

Regional and National Health Care Data Repositories

James H. Harrison, Jr, MD, PhD^{a,*},
Raymond D. Aller, MD^b

^a*Departments of Public Health Sciences and Pathology, University of Virginia, Suite 3181
West Complex, 1335 Hospital Drive, Charlottesville, VA 22908, USA*

^b*Automated Disease Surveillance Section, Acute Communicable Disease Control Program,
Los Angeles County Department of Public Health, 313 N. Figueroa Street,
Room 222, Los Angeles, CA 90012, USA*

Large-scale analysis of patient data, including data mining, has the potential to improve substantially the understanding of disease presentation, care delivery processes, and the variability of disease progression and response to therapy as they happen in the real world. The inherent limitations of clinical trials in this respect and the benefits of direct analysis of patient data were described by Feinstein [1] more than 20 years ago and have been reviewed more recently from the perspective of evidence-based medicine [2]. Direct patient data analyses gain power with increasing data volume, completeness, and accuracy and under conditions that limit selection bias. For this reason, data sets to address community health and health care process questions optimally are aggregated at the population level and cut across traditional socioeconomic categories and geographic divisions. Such comprehensive health-related population data sets have not been available previously in the United States because of complex social, political, and business-related barriers. Recently, the societal benefits of such data have been articulated clearly in a call for the creation of a national framework for the appropriate “secondary use of health data” [3] and a national health data warehouse [4]. These efforts may bear fruit in the future; in the meantime, useful although limited analyses have been performed against large health-related data sets created for other purposes, as described in more detail later.

* Corresponding author.

E-mail address: james.harrison@virginia.edu (J.H. Harrison).

Clinical laboratory data may play several important roles in aggregated health-related databases. Some of the early work in health data warehousing was pioneered from laboratory medicine by Altshuler [5,6] and Eggert and Emmerich [7]. Since that time, most laboratory medicine publications on data warehousing have focused on aggregating data from a single care provider, such as a hospital [8] or a regional organization [9,10], to be used for local analysis. Although these analyses can have substantial value for local laboratory operations [10], improvement of health care quality [11], and support of clinical research [12], the approach has been underutilized [13]. In databases that aggregate laboratory data with other health-related data, the content of laboratory data often is limited and targeted to a particular purpose. Even in administrative and other databases that may contain only laboratory orders or specific disease-targeted laboratory results, however, these records serve as important objective values that can confirm or question the accuracy of other values in the database such as diagnostic or procedure codes [14,15].

Existing large-scale health-related databases can be characterized broadly as Medicare/insurance claims databases, single-provider data warehouses, regional/national collaborative provider data repositories, and public health and government population databases. Most of these data sets were collected for purposes other than the delivery of health care, and they generally contain summary data derived from patients' health records rather than a comprehensive clinical record. Thus they suffer shortcomings and biases related to their design, purpose for data acquisition, and data selection and summarization procedures [14,16]. Nonetheless, they have been useful for obtaining a global view of particular aspects of health care and characteristics of patients in a number of large-scale studies. The strengths and weaknesses of these databases, and their relationships with clinical laboratories and laboratory data, are considered here.

Claims data

Medicare and large insurance claims databases can provide coverage of large patient populations. Within a claims database, records for patients can be linked longitudinally through patient identity, yielding a temporal view of patient care. The most useful data in these sets is generally based on physician, facilities, and pharmacy claims (Table 1) [17]. Physician claims contain American Medical Association Current Procedural Terminology (CPT) codes describing procedures performed by physicians and other care providers, International Classification of Disease (ICD, most commonly ICD-9-CM) codes describing diagnoses, and dates of service. Facilities claims contain ICD and CPT codes, including Health Care Common Procedure Coding System (HCPCS) codes representing services, supplies, therapeutics, and other items, and admit/discharge dates if applicable. Pharmacy claims include National Drug Code and/or HCPCS identifiers of

Table 1
Contents of common regional data sets

Data set	Typical contents ^a	Features
Claims data, from care providers and payers	ICD-9-CM codes ^b CPT codes ^c HCPCS codes ^d NDC codes ^e Charges	Charged items only Some diagnoses may be present only as “rule out” Results of observations and tests not included Time precision at the day level
Single and cooperative provider repositories	ICD/CPT codes Problem lists Clinical orders Observation/test results Procedure descriptions Textual case summaries	Full clinical detail with precise timestamps
Public health and governmental databases	Diagnoses/codes Observation/test results Follow-up data as specified, including geographic and environmental data Large-scale survey results	Detailed data confined to that mandated for reporting and follow-up or included in the survey

Abbreviations: CPT, Current Procedural Terminology codes; HCPCS, Healthcare Common Procedure Coding System; ICD-9, International Classification of Disease, edition 9; NDC, National Drug Code.

^a Actual contents will be site dependent, particularly for clinical repositories from providers.

^b Clinical diagnoses.

^c Physician and other provider procedures.

^d Services and supplies.

^e Therapeutics.

therapeutics, a therapeutic class categorization, and prescription date. Time-stamps generally provide only the day of the charge (which may not be the day of the clinical event) and thus do not support precise short-term calculations such as most pharmacokinetics evaluations or confirmation of a rapid response to an event. Claims databases have been used in a large number of health care process and outcomes studies (eg, Ref. [18]), and similar data are used by the Health Plan Employer Data and Information Set (HEDIS) to compare quality of care among providers [19].

A number of problems exist with claims data, and their use is controversial [15,16,20]. By their nature, claims records exist only for items that generate a financial transaction, and there is an incentive for the choice of codes to support the operation of the financial system. Many codes are added after patient care during a separate clinical records review, rather than by caregivers. Claims forms may provide a limited number of “slots” for codes, forcing code selection based on the needs of the current transaction. Diagnostic codes attached to procedures and laboratory tests may represent “rule-out” possibilities rather than confirmed diagnoses. Finally, the ICD coding system is not comprehensive for all clinical conditions and does

not support accurate description of differences in severity, presentation, or course among patients who have the same disease.

For these reasons, the sensitivity and specificity of claims data for identifying conditions in patients has been questioned. Results differ among diagnoses reviewed, although specificity is generally better than sensitivity (it is more common to fail to code an existing condition than to code a non-existent one [20]), and the coding of comorbidities seems particularly problematic [21]. Suboptimal sensitivity and/or specificity of coding have been reported for cardiovascular diseases [22,23], acute renal failure [24], pneumonia [25], pediatric asthma [26], migraine [27], gout [28], and mild traumatic brain injury [29], among others. Several investigators have reported improved sensitivity and specificity for identifying particular conditions using algorithms that categorize patients based on multiple criteria in claims databases rather than by the presence or absence of specific codes [17,30]. In general, claims data should be used carefully, and the results should be interpreted conservatively, with an understanding of the biases and inaccuracies inherent in the data sets.

Single-provider data warehouses

Many health care organizations have or are developing local research databases, but only a few large, integrated health care organizations and national laboratories are able to marshal data sets that approach large claims databases in size. Data sets from single health care organizations can contain detailed clinical data that are linked through patient identity and thus present a longitudinal view of patient care (see Table 1). National laboratory data may not guarantee unique identification of patients outside specific testing transactions and thus are less useful for longitudinal studies. Because of the increased detail offered by these data sets, they may support a substantially more comprehensive analysis from a medical perspective than claims databases. Access to the data, however, may be limited to internal researchers whose activities are directed according to organizational interest. Barriers to general analysis of such data sets include a lack of financial incentive for the organization to support and maintain data access and processing for arbitrary projects, a lack of financial incentive to support governmentally mandated requirements for research use of patient data, and the perception of organizational risk related to loss of patient confidentiality even if explicit patient identifiers are removed.

Several large organizations have published analyses from their internal data sets. Kaiser Permanente (www.kaiserpermanente.org), which covers a substantial number of patients in the Western United States, has implemented an organizational data warehouse [31] that identifies patients for inclusion in a Cardiovascular Risk Factor Management Program [32] and is being used to standardize clinical care across diverse patient populations [33]. Large national laboratories have used data warehouses consisting of

test orders, diagnoses, and results to calculate population-based reference ranges [10] and to support infectious disease surveillance [34]. The US Veterans Administration (VA), which supports what is perhaps one of the world's largest integrated electronic health records, maintains a centralized database into which data from local VA health care sites can be aggregated for analysis. Although this database is not yet comprehensive, large targeted data sets such as the Diabetes Epidemiology Cohort have been analyzed to support improved patient care, identify process problems, and target appropriate education to clinicians [35].

Regional and national collaborative provider databases

Most health care providers are not large enough to create regional clinical repositories individually but could build similar databases through cooperative data-sharing agreements involving multiple providers (see Table 1). The utility of these regional databases, along with the need for sharing clinical and administrative data between the multiple parties typically involved in patient care, led to a proposal some years ago for the development of cooperative community health information networks (CHINs). Although there have been a few notable successes in developing community-based data warehouses [36], these are unusual, and most initial work in CHIN development ultimately did not come to fruition for a variety of technical, financial, administrative, and political reasons [37–39]. The need for sharing health information to improve care delivery remains, however, and has driven the development of a modified successor of CHINs termed “regional health information organizations” (RHIOs) [39,40]. A primary goal of the RHIO approach is the communication of clinical information across organizations to support patient care. This goal may be met either through establishment of an accessible central database of shared information or by peer-to-peer communications without a central database; examples of both approaches exist [40]. When actual or federated virtual databases are created, RHIOs could provide an excellent foundation for data mining. Only time will tell whether the RHIO movement will be successful; several have been operating for many years, predating their identification as “RHIOs.” Many others have been formed only recently: the Massachusetts Health Data Consortium's national list of RHIOs included 53 organizations as of July 2006 [41].

Health care organizations also have formed voluntary data-sharing consortia based on implementation of particular information systems rather than on regional location. This approach has the advantage of reducing the challenges of reconciling data models and data representation across systems by adopting a vendor's proprietary designs as a standard for data sharing, but it also relies on the vendor's continued viability and may be problematic if sites change vendors. Furthermore, although this strategy increases the volume of data available for analysis, it does not capture

a comprehensive longitudinal care record within a region, because data transmitted to the central repository are de-identified (and thus data from one patient cannot be linked across care providers), and vendors generally do not have complete regional coverage. Thus vendor-based data aggregation is not likely to be satisfactory in the long term. Examples of this approach include hospitals using the ambulatory electronic medical record from GE Healthcare Information Technologies (Milwaukee, Wisconsin) [42] and the PedCath pediatric catheterization database (Scientific Software Solutions, Inc., Charlottesville, Virginia) [43].

The development of collaborative data sharing has been limited for a variety of reasons. Among the most important are a lack of clear financial incentives for institutions that bear the cost of creating and maintaining the data repositories and uncertainty concerning risks and benefits (eg, balancing patient confidentiality risks against organizational and societal benefits). As discussed in relation to data warehousing elsewhere in this issue, developing and maintaining methods for accurately communicating and cleansing data and for reconciling multiple data models and data representation strategies is demanding. Once created, data warehouses probably will support work that will lead to recommendations for improved care. This information may benefit society, but whether the called-for improvements in care would be financially beneficial or detrimental to health care organizations or would provide competitive advantage over other organizations that also may be participating in the data-sharing effort is unclear. In addition, linking patients' data longitudinally across multiple providers requires patient identity information of some type. Thus aggregated databases carry a risk of loss, albeit perhaps low, of patient confidentiality. If confidentiality were breached, the responsibility of the organization providing the data is not completely clear under current regulations; in any case, the perception of poor stewardship of health information would be detrimental. The incentive is limited for an institution to participate in a resource-intensive effort that has a low but real risk for a negative outcome and an uncertain likelihood of a positive outcome (for the institution). It is understandable that most aggregated databases to date are intraorganizational: data processing, communication, and maintenance issues are simpler, and analysis can be targeted to particular projects of interest to the organization. Until the incentives for provider data sharing can be brought into alignment with public needs, public mandates probably will remain the most effective method for aggregating clinical data from multiple organizations into useful form.

Public health and related databases

In nations with different approaches to health care financing and different regulations related to data stewardship and privacy, the creation of large-scale clinical repositories for analysis has proceeded more rapidly than in the United States. These projects have been initiated within health care

organizations or governmentally mandated at the regional or national level, or they may involve data sharing between nations. Examples include regional [44] and national [45] general-purpose clinical data warehouses, national [46] and international [47] disease-specific data warehouses, national public health databases [48], and international federated genome data sets for twin studies [49].

In the United States, reporting specific laboratory, diagnosis, and other data identified as important to public health has been mandated for many years, and these data are gathered into regional databases designed for public health surveillance and case follow-up (Fig. 1; see Table 1). These databases typically contain geographic distributions of symptoms or disease (Fig. 2) and may contain non-health care information from a variety of sources that is pertinent to understanding disease transmission (Fig. 3). The unique geographic aspects of these databases make them particularly useful for analysis using geographic information systems and spatial data mining [50].

In the past public health reporting has been a hierarchical process managed by local and state public health authorities, with manual entry of paper forms from health care providers into multiple disease- and program-specific database systems. These processes and systems are being replaced by electronic reporting and integrated data warehousing capabilities. A recent survey [51] indicated that 69% of state public health laboratories had integrated data management capabilities, and an additional 16% planned to acquire these capabilities by the end of the decade. States at the leading edge of this development are relatively advanced, with electronic reporting, data transfer collaborations with in-state RHIOs, and data warehousing at the state level [52].

Data collected by the states is ultimately reported to the US Centers for Disease Control (CDC), which supports active programs that aggregate the data for a variety of special purposes. For example, the Public Health Information Network [53] provides an architecture for reporting and responding to infectious disease and other health data that may signal public health emergencies. CDC Wonder (wonder.cdc.gov) is a system that provides access to a large selection of public health-related summary data based on the state reports. Special programs such as the Vaccine Adverse Event Reporting System collect data about a particular health-related issue that then are accessible for data mining [54,55]. Public health data reporting and databases are exempt from Health Insurance Portability and Accountability Act (HIPAA) privacy restrictions but have their own accessibility requirements. In general, detailed public health data are accessible only to public health practitioners, and thus data mining involves collaboration with appropriate public health officials.

A number of other special-purpose governmental databases containing survey data related to population health and health care services use (reviewed in Ref. [56]) or to Food and Drug Administration product surveillance [57] also may be available for analysis. All these public health and

County of Los Angeles • Department of Health Services • Public Health *Please post.*

REPORTABLE DISEASES AND CONDITIONS

Title 17, California Code of Regulations (CCR), § 2500

It shall be the duty of every healthcare provider, knowing of or in attendance on a case or suspected case of any diseases or conditions listed below, to report to the local health officer for the jurisdiction where the patient resides. Where no healthcare provider is in attendance, any individual having knowledge of a person who is suspected to be suffering from one of the diseases or conditions listed below may make such a report. "Healthcare provider" encompasses physicians, surgeons, veterinarians, podiatrists, nurse practitioners, physician assistants, registered nurses, nurse midwives, school nurses, infection control practitioners, medical examiners, coroners, dentists and chiropractors.

Urgency Reporting Requirements:
 ☐ = Report immediately by telephone.
 ☐ = Report by mailing, telephoning or electronically transmitting a report within 1 working day of identification of the case or suspected case.
 ☐ = Report by telephone within 1 hour followed by a written report submitted by facsimile or electronic mail within 1 working day.
 ☐ = Report within 7 calendar days from the time of identification by mail, telephone or electronic report.

REPORTABLE DISEASES

☐ Acquired Immune Deficiency Syndrome (AIDS) * ☐ Amebiasis ☐ Amnionitis ☐ Anthrax ☐ Babesiosis ☐ Botulism: Infant, Foodborne, or Wound ☐ Brucellosis ☐ Campylobacteriosis ☐ Chancroid * ☐ Chlamydial Infections * ☐ Cholera ☐ Ciguatera Fish Poisoning ☐ Coccioidiomycosis ☐ Colorado Tick Fever ☐ Conjunctivitis, Acute Infections of the Newborn, specify etiology ☐ Cryptosporidiosis ☐ Cysticercosis ☐ Dengue ☐ Diarrhea of the Newborn, outbreaks only ☐ Diphtheria ☐ Domestic Acid Poisoning (Anemic Shellfish Poisoning) ☐ Echinococcosis (Hydatid Disease) ☐ Ecthymosis ☐ Encephalitis, specify etiology: Viral, Bacterial, Fungal, Parasitic ☐ <i>Escherichia coli</i> O157:H7 Infections ☐ Foodborne Diseases ☐ 2 or more cases from separate households with same suspected source ☐ Giardiasis ☐ Gonococcal Infections * ☐ <i>Haemophilus influenzae</i>, Invasive Disease ☐ Hantavirus Infections ☐ Hemolytic Uremic Syndrome ☐ Hemorrhagic Fever, Viral (e.g., Crimean-Congo, Ebola, Lassa, and Marburg viruses)	☐ Hepatitis: ☐ Hepatitis A ☐ Hepatitis B, specify Acute or Chronic ☐ Hepatitis C, specify Acute or Chronic ☐ Hepatitis D (Delta) ☐ Hepatitis Other, Acute ☐ Human Immunodeficiency Virus (HIV) * ☐ Influenza, pediatric—ICU cases or deaths only ☐ Kawasaki Syndrome (Mucocutaneous Lymph Node Syndrome) ☐ Legionellosis ☐ Leprosy (Hansen's Disease) ☐ Leptospirosis ☐ Listeriosis ☐ Lyme Disease ☐ Lymphocytic Choriomeningitis ☐ Malaria ☐ Measles (Rubella) ☐ Meningitis, specify etiology: Viral, Bacterial, Fungal, or Parasitic ☐ Meningococcal Infections ☐ Mumps ☐ Non-Gonococcal Urethritis (report laboratory confirmed Chlamydia as Chlamydia) * ☐ Paralytic Shellfish Poisoning ☐ Pediatric Inflammatory Disease (PID) * ☐ Pertussis (Whooping Cough) ☐ Plague, Human or Animal ☐ Poliomyelitis, Paralytic ☐ Psittacosis ☐ Q Fever ☐ Rabies, Human or Animal ☐ Relapsing Fever ☐ Rubeola ☐ Rocky Mountain Spotted Fever ☐ Rubella: ☐ Acute Rubella (German Measles) ☐ Congenital Rubella Syndrome	☐ Salmonellosis (other than Typhoid Fever) ☐ SARS (Severe Acute Respiratory Syndrome) ☐ Scabies (Atypical or Crusted) * ☐ Scombroid Fish Poisoning ☐ Shigellosis ☐ Smallpox ☐ Streptococcal Infections: ☐ Outbreaks of any type ☐ Individual case in a food handler ☐ Individual case in a dairy worker ☐ Invasive Group A Streptococcal Infections including Streptococcal Toxic Shock Syndrome and Necrotizing Fasciitis * (Do not report individual cases of pharyngitis or scarlet fever.) ☐ <i>Streptococcus pneumoniae</i>, Invasive * ☐ Swimmer's Itch (Schistosomal Dermatitis) ☐ Syphilis * ☐ Tetanus ☐ Toxic Shock Syndrome ☐ Toxicoplasmiasis ☐ Trichinosis ☐ Tuberculosis * ☐ Tularemia ☐ Typhoid Fever, cases and carriers ☐ Typhus Fever ☐ Varicella: ☐ Varicella, Fatal Cases ☐ Varicella, Hospitalized Cases (Do not report cases of herpes zoster/shingles.) ☐ <i>Vibrio</i> Infections ☐ Water-associated Disease ☐ West Nile Virus ☐ Yellow Fever ☐ Yersiniosis
--	--	---

Notification Required of Laboratories (CCR § 2503)

☐ Anthrax ** ☐ Botulism ** ☐ Brucellosis ** ☐ Chlamydial Infections * ☐ Cryptosporidiosis ☐ Diphtheria * ☐ Encephalitis, arboviral ☐ <i>Escherichia coli</i> O157:H7 or Shiga toxin-producing <i>E. coli</i> O157:HNM * ☐ Gonorrhea * ☐ Hepatitis A, Acute Infections, by IgM antibody test or positive viral antigen test	☐ Hepatitis B: ☐ Acute Infections, by IgM anti-HBc antibody test ☐ Surface Antigen Positivity (specify gender) ☐ Hemorrhagic Fever, Viral (e.g., Crimean-Congo, Ebola, Lassa, and Marburg viruses) * ☐ Human Immunodeficiency Virus (HIV) * ☐ Listeriosis * ☐ Lyme Disease ☐ Malaria * ☐ Measles (Rubella), Acute Infections, by IgM antibody test or positive viral antigen test	☐ Plague, Animal or Human ** ☐ Rabies, Animal or Human ☐ Salmonella * ☐ Smallpox * ☐ <i>Streptococcus pneumoniae</i>, Invasive * ☐ Syphilis * ☐ Tuberculosis ** ☐ Tularemia ** ☐ Typhoid and other <i>Salmonella</i> Species * ☐ <i>Vibrio</i> Species Infections * ☐ West Nile Virus
---	---	---

* Reportable to the Los Angeles County Department of Health Services.
 ** Incubated isolates and isolates taken must be forwarded to the Los Angeles County DHS Public Health Laboratory for confirmation. Healthcare providers must still report all such cases separately.
 * Laboratories receiving specimens for the diagnosis of these diseases must immediately contact the California Department of Health Services, for bacterial testing call 510-412-3700, for viral testing call 510-307-4575. For botulism testing, contact Acute Communicable Disease Control at 213-240-7941.

Non-communicable Diseases or Conditions

☐ Alzheimer's Disease and Related Conditions (CCR § 2802, § 2806, § 2810) ☐ HIV Epidemiology Program 213-331-8318 www.lapublichealth.org/hiv/index.htm	☐ Disorders Characterized by Loss of Consciousness (CCR § 2806, § 2810) ☐ STD Program 213-744-3070 www.lapublichealth.org/std/index.htm	☐ Pesticide-Related Illnesses (Health and Safety Code, § 105200) ☐ TB Control Program 213-744-6271 (for reporting) 213-744-6160 (general) www.lapublichealth.org/tb/index.htm
---	--	--

To report a case or outbreak of any disease contact the Communicable Disease Reporting System Hotline
Tel: 888-397-3993 • Fax: 888-397-3778

(Rev. 10/05)

Fig. 1. Reportable diseases and health conditions, Los Angeles County, California. Reporting may be performed by submission of paper forms, telephone, or electronic communication. The submitted information, along with additional follow-up data, is incorporated into a regional public health data warehouse. (Courtesy of County of Los Angeles, Department of Health Services, Public Health, Los Angeles, CA.)

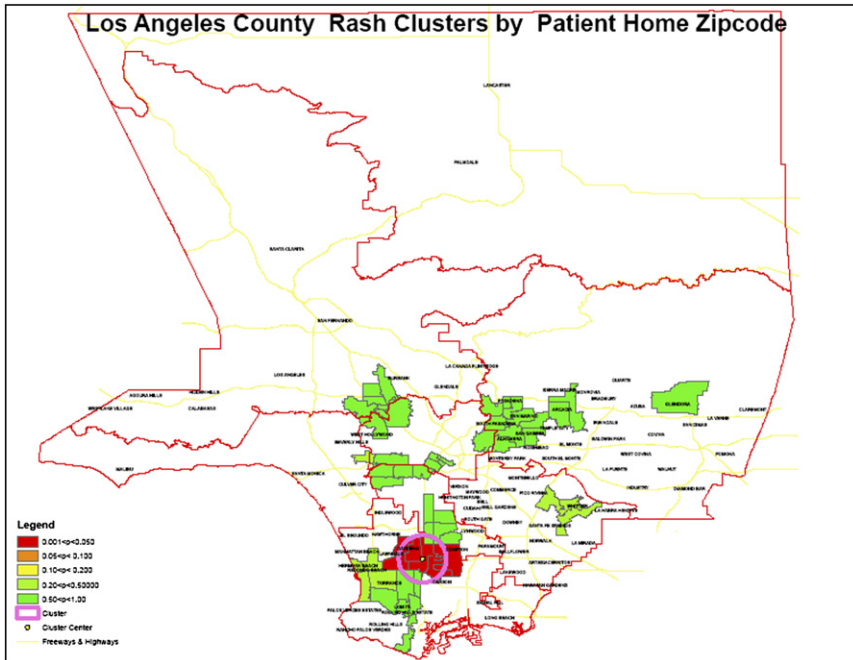


Fig. 2. Prevalence of a rash symptom in patients seen in emergency departments in Los Angeles County. Color coding indicates how statistically abnormal the rate is in each zip code of patient residence. The bright red area indicates a probable outbreak of a rash-related illness.

governmental data sets are limited, in that they are not complete health records and contain only data that are mandated for survey, reporting, or capture during follow-up. Thus they are not regional or national clinical repositories in the sense of some of the European systems, but they can be useful for addressing particular questions in appropriate domains.

Challenges inherent in regional databases

Regional data warehouses, whether they house clinical data from collaborating care providers or public health data, offer all the challenges previously described in this issue for data warehouse construction, with the added complexity of multiple independent data-contributing organizations. Furthermore, regional warehouses may receive useful data from organizations or locations distinct from traditional health care providers, for example environmental or veterinary laboratories (see Fig. 3), nursing homes or home health care providers, or remote sensors such as home blood-glucose monitors. Some users may be outside the traditional health care provider or health services researcher roles, and data presentation or analysis capabilities should support their needs. For example, public health data are important for highlighting geographic clusters of symptoms in biosurveillance (see Fig. 2)

Communications and data loading

Traditionally, national hospital discharge databases accumulated data in yearly chunks. Data processing and loading often took significant additional time, so data might be several years old by the time they became accessible for study. The applicability of conclusions based on such older data to current decisions was legitimately questioned. For a regional warehouse to be useful for business planning or process improvement, the authors have proposed that data be accumulated and distributed at least quarterly. For detection of disease outbreaks, data should be loaded at least daily and surveyed every day. Finally, if a database will be a hub that communicates patient data to care providers, it should be designed as a transactional system, with updates from contributing systems within a few minutes or less. Thus, depending on their purpose, regional systems can span a broader range of updating requirements than typical for intraorganizational data warehouses, which generally are updated weekly or monthly.

Data warehouses that are updated periodically may receive data as files that are manually or automatically downloaded from contributing systems and processed to an appropriate form for loading before or after electronic transmission by standard means such as secure File Transfer Protocol. Ongoing developments in standard data formats based on Health Level 7 (HL7, www.hl7.org) and Extensible Markup Language (XML, www.w3.org/XML) and the use of standard HL7 parsers and XML processing software may simplify these basic import-export steps, although the challenges described later related to data representation also will need to be addressed before a fully standard approach is practical. For more rapid updates related to biosurveillance and clinical communication, real-time transactional interfaces will be necessary. In the United States, most clinical data interfaces use the HL7 version 2 framework [58], which provides a transaction message structure but does not define its data content fully. Thus system interfaces using HL7 require negotiation and adaptation, and creating separate interfaces for each contributing system can be a significant expense in the construction of regional data warehouses. It is hoped that HL7 version 3 will be able to standardize clinical interfaces better and reduce their costs through definition of a common data model and data element representations in addition to a messaging framework [59]. The Clinical Data Interchange Standards Consortium (www.cdisc.org) and the Cancer Biomedical Informatics Grid (cabig.nci.nih.gov) are engaged in similar efforts to standardize data models and representation related to clinical research. As these efforts progress and coordinate, they ultimately may decrease substantially the expense and effort currently required to connect clinical information systems for data sharing.

Reconciliation of data representation

In the absence of a standard data representation specification such as that in HL7 version 3, data transferred using a standard messaging or file format

may represent values in multiple ways. For example, most laboratories have created their own unique sets of test codes and text mnemonics, and other data similarly may have site-specific representation. With a few exceptions, it generally is necessary to build a translation table into each source system, mapping its vocabulary to a standard target vocabulary used in the regional warehouse. The Logical Observation Identifier Names and Codes (LOINC) has established a well-accepted standard for laboratory test names [60], and organism names in microbiology can be converted to the Systematized Nomenclature of Medicine – Clinical Terminology (SNOMED CT [61]). The task of fully coding a laboratory database for LOINC and SNOMED nomenclature is significant, and most laboratories create mapping tables only for tests that are “exported.” This approach can cause problems, because each laboratory’s mapping table must be maintained as tests change and new tests are incorporated into the regional database. Failure to maintain the table appropriately can cause failure of data transmission and require retroactive identification and transmission of unreported results. In the future, as an increasing number of regulatory agencies and other organizations require the use of communication standards, laboratories should be able to provide conversion to standard terminologies as a part of routine data export. Similar data-representation considerations apply to other clinical systems that supply data to a regional warehouse.

Linkage of data by patient

A health information data warehouse that is optimally useful for understanding disease presentation and course as well as care processes should contain accurate longitudinal representations of disease, care, and therapeutic response for each patient. When data come from multiple care providers who share no common patient identification mechanism, data linkage requires transmission of some type of patient identifier [62]. These identifiers may be stripped from the data before the data are used for analysis, but it is necessary to maintain the identifiers and their link to the patient data in a secure environment in the data warehouse so that future data may be linked to data from the same patient that are already in the warehouse, irrespective of the data source. Although this architecture is used in regional and national European health data warehouses [45,46], the authors are not aware that it has been deployed successfully in a multiorganizational health data warehouse in the United States.

Importance, standardization, and normalization of clinical laboratory data

Laboratory data occupy an important position in health data warehouses. Much of the data in these databases is coded (eg, ICD-9-CM codes), and the codes typically are entered by people who are performing a rapid review of all or part of the medical record for the purpose of financial

transactions [16]. In databases that contain textual records, physicians' comments in text may be unclear, ambiguous, or misleading. Laboratory data are objective and clinically meaningful and generally are transmitted accurately. Laboratory data can serve as validation for correctly coded data and as a flag for incorrect codes. Laboratory data also can support risk stratification and inference of disease severity or unusual presentation, which are poorly represented in the currently common coding schemes [63]. Finally, with appropriate analysis techniques, laboratory data can allow correct classification of patients to conditions for which codes do not exist or when codes are omitted [64].

The aggregation of data from multiple laboratories into data warehouses offers several important challenges in addition to the previously mentioned differences in naming of test codes. Results that are expressed as categorical values or discrete scales may differ between laboratories based on different category names and scales. These values must be reconciled or standardized to a common set of values or scale. Many clinical laboratory tests that yield numeric values are not standardized to reference material, and the variability of results between test methodologies, and even between laboratories using the same methodology, has been well described [65–67]. The authors have found that many analytes can yield differences between laboratories in a range that would affect clinical interpretation and yield spurious associations in data mining. Merely comparing the results to reference ranges is insufficient for rigorous manual evaluation and is inadequate for data mining. Statistical methods to normalize result values between different laboratories are available and should be considered for general application across the data warehouse [68,69].

Summary

Although the United States lags behind Europe in creating large-scale, detailed health data repositories at the regional level, several types of repositories exist in this country that may provide data useful for particular data-mining applications. Regional repositories or data warehouses have special requirements and constraints that distinguish them from intraorganizational data warehouses. These differences are related to their intended purpose, the diversity of data they may receive, and their connection to multiple unrelated data providers. Public health databases differ further because they are exempt from HIPAA privacy requirements and operate under their own security and confidentiality mandates. As objective data that are highly likely to be transmitted correctly, clinical laboratory results play an important role in all these databases in validating other data, including diagnosis and procedure codes, and in supporting inference of disease severity and risk-related information that is poorly handled in coding systems. Aggregation of laboratory results from multiple data providers yields specific challenges related to representation and normalization of data that must be

addressed successfully for these databases to be optimally useful. As data communications related to public health continue to develop, more laboratories will become direct contributors to public health databases through electronic data transfer. If the United States is successful in current efforts to develop a general administrative and political framework for the secondary analysis of health data, laboratories also are likely to become important and valuable contributors to multiple regional databases targeting improvement in community health, better understanding of therapeutic response, and improvement in the medical process.

References

- [1] Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. *Ann Intern Med* 1983;99(4):544–50.
- [2] Grossman J, Mackenzie FJ. The randomized controlled trial: gold standard, or merely standard? *Perspect Biol Med* 2005;48(4):516–34.
- [3] Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. *J Am Med Inform Assoc* 2007;14(1):1–9.
- [4] Blewett LA, Parente ST, Finch MD, et al. National health data warehouse: issues to consider. *J Healthc Inf Manag* 2004;18(1):52–8.
- [5] Altshuler CH. Building a database for monitoring and facilitating health care. *Clin Lab Med* 1983;3(1):179–204.
- [6] Altshuler CH. Use of comprehensive laboratory data as a management tool. *Clin Lab Med* 1985;5(4):673–96.
- [7] Eggert AA, Emmerich KA. Long-term data storage in a clinical laboratory information system. *J Med Syst* 1989;13(6):347–54.
- [8] Kamal J, Rogers P, Saltz J, et al. Information warehouse as a tool to analyze Computerized Physician Order Entry order set utilization: opportunities for improvement. *AMIA Annu Symp Proc* 2003;336–40.
- [9] Maizlish NA, Shaw B, Hendry K. Glycemic control in diabetic patients served by community health centers. *Am J Med Qual* 2004;19(4):172–9.
- [10] Bock BJ, Dolan CT, Miller GC, et al. The data warehouse as a foundation for population-based reference intervals. *Am J Clin Pathol* 2003;120(5):662–70.
- [11] Bates DW, Pappius E, Kuperman GJ, et al. Using information systems to measure and improve quality. *Int J Med Inform* 1999;53(2–3):115–24.
- [12] Gilbertson JR, Gupta R, Nie Y, et al. Automated clinical annotation of tissue bank specimens. *Medinfo* 2004;11(Pt 1):607–10.
- [13] Aller RD. The clinical laboratory data warehouse. An overlooked diamond mine. *Am J Clin Pathol* 2003;120(6):817–9.
- [14] McCarthy EP, Iezzoni LI, Davis RB, et al. Does clinical evidence support ICD-9-CM diagnosis coding of complications? *Med Care* 2000;38(8):868–76.
- [15] Solberg LI, Engebretson KI, Sperl-Hillen JM, et al. Are claims data accurate enough to identify patients for performance measures or quality improvement? The case of diabetes, heart disease and depression. *Am J Med Qual* 2006;21:238–45.
- [16] O'Malley KJ, Cook KF, Price MD, et al. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;40(5 Pt 2):1620–39.
- [17] Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+choice health plans that have chronic medical conditions. *Health Serv Res* 2004;39(6 Pt 1):1839–57.

- [18] Udvarhelyi IS, Gatsonis C, Epstein AM, et al. Acute myocardial infarction in the Medicare population. Process of care and clinical outcomes. *JAMA* 1992;268(18):2530–6.
- [19] National Committee for Quality Assurance. Health Plan Employer Data and Information Set 3.0. Washington, DC: National Committee for Quality Assurance; 1998.
- [20] Ahmed F, Janes GR, Baron R, et al. Preferred provider organization claims showed high predictive value but missed substantial proportion of adults with high-risk conditions. *J Clin Epidemiol* 2005;58:624–8.
- [21] Kern EF, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res* 2006;41(2):564–80.
- [22] Birman-Deych E, Waterman AD, Yan Y, et al. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care* 2005;43(5):480–5.
- [23] Bullano MF, Kamat S, Willey VJ, et al. Agreement between administrative claims and the medical record in identifying patients with a diagnosis of hypertension. *Med Care* 2006;44(5):486–90.
- [24] Waikar SS, Wald R, Chertow GM, et al. Validity of international classification of diseases, ninth revision, clinical modification codes for acute renal failure. *J Am Soc Nephrol* 2006;17(6):1688–94.
- [25] Aronsky D, Haug PJ, Lagor C, et al. Accuracy of administrative data for identifying patients with pneumonia. *Am J Med Qual* 2005;20(6):319–28.
- [26] Dombkowski KJ, Wasilevich EA, Lyon-Callo SK. Pediatric asthma surveillance using Medicaid claims. *Public Health Rep* 2005;120(5):515–24.
- [27] Kolodner K, Lipton RB, Lafata JE, et al. Pharmacy and medical claims data identified migraine sufferers with high specificity but modest sensitivity. *J Clin Epidemiol* 2004;57(9):962–72.
- [28] Harrold LR, Saag KG, Yood RA, et al. Validity of gout diagnoses in administrative data. *Arthritis Rheum* 2007;57(1):103–8.
- [29] Bazarian JJ, Veazie P, Mookerjee S, et al. Accuracy of mild traumatic brain injury case ascertainment using ICD-9 codes. *Acad Emerg Med* 2006;13(1):31–8.
- [30] Mapel DW, Frost FJ, Hurley JS, et al. An algorithm for the identification of undiagnosed COPD cases using administrative claims data. *J Manag Care Pharm* 2006;12(6):457–65.
- [31] Hollis J. Deploying an HMO's data warehouse. *Health Manag Technol* 1998;19(8):46–8.
- [32] Joyce JS, Fetter MM, Klopfenstein DH, et al. The Kaiser Permanente Northwest Cardiovascular Risk Factor Management Program: a model for all. *The Permanente Journal* 2004;9(2):19–26.
- [33] Sequist TD, Cullen T, Ayanian JZ. Information technology as a tool to improve the quality of American Indian health care. *Am J Public Health* 2005;95(12):2173–9.
- [34] Koski E, Teates KS, Tellez P, et al. Exploring the role of Quest Diagnostics corporate data warehouse for timely influenza surveillance. *Advances in Disease Surveillance* 2006;1:41.
- [35] Kupersmith J, Francis J, Kerr E, et al. Advancing evidence-based care for diabetes: lessons from the Veterans Health Administration. *Health Aff (Millwood)* 2007;26(2):w156–68.
- [36] Berndt DJ, Hevner AR, Studnicki J. The Catch data warehouse: support for community health care decision-making. *Decision Support Systems* 2003;35:367–84.
- [37] Starr P. Smart technology, stunted policy: developing health information networks. *Health Aff (Millwood)* 1997;16(3):91–105.
- [38] Payton FC, Brennan PF. How a community health information network is really used. *Commun ACM* 1999;42(12):85–9.
- [39] Berberabe T. Information: it's better when you share. *Manag Care* 2005;14(2):30, 35–37.
- [40] Solomon MR. Regional health information organizations: a vehicle for transforming health care delivery? *J Med Syst* 2007;31(1):35–47.
- [41] Massachusetts Health Data Consortium. Active Regional Health Information Organizations (RHIO) List. 2006. Available at: http://www.mahealthdata.org/data/library/20061127_ActiveRHIOs.pdf. Accessed August 19, 2007.

- [42] Wright A, Ricciardi TN, Zwick M. Application of information-theoretic data mining techniques in a national ambulatory practice outcomes research network. *AMIA Annu Symp Proc* 2005;829–33.
- [43] Everett AD, Ringel R, Rhodes JF, et al. Development of the MAGIC congenital heart disease catheterization database for interventional outcome studies. *J Interv Cardiol* 2006; 19(2):173–7.
- [44] Currie CJ, McEwan P, Peters JR, et al. The routine collation of health outcomes data from hospital treated subjects in the Health Outcomes Data Repository (HODaR): descriptive analysis from the first 20,000 subjects. *Value Health* 2005;8(5):581–90.
- [45] van Bommel JH, van Mulligen EM, Mons B, et al. Databases for knowledge discovery. Examples from biomedicine and health care. *Int J Med Inform* 2006;75(3–4):257–67.
- [46] Ben Said M, le Mignot L, Mugnier C, et al. A multi-source information system via the Internet for end-stage renal disease: scalability and data quality. *Stud Health Technol Inform* 2005;116:994–9.
- [47] Steil H, Amato C, Carioni C, et al. EuCliD—a medical registry. *Methods Inf Med* 2004; 43(1):83–8.
- [48] Hristovski D, Rogac M, Markota M. Using data warehousing and OLAP in public health care. *Proc AMIA Symp* 2000;369–73.
- [49] Muilu J, Peltonen L, Litton JE. The federated database—a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe. *Eur J Hum Genet* 2007;15(7):718–23.
- [50] Scotch M, Parmanto B. Development of SOVAT: a numerical-spatial decision support system for community health assessment research. *Int J Med Inform* 2006;75(10–11): 771–84.
- [51] Inhorn SL, Wilcke BW Jr, Downes FP, et al. A comprehensive Laboratory Services Survey of State Public Health Laboratories. *J Public Health Manag Pract* 2006;12(6):514–21.
- [52] Hanrahan LP, Foldy S, Barthell EN, et al. Medical informatics in population health: building Wisconsin's strategic framework for health information technology. *WMJ* 2006;105(1): 16–20.
- [53] Loonsk JW, McGarvey SR, Conn LA, et al. The Public Health Information Network (PHIN) Preparedness initiative. *J Am Med Inform Assoc* 2006;13(1):1–4.
- [54] Banks D, Woo EJ, Burwen DR, et al. Comparing data mining methods on the VAERS database. *Pharmacoepidemiol Drug Saf* 2005;14(9):601–9.
- [55] Iskander J, Pool V, Zhou W, et al. Data mining in the US using the Vaccine Adverse Event Reporting System. *Drug Saf* 2006;29(5):375–84.
- [56] Zeni MB, Kogan MD. Existing population-based health databases: useful resources for nursing research. *Nurs Outlook* 2007;55(1):20–30.
- [57] Cabell CH, Noto TC, Krucoff MW. Clinical utility of the food and drug administration electrocardiogram warehouse: a paradigm for the critical pathway initiative. *J Electrocardiol* 2005;38(Suppl 4):175–9.
- [58] Health Level Seven, Inc. About HL7. Available at: <http://www.hl7.org/about/hl7about.htm>. Accessed August 19, 2007.
- [59] Blobel BGME, Engel K, Pharow P. Semantic interoperability—HL7 version 3 compared to advanced architecture standards. *Methods Inf Med* 2006;45(4):343–53.
- [60] Regenstrief Institute, Inc. Logical Observation Identifiers Names and Codes (LOINC®). Available at: <http://www.regenstrief.org/medinformatics/loinc/>. Accessed August 19, 2007.
- [61] National Library of Medicine. SNOMED Clinical Terms® (SNOMED CT®). Available at: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. Accessed August 19, 2007.
- [62] Black N. Secondary use of personal data for health and health services research: why identifiable data are essential. *J Health Serv Res Policy* 2003;8(3 Suppl 1):36–40.
- [63] Emons MF. Integrated patient data for optimal patient management: the value of laboratory data in quality improvement. *Clin Chem* 2001;47(8):1516–20.

- [64] Post AR, Harrison JH Jr. PROTEMPA: a method for specifying and identifying temporal sequences in retrospective data for patient selection. *J Am Med Inform Assoc* 2007;14(5): 674–83.
- [65] Ricos C, Domenech MV, Perich C. Analytical quality specifications for common reference intervals. *Clin Chem Lab Med* 2004;42(7):858–62.
- [66] Westgard JO, Westgard SA. The quality of laboratory testing today: an assessment of sigma metrics for analytic quality using performance data from proficiency testing surveys and the CLIA criteria for acceptable performance. *Am J Clin Pathol* 2006;125(3):343–54.
- [67] Viljoen A, Twomey PJ. True or not: uncertainty of laboratory results. *J Clin Pathol* 2007; 60(6):587–8.
- [68] Karvanen J. The statistical basis of laboratory data normalization. *Drug Inf J* 2003;37: 101–7.
- [69] Ruvuna F, Flores D, Mikrut B, et al. Generalized lab norms for standardizing data from multiple laboratories. *Drug Inf J* 2003;37:61–79.