium on Operating Systems Design and Implementation (OSDI). 2006 Nov 6-8. Available from: http://labs.google.com/papers/bigtable-osdi06.pdf

[65] Ingersoll G. NoSQL, Lucene and Solr [Internet]. San Mateo(CA): Lucid Imagination. 2010 Apr 30 [cited 2011 Mar 28]. http://www.lucidimagination.com/blog/2010/04/30/nosql-lucene-and-solr/

[66] Neamatullah I, Douglass MM, Lehman LW, Reisner A, Villarroel M, Long WJ, Szolovits P, Moody GB, Mark RG, Clifford GD. Automated de-identification of free-text medical records. BMC Med Inform Decis Mak. 2008 Jul 24;8:32. PMID: 18652655.

[67] Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L.Rapidly retargetable approaches to de-identification in medical records. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):564-73. Epub 2007 Jun 28. PMID: 17600096

[68] Roberts A, Gaizauskas R, Hepple M, Guo Y. Combining terminology resources and statistical methods for entity recognition*:* an evaluation*.* Proc of the 6th Intern Conf on Language Resources and Evaluation (LREC 2008), European Language Resources Association (ELRA) 2008 May 28. ISBN: 2-9517408-4-0. Available from: ftp://ftp.dcs.shef.ac.uk/home/robertg/papers/lrec08-entityrec.pdf

[69] LingPipe 4.0.1 [Internet]. New York(NY): Alias-i. c.2008 [cited 2011 Mar 28]. Available from: http://alias-i.com/lingpipe

[70] Carpenter, Bob. LingPipe for 99.99% Recall of Gene Mentions. Proceedings of the 2nd BioCreative workshop. Valencia, Spain. 2007. Available from: http://www.colloquial.com/carp/Publications/biocreative-8-alias-i.pdf

[71] Computerisation of personal health records. Health Visit. 1978 Jun; 51(6): 227. PMID 248054.

[72] Silvestre AL, Sue VM, Allen JY. If you build it, will they come? The Kaiser Permanente model of online health care. Health Aff. 2009 Mar-Apr; 28(2): 334-344.

[73] Weingart S, Rind D, Tofias Z, Sands DZ. Who uses the patient internet portal? The PatientSite experience. J Am Med Inform Assoc. 2006 Jan-Feb; 13(1): 91-95.

[74] Riva A, Mandl KD, Oh DH, Nigrin DJ, Butte A, Szolovits P, Kohane IS. The personal internetworked notary and guardian. Int J Med Inform. 2001 Jun;62(1):27-40. PMID: 11340004

[75] Mandl KD, Kohane IS. Tectonic shifts in the health information economy. NEJM. 2008 Apr 17;358(16):1732-37.

[76] IIS: HL7 Standard Code Set. CVX -- Vaccines Administered [Internet]. Atlanta, GA: Centers for Disease Control and Prevention; [updated 2011 Feb 15; cited 2011 Mar 24]. Available from: http://www2a.cdc.gov/nip/IIS/IISStandards/vaccines.asp?rpt=cvx

[77] Office of the National Coordinator for Health Information Technology. Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule, 45 C.F.R. Part 170. (2010). Available from: http://edocket.access.gpo.gov/2010/pdf/2010-17210.pdf

[78] National Library of Medicine. MedlinePlus. Bethesda, M.D.: National Library of Medicine; [updated 2011 Mar 17; cited 2011 Mar 24]. Available from: http://www.nlm.nih.gov/medlineplus/.

[79] Tufte E. Sparkline theory and practice [Internet]. ET Notebooks (Ask ET); 2004 May 27 [updated 2011 Feb 25; cited 2011 Mar 25]. Available from: http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001OR

Mentoring

Since becoming the Director of the Lister Hill National Center for Biomedical Communications, I have mentored the following people:

Name	Dates at LHC	Type of Fellowship/ Project at LHC	Came from (Position/ Institution)	Went on to (Position/Institution)
Bruce Bray	9/2007-12/2008	Visiting Scientist/ PHR; Image DBs	Professor/ U. Utah	Professor/ U. Utah
Jeffrey Friedlin	3/2007-5/2007	Postdoctoral Fellow/ De-identification	Ph.D. student/ Regenstrief Institute	Asst Professor/ Regenstrief Institute
Emily Maxson	10/2009-7/2010	Medical Student/ Project at HHS Office of the Secretary	Medical Student/ Duke U.	Medical Student/ Duke U.
Howard Ching	4/2010-5/2010	Medical Student/ MIMIC II (Large Medical Database)	Medical Student/ Tufts U	Anesthesiology Residency/ NYU
Paula Scariati	6/1/2010-7/23/2010	NLM Rotation/ MIMIC II (Large Medical Database)	Graduate Student (fellowship) /Oregon Health Sciences U	Continued studies as Graduate Student /Oregon Health Sciences U
Swapna Abhyankar	2/2010-present	Clinical Postdoctoral Fellow/ PHR; MIMIC II; and newborn screening	Practicing pediatrician in Maryland / M.D. from U. Michigan	Currently 2nd-year NLM Clinical Postdoctoral Fellow; Continues clinical pediatric practice (part-time)

Clinical Informatics Training Sessions

I hold weekly clinical informatics training sessions that are attended by the postdoctoral fellows I mentor, as well as medical students serving in clinical electives at Lister Hill Center, and physicians at LHC for fellowships or rotations (many of whom also have other LHC mentors):

Name	Arrive	Depart	Medical School	Went on to	Mentor	LHC Project	Role at LHC
Xia Jing	3/1/2010	3/2/2012	U of Salford	In progress at NLM	Cimino	Infobutton	Post-M.D. Fellow
Swapna Abhyankar	2/1/2010	2/2/2012	U Michigan	In progress at NLM	McDonald	PHR, NBS	Post-M.D. Fellow
Christopher Miller	11/17/2008	11/17/2011	Case Western	In progress at NLM	Rindflesch/Hanna	NLP	Post-M.D. Fellow
Alex Gavino	10/18/2010	10/18/2011	U. Philippines	In progress at NLM	Fontelo	Mobile Computing	Post-M.D. Fellow
Marc Curvin	12/6/2010	12/30/2010	All Saints U Medicine, Baltimore	Internal med residency (St. Elizabeth Med Ctr, KY)	Rindflesch	NLP	Medical Student Clinical Elective
Michael Cairelli	10/25/2010	12/17/2010	Philadelphia Col Osteo Med	Philadelphia Col Osteo	Rindflesch	NLP	Medical Student Clinical Elective
Lincoln Sheets	11/3/2010	11/25/2010	U Missouri	U. Missouri	Fontelo	EBM	Med Stud Clin Elective
Lu Wang	6/3/2010	8/31/2010	Harvard	Harvard	Ciminio	Infobutton	Post-M.D. NLM Rotation Student
Ryan Banez	8/24/2009	8/24/2010	U Philippines	U. Philippines, faculty	Fontelo	mobile computing	Post-M.D. Fellow

Paula Scariati	6/1/2010	7/23/2010	Oregon Health Science	Cont. at OHSU	Demner-Fushman	Large DB Discovery	Post-M.D. NLM Rotation Student
Emily Maxson	10/2009	7/2010	Duke U	Continued Med School at Duke U	McDonald	Project at HHS OS ONC	Medical Student
Howard Ching	4/1/2010	5/18/2010	Tufts U	Residency Anesthesiology NYU	McDonald/ Demner-Fushman	Large Med DB	Medical Student Clinical Elective
Alex Iannone	8/3/2009	12/1/2009	Des Moines U	U Iowa; Internal Med Residency	Fontelo	mobile computing	Medical Student Clinical Elective
Vivienne Zhu	6/8/2009	8/7/2009	Regenstrief	Regenstrief	Fung	RxTerms	Post-M.D. NLM Rotation Student
Olayinka Ajayi	6/8/2009	7/31/2009	Johns Hopkins	Johns Hopkins; Nigeria NGO	Fontelo	Telemedicine	Post-M.D. NLM Rotation Student
Tsvi Aranoff	1/21/2009	5/31/2009	New York	FDA medical officer	Rindflesch	NLP	Post-M.D. NLM Rotation Student
Erick Ducut	3/1/2007	5/15/2009	U Philippines	U. Philippines, faculty	Fontelo	Interactive Medline	Post-M.D. Fellow
Lucas McCarthy	3/2/2009	4/24/2009	Albert Einstein	Stanford U-resident IM	Demner-Fushman	Text Extraction EMR	Medical Student Clinical Elective
Elizabeth Campbell	1/5/2009	3/31/2009	Johns Hopkins	Johns Hopkins	Fung	Medical Vocab	Post-M.D. NLM Rotation Student
Krystl Haerian	5/19/2008	2/27/2009	UMBC	Columbia, MS program	Cimino	EMR	Post-M.D. Fellow
Kevin Lai	8/4/2008	8/29/2008	U. Kentucky	U. Kentucky	Fung	RxTerms	Medical Student Clinical Elective
Caroline Wright	2/28/2008	4/18/2008	Georgetown U	Georgetown U	Lynch	PHR	Medical Student Clinical Elective
Kevin Maloy	3/1/2008	3/31/2008	Georgetown U	Georgetown, EMR resident	Fung	PHR	Medical Student Clinical Elective
Sneha Thakkar	2/4/2008	3/28/2008	U. Maryland	U. Maryland	Rindflesch	NLP	Medical Student Clinical Elective
Michael Nguyen	11/5/2007	11/30/2007	U. Philippines	U Philippines	Browne	De-identification	Medical Student Clinical Elective
Cynthia Luk	8/27/2007	9/21/2007	McGill U	McGill U	Lynch	PHR	Medical Student Clinical Elective
Sergio Leon	7/1/2006	7/1/2007	Prince Georges Hospital	U. Mass, Rheumatology residency	Fontelo	Wireless EB CIS	Post-M.D. Fellow
Caroline Ahlers	3/27/2006	5/31/2007	Ross U Medicine	Residency Georgetown U psychiatry	Rindflesch	NLP	Post-M.D. Fellow
Jeffrey Friedlin	3/12/2007	5/5/2007	Regenstrief	Faculty; Regenstrief	McDonald	De-identification	Post-M.D. NLM Rotation Student
Dina Demner-Fushman	11/1/2004	5/1/2007	U Maryland, College Park	NLM as researcher	Hauser	Q&A research	Post-M.D. Fellow
Sylvia Park	2/5/2007	4/30/2007	Johns Hopkins U	JHU residency	Fontelo	Consumer Health	Post-M.D. NLM Rotation Student
Mingchih Kao	2/5/2007	3/30/2007	U. Michigan	Yale U internship & residency	Zarin	ClinicalTrials	Medical Student Clinical Elective
Vinod Chacko	2/12/2007	3/11/2007	Drexel U	Drexel U	Hauser	PubMed M.D.	Medical Student Clinical Elective
Vivian Lee	2/5/2007	3/2/2007	Vanderbilt	U. Chicago, internship & residency	Aronson	IR	Medical Student Clinical Elective
Johnny Mei	1/4/2007	1/31/2007	Ben Gurion	OHSU; Med. Informatics Grad	Fontelo	Wireless EBM	Post-M.D. NLM Rotation Student
Michael Muin	9/13/2004	11/30/2006	Lyceum-Northwestern U	U. Philippines, faculty	Fontelo	Interactive Medline	Post-M.D. Fellow

Curriculum Vitae

Clement Joseph McDonald, M.D.

Director, Lister Hill National Center for Biomedical Communications

Office: Lister Hill Center, 8600 Rockville Pike, Bldg. 38A / Rm 7N707, Bethesda, M.D. 20894
(301) 496-4441

Education/Training:

1970 - 1972	Resident, Internal Medicine, Cook County Hospital and University of Wisconsin
1968 - 1970	Fellow, National Institutes of Health. Managed development of the first clinical laboratory computer system at the clinical center in Bethesda, M.D.
1967 - 1968	M.S., Northwestern University, Biomedical Engineering. Focus on computers and mathematics. Thesis: Computer Diagnosis of the Acute Abdomen by Computer Pattern Recognition Methods
1965 - 1966	Intern, Internal Medicine, Boston City Hospital, Harvard Medical Service
1964 - 1965	M.D., University of Illinois. First in class rank.
1958 - 1961	B.S., University of Notre Dame. Cum laude. Completed 132 semester-hour degree in three years

Positions:

2006 - Present	Director, Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health (NIH)
2006 - Present	Scientific Director, Lister Hill National Center for Biomedical Communications
1997 - 2006	Director, Regenstrief Institute for Health Care
1993 - 1997	Co-Director, Regenstrief Institute for Health Care, Indianapolis, Indiana.
1989 - 1998	Associate Administrator, Wishard Memorial Hospital
1972 - Present	Medical staff, Wishard Memorial Hospital
1976 - 2006	Chief, Computer Science Research Group, Regenstrief Institute for Health Care, Indianapolis, Indiana
1976 - 1993	Medical staff, Veterans Affairs Medical Center

Professional Activities:

2010 - Present	Member, NIH Office of Rare Diseases Research (ORDR) Global Rare Disease Registry (GRDR) Workgroup, Common Data Elements (CDE) Steering Committee
2010 - Present	Member, CDC Electronic Laboratory Reporting (ELR) Standards Working group
2010 - Present	Member, HHS Secretary's Advisory Committee on Heritable Disorders in Newborn and Children (SACHDNC) Health Information Technology Workgroup (HIT WG)
2009 - Present	Member, HHS Health Information Technology (HIT) Standards Committee Task Force on Vocabulary
2009 - Present	Member, Healthcare Informatics Advisory Board, FNIH/FDA Observational Medical Outcomes Partnership (OMOP)
2008 - Present	Member, NIH Clinical And Translational Science Awards (CTSA) Informatics Key Function Committee (IKFC) Directors and Group Leads
2007 - Present	Member, Trans-NIH Biomedical Informatics Coordinating Committee (BMIC)
2007 - Present	Member, NIH Clinical Center Clinical Research Information System (CRIS) Steering Committee
2006 - 2006	Regent, National Library of Medicine (NLM)
2005 - 2005	Member, Joint Commission's Health Information Technology Advisory Panel (HITAP)
2005 - 2005	Member, NCQA's Research Advisory Council (RAC)

2004 - 2005 Member, American Medical Association (AMA) E-Medicine Advisory Committee (EMAC)
2003 - 2008 Regent, American College of Physicians (ACP)
2003 - 2006 Member, Annals of Internal Medicine's Editorial Board
2002 - 2003 Member, Board of Directors, Clinical Data Interchange Standards Consortium (CDISC)
2001 - 2002 Member, Indiana University Cancer Center
2001 - 2003 Member, Board of Directors, Institute for Medical Knowledge Implementation (IMKI)
2000 - 2003 Committee Member, Robert Wood Johnson Foundation, *Pursuing Perfection: Raising The Bar For Healthcare Performance* Grant
2000 - 2002 Co-Chair HL7, HIPAA Attachment Sig
1996 - 2004 Member, National Committee on Vital and Health Statistics (NCVHS)
1996 - 1997 Member, The Koop Foundation's Action Team and Leadership Council
1995 - 1997 Member, American Medical Association's Education Advisory Committee
1995 - 1999 Member, International Union of Pure and Applied Chemistry, Commission on Nomenclature, Properties and Units (VII.C.1)
1995 - 1997 Member, Joint Commission Council on Performance Measurement
1995 - 1998 Member, VHA (Volunteer Hospital Association) Information Executive Council
1995 - 1996 Member, HCFA ICD-10-PCS Advisory Panel
1994 - Present Chairman, LOINC Committee
1994 - Present Member, Association of American Physicians
1994 - 1996 Member, Editorial Advisory Board, Journal of Health Data Management
1994 - 1996 Member, European Committee for Standardization (CEN) TC251 WG3 (Healthcare Communications and Messages) PT3-022 Request and Report Messages for Diagnostic Service Departments
1994 - 1996 Chairman, Advisory Panel for the Information Technology and Health Care System, for the Office of Technology Assessment of the United States Congress
1992 - 1995 Chairman, American National Standards Institute (ANSI), Health Information Standards Planning Panel (HISPP)
1992 - 1995 Chairman, American National Standards Institute (ANSI), Message Standards Developers Subcommittee (MSDS)
1991 - 1997 Member, Board of Directors, InterStudy, Excelsior, Minnesota
1992 - 1993 President, American Medical Informatics Association (AMIA).
1990 - 1993 Member, Dept. of Defense Peer Review Group, BATTELLE Corporation Health Care Systems
1990 - 1992 Member of the Committee on Clinical Practice Guidelines, Institute of Medicine
1989 - 1990 Chairman, American College of Physicians, Medical Informatics Subcommittee
1989 - 1990 Member, American College of Physicians, Educational Policy Subcommittee
1989 - 2000 Associate Editor, *M.D. Computing*, a Springer-Verlag International Journal
1987 - 2003 Co-Chairman, Founding Chairman, Health Level-Seven, Orders/Observations Technical Committee
1985 - 1996 Chairman, ASTM E-31.11 Subcommittee for Developing Standards for Electronic Transmission of Clinical Data
1985 - 1988 Secretary, Executive Board, Symposium on Computer Applications in Medical Care
1984 - 1986 Member, Executive Board, American College of Medical Informatics
1983 - 1989 Founding Editor, *M.D. Computing*, a Springer-Verlag International Journal
1977 - 1981 Member, Technology Study Section, National Center for Health Services Research

Academic Appointments:

1999 - 2006 Regenstrief Professor of Medical Informatics, Indiana University School of Medicine

1992 - 2006 Distinguished Professor of Medicine, Dept. of Medicine, Indiana University School of Medicine
1981 - 1992 Professor of Medicine, Department of Medicine, Indiana University School of Medicine
1974 - 1981 Associate Professor of Medicine, Department of Medicine, Indiana University School of Medicine
1972 - 1974 Assistant Professor of Medicine, Department of Medicine, Indiana University School of Medicine

Journal Reviewer:

The New England Journal of Medicine
Annals of Internal Medicine
Journal of the American Medical Association
Medical Care
The Journal of General Internal Medicine
The National Library of Medicine
ACP (American College of Physicians) Journal Club
Journal of American Medical Informatics Association

Grant Reviewer:

National Institutes of Health
Robert Wood Johnson Foundation

Specialty Board Status:

1972 Diplomat, American Board of Internal Medicine

Licensure and Certifications:

1972 Medical License #003848, Indiana
1965 Medical License #036-040726, Illinois

Honors:

2010 NIH Directors' Award
2009 Indiana Univesity President's Medal for Excellence
2006 Glenn W. Irwin, Jr., M.D. Research Scholar Award
2005 Honoree, Hoosier Heritage Lifetime Achievement Award 2005
2004 Morris Collen Award, MedInfo 2004, American Medical Informatics Association
2003 Indiana Business Journal Health Care Heroes Award
2002 American Medical Informatics Association (AMIA) Presidents' Award
2002 Who's Who in Health Care, Indianapolis Business Journal
2002 Association of Medical Directors of Information Systems, AM.D.IS Achievement Award
2001 Outstanding Researcher Award, GIM/Geriatrics Division Awards for 2001, Indiana University School of Medicine
2000 Hickam Award, Annual Mid-West SGIM Meeting
1994 - Present Member, National Academy of Science's Institute of Medicine
1994 - Present Member, Council of the Association of American Physicians
1993 American Health Informatics Management Association, Distinguished Service Award
1993 Blue & Company/Hudson Institute, Horizon Award
1993 Fellow, American Institute for Medical and Biological Engineering
1992 Distinguished Professor of Medicine, Indiana University
1991 Computers in Health Care, Pioneer Award

1991	Hospital Information Management Systems Society, Award for Outstanding Achievement
1984	Fellow, American College of Physicians
1988	Founding Fellow, American Medical Informatics Association

Patient Care Service:

1972 - 2006	Attending, General Internal Medicine Inpatient Service, Wishard Memorial Hospital
1972 - 1990	Attending, General Internal Medicine Inpatient Service, Veterans Affairs Medical Center
1972 - 2006	Attending, General Medicine Clinic, Regenstrief Health Center, Wishard Memorial Hospital
1972 - 1995	Attending, Emergency Room Medicine Service, Wishard Memorial Hospital

Developed and studied the effect of computer-stored medical record, sharing expression rule-based physician reminder and physician order-entry systems on the process of medical care. Developed database management systems, automated clinical laboratory, pharmacy and appointment scheduling systems. Fostered and stimulated and codes for standards for electronic exchange of clinical information between independent computer systems.

Publications (2006-Present; excerpted from a set of 283 publications since 1968):

2011

McDonald CJ, Abhyankar S. Clinical Decision Support and Rich Clinical Repositories: A Symbiotic Relationship: Comment on "Electronic Health Records and Clinical Decision Support Systems." Arch Intern Med, Jan 2011. PMID: 21263079.

Mitra R, Lee J, Jo J, Milani M, McClintock J, Edenberg H, Kesler K, Rieger K, Badve S, Cummings O, Mohiuddin A, Thoas D, Luo X, Juliar B, Li L, Mesaros C, Blair I, Srirangam A, Kratzke R, McDonald CJ, Kim J, Potter D. Prediction of Postoperative Recurrence-Free Survival in Nonsmall Cell Lung Cancer by Using an Internationally Validated Gene Expression Model. Clin Cancer Res 2011 [Epub ahead of print]. PMID: 21242119

2010

Lin MC, Vreeman DJ, McDonald CJ, Huff SM. Correctness of Voluntary LOINC Mapping for Laboratory Test in Three Large Institutions. AMIA Annu Symp Proc. 2010; 2010**:** 447–451**.** PMID21347018: PMC3041457

Abhyankar S, Lloyd-Puryear MA, Goodwin RM, Copeland S, Eichwald J, Therrell B, Zuckerman AE, Dowing G, McDonald CJ. Standardizing Newborn Screening Results for Health Information Exchange. AMIA Annu Symp Proc. 2010; 2010: 1–5. PMID21346929: PMC3041276

Vreeman DJ, McDonald CJ, Huff SM. Representing Patient Assessments in LOINC. AMIA Annu Symp Proc. 2010; 2010**:** 832–836. PMID21347095: PMC3041404

Friedlin J, McDonald CJ. An Evaluation of medical knowledge contained in Wikipedia and its use in the LOINC database. J Am Med Inform Assoc 2010; 17:283-287. PMID:20442145 : PMC2974620.

Lin MC, Vreeman DJ, McDonald CJ, Huff SM. **A** Characterization of Local LOINC Mapping for Laboratory Tests in Three Large Institutions. Methods Inf Med 2010 Aug 20; 46(5). PMID20725694: PMC 3034110.

Fung KW, McDonald CJ, Srinivasan S. The UMLS-Core Project A Study of the Problem List Vocabularies Used in Large Health Care Institutions. J Am Med Inform Assoc 2010 Nov 1;17(6):675-80. PMID20962130 :PMC3000762

Downs SM, van Dyck PC, Rinaldo P, McDonald C, Howell RR, Zuckerman A, Downing G. Improving Newborn Screening Laboratory Test Ordering and Results Reporting Using Health Information Exchange. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):13-8. PMID20064796 : PMC2995628

2009

McDonald CJ. Perspective: Protecting Patients in Health Information Exchange: A Defense of the HIPAA Privacy Rule. Health Affairs (2009); 28(2): 447–449. PMID19276002 : PMC2953709.

Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-72. PMID19683066 : PMC2757540.

Simonaitis L, McDonald CJ. Using National Drug Codes and Drug Knowledge Bases to Organize Prescription Records from Multiple Sources. Am J Health Syst Pharm. 2009 Oct 1;66(19):1743-53. PMID19767382 : PMC2965522

2008

Fung KW, McDonald CJ, Bray BE. Rx Terms a Drug Interface Terminology Derived from RxNorm. AMIA Annu Symp Proc 2008 Nov 6:227-31. PMID18998891: PMC2655997

Hansen RA, Tu W, Wang J, Ambuehl R, McDonald CJ, Murray M.D.. Risk of Adverse Gastrointestinal Events from Inhaled Corticosteroids. Pharmacotherapy. 2008 Nov;28(11):1325-34.PMID18956992 : PMC2648528.

Friedlin FJ, McDonald CJ. A Software Tool for Removing Patient Identifying Information from Clinical Documents. J Am Med Inform Assoc. J Am Med Inform Assoc. 2008 Sep-Oct;15(5):601-10. Epub 2008 Jun 25.PMID18579831 : PMC2528047.

Kho AN, Dexter PR, Warvel Js, Belsito Aw, Commiskey M, Wilson SJ, Hui SL, McDonald CJ. An Effective Computerized Reminder for Contract Isolation of Patients Colonized or Infected with Resistant Organisms. Int J Med Inform. 2008 Mar;77(3):194-8. Epub 2007 Mar 29.PMID17398145 : PMC2974622.

Kho AN, Lemmon L, Commiskey M, Wilson SJ, McDonald CJ. Use of a Regional Health Information Exchange to Detect Crossover of Patients with MRSA between Urban Hospitals. J Am Med Inform Assoc. 2008 Mar-Apr;15(2):212-6. Epub 2007 Dec 20. PMID18096903 : PMC2274796.

Overhage JM, Grannis S, McDonald CJ. A Comparison of the Completeness and Timeliness of Automated Electronic Laboratory Reporting and Spontaneous Reporting of Notifiable Conditions. Am J Public Health. 2008 Feb;98(2):344-50. Epub 2008 Jan 2. PMID18172157 : PMC2376898.

2007

Drake TA, Braun J, Marchevsky A, Kohane IS, Fletcher C, Chueh H, Beckwith B, Berkowicz D, Kuo F, Zeng QT, Balis U, Holzbach A, McMurry A, Gee CE, McDonald CJ, Schadow G, Davis M, Hattab EM, Blevins L, Hook J, Becich M, Crowley RS, Taube SE, Berman J; Shared Pathology Informatics Network. A System for Sharing Routine Surgical Pathology Specimens Across Institutions: The Shared Pathology Informatics Network. Hum Pathol. 2007 Aug;38(8):1212-25. Epub 2007 May 8.PMID: 17490722.

Kho AN, Hui S, Kesterson JG, McDonald CJ. Which Observations from the Complete Blood Cell Count Predict Mortality for Hospitalized Patients? J Hosp Med. 2007 Jan;2(1):5-12.PMID: 17274042.

Mamlin BW, Overhage JM, Tierney WM, Dexter PR, McDonald CJ. Clinical Decision Support Within the Regenstrief Medical Record System. Book chapter – Clinical Decision Support Systems - Theory and Practice - Series: Health Informatics, Berner, Eta S. (Ed.) 2nd ed., 2007, 190-21.

2006

McDonald CJ, Blevins L, Dexter P, Schadow G, Hook J, Abernathy G, Dugan T, Martin A, Phillips R, Davis M. Demonstration of the Indianapolis SPIN Query Tool for De-identified Access to Content of the Indiana Network for Patient Care's (a Real RHIO) Database. AMIA Annu Symp Proc. 2006:1194. PMC78939605

Kroth PJ, Dexter PR, Overhage JM, Knipe C, Hui SL, Belsito A, McDonald CJ. A Computerized Decision Support System Improves the Accuracy of Temperature Capture from Nursing Personnel at the Bedside. AMIA Annu Symp Proc. 2006:444-8.PMID: 17238380 : PMC 1839332.

Biondich PG, Downs SM, Carroll AE, Shiffman RN, McDonald CJ. Collaboration Between the Medical Informatics Community and Guideline Authors: Fostering HIT Standard Development that Matters. AMIA Annu Symp Proc. 2006:36-40.PMID: 17238298 : PMC1839341.

Christensen JD, Hutchins GC, McDonald CJ. Computer Automated Detection of Head Orientation for Prevention of Wrong-Side Treatment Errors. AMIA Annu Symp Proc. 2006:136-40.PMID: 17238318 : PMC1839503.

Simonaitis L, Belsito A, Warvel J, Hui S, McDonald CJ. Extensible Stylesheet Language Formatting Objects (XSL-FO): a Tool to Transform Patient Data into Attractive Clinical Reports. AMIA Annu Symp Proc. 2006:719-23. PMID: 17238435: PMC 1839487.

Vreeman DJ, McDonald CJ. A Comparison of Intelligent Mapper and Document Similarity Scores for Mapping Local Radiology Terms to LOINC. AMIA Annu Symp Proc. 2006:809-13.PMID: 17238453 : PMC1839677.

Friedlin J, McDonald CJ. A Natural Language Processing System to Extract and Code Concepts Relating to Congestive Heart Failure from Chest Radiology Reports. AMIA Annu Symp Proc. 2006:269-73.PMID: 17238345: PMC1839544

McDonald, CJ. Computerization Can Create Safety Hazards: A Bar-coding Near miss. Ann Intern Med. 2006 Apr 4;144(7):510-6. Erratum in: Ann Intern Med. 2006 Aug 1;145(3):235. PMID: 16585665.

Tang PC, McDonald CJ. Computer-based patient-record systems. In Shortliffe EH and Cimino JJ (Eds), Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 3rd edition. Springer 2006: 447-475.

Thomsen RW, Riis A, Nørgaard M, Jacobsen J, Christensen S, McDonald CJ, Sørensen HT. Rising Incidence and Persistently High Mortality of Hospitalized Pneumonia: A 10-year Population-based Study in Denmark. J Intern Med. 2006 Apr;259(4):410-7.PMID: 16594909.

Curriculum Vitae

Swapna Abhyankar, M.D.

Clinical Informatics Postdoctoral Fellow, Lister Hill National Center for Biomedical Communications

Education/Training

Clinical Informatics Postdoctoral Fellowship, February 2010 to present
Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, M.D.

Internship and Residency in Pediatrics, July 2002 to June 2005
Children's National Medical Center, Washington, DC

Doctor of Medicine, August 1998 to June 2002
University of Michigan Medical School, Ann Arbor, MI

Bachelor of Science in Earth Systems, September 1991 to June 1995
Stanford University, Stanford, CA
Graduated with Distinction

Technical and Standards Knowledge

Cerner PowerChart, eClinicalWorks, Oracle SQL, basic Eclipse/Java, UMLS, Microsoft Office Suite, Website design/implementation. HL7 v2.x, LOINC, SNOMED CT, ICD-9-CM, and ICD-10-CM.

Clinical Practice Experience

Staff Pediatrician, August 2005 to January 2010 (full-time), February 2010 to present (part-time)
Primary Pediatrics, Laurel, M.D.
Provide outpatient primary care services to children from birth through 21 years of age. Also provide newborn and inpatient pediatric care at Holy Cross Hospital in Silver Spring, M.D. In addition to clinical duties, have designed, implemented, and continuously update the Primary Pediatrics website (www.primarypedsmd.com). Designed all new encounter forms including well child and sick visit forms as well as newborn and ADHD management forms, and implemented the same content as templates in eClinicalWorks during the transition from paper charts to the office electronic health records. Created and implemented a book exchange to provide books for our underserved patients.

Leadership Experience

Secretary-Treasurer, Department of Pediatrics, January 2009 to January 2010
Holy Cross Hospital, Silver Spring, M.D.

Chair, Newborn Committee, January 2009 to January 2010
Holy Cross Hospital, Silver Spring, M.D.

Professional Memberships

Fellow, American Academy of Pediatrics (Maryland Chapter; Council on Clinical Information Tech)

Licensure/Certifications

Board Certified in Pediatrics, 2006 to present
Maryland Medical License #D0062820, 2005 to present
Active Maryland Controlled Substance Number and Federal DEA Number, 2005 to present

Publications

Abhyankar S, Lynch P, Wang Y, Goodwin RM, Jenders RA, McDonald CJ. The L-PHR: a Standards-Based Personal Health Record for Your Family. AMIA Annu Symp Proc. 2011. *(Submitted)*

McDonald CJ, Abhyankar S**.** Clinical Decision Support and Rich Clinical Repositories: A Symbiotic Relationship. *Arch Intern Med.* 2011; 0(2011): 20105181-2. Epub 2011 Jan 24.

Abhyankar S, Lloyd-Puryear MA, Goodwin RM, Copeland S, Eichwald J, Therrell B, Zuckerman A, Downing G, McDonald CJ. Standardizing Newborn Screening Results for Health Information Exchange. Proceedings of the AMIA 2010 Annual Symposium; 2010 Nov 13-17; Washington, DC.; c2010.

Other Professional Experience

Physician Coach, August 2008 to November 2008
Holy Cross Hospital, Silver Spring, M.D.
Provided training for physicians in a wide variety of specialties at Holy Cross Hospital prior to and during the implementation of the new Genesis electronic health record system. Took the lead in preparing training materials for the Pediatrics Department.

Intern, Summer Biomedical Research Program, June 1999 to August 1999
University of Michigan, Ann Arbor, MI
Worked on an ongoing joint Michigan-Israel population-based study of colorectal cancer. Used epidemiologic methods for analyzing case/control studies to determine the relationship between metabolic polymorphisms, nutritional intake, and the development of colorectal cancer. Helped develop an Oracle database for the study data.

Database Consultant, Anatomical Donations Program, June 1999 to October 1999
University of Michigan, Ann Arbor, MI
Designed a Y2K-compliant Filemaker Pro database for the Anatomical Donations Program.

Acting Business Manager, Division of Emergency Medicine, February 1998 to June 1998
Stanford University, Stanford, CA
Worked with Division Chief on management of finances and administrative services. Oversaw professional fee billing operations. Managed transition to new billing company. Maintained communication with other hospital departments involved in Emergency Department (ED) operations achieve operational and patient care goals. Provided tech support for the ED computer network.

Clinic Assistant, Stanford Emergency Department, July 1997 to February 1998
UCSF Stanford Health Care, Stanford, CA
Served as a liaison between the ED, employers, insurance companies, and patients for Workers' Compensation claims. Helped educate the ED physicians on documentation requirements for Workers' Compensation cases. Participated on a committee for streamlining patient chart flow through the ED. Provided technical support for the Emergency Medicine computer network.

Research Assistant/Programmer Analyst, SFO Medical Service, April 1996 to June 1997
University of California, San Francisco, San Francisco, CA
Supported faculty and staff with research studies, by performing literature searches, analyzing data, and drafting grant proposals, abstracts, and related documents. Trained and supported staff on all computer functions and software applications; provided 24-hour on-call assistance for the same. Beta-tested ClinAssist patient information software. Designed and maintained SFO Medical Service web site.