Published in final edited form as:

J Biomed Inform. 2007 June ; 40(3): 252–269.

Anatomy of Data Integration

Olga Brazhnik[1] and **John F. Jones**

Center for Information Technology, National Institutes of Health

Abstract

Producing reliable information is the ultimate goal of data processing. The ocean of data created with the advances of science and technologies calls for integration of data coming from heterogeneous sources that are diverse in their purposes, business rules, underlying models and enabling technologies. Reference models, Semantic Web, standards, ontology, and other technologies enable fast and efficient merging of heterogeneous data, while the reliability of produced information is largely defined by how well the data represent the reality. In this paper we initiate a framework for assessing the informational value of data that includes data dimensions; aligning data quality with business practices; identifying authoritative sources and integration keys; merging models; uniting updates of varying frequency and overlapping or gapped data sets.

Index Terms

Informational value of data; Clinical data integration; Database architecture; Data processing; Conceptual modeling

I. Introduction

The data producing era when keynote speakers impressed audiences by referring to petabytes of accumulated data has transformed into the epoch of data integration. From the challenge of technical management of huge data volumes the focus has shifted to their informational value. The overabundance of data accumulated by science and business needs to be brought together and integrated in order to extract information and gain knowledge. Data integration aims at answering complex questions, e.g. how a genetic background or exposure to some environmental factors influences development of certain diseases. Projects vary from a simple union of structurally similar datasets acquired at different locations and times to interdisciplinary integration, however, they all bring together diverse data through mapping and merging of concepts, models, controlled vocabularies, datasets, data elements and data values. While frameworks of standards, ontologies, reference models, semantic web and other modern technologies enable efficient fusion of heterogeneous data, they inherently create a danger of producing unreliable information if the data are of varying quality or if the process of integration is based on erroneous assumptions. In this paper we identify data dimensions, authoritative sources, integration keys and business processes that affect the reliability of produced information and are often overlooked in the mainstream integration approaches.

This paper consists of four parts and a conclusion. The following introductory sections establish terminology, identify data dimensions in the field of clinical informatics, and describe the goal

[1] Corresponding author: Dr. Olga Brazhnik, NIH/CIT, 10401 Fernwood Rd, rm 3NW03, Bethesda, MD 20817; brazhnik@nih.gov; phone: 301-496-7370; fax: 301-480-1385.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

of this paper. Part two discusses goals of data integration, supporting technology architecture and a conceptual data model. Part three identifies factors affecting informational value of collected data in every integration layer, i.e. data sources, elements, data sets and values. Part four focuses on preserving the informational value of data presented to users. The conclusion summarizes main points of the paper and draws a broader picture for applications of discussed methods.

A. Glossary of terms

Every term must be identified by its name, definition, its use, the timeframe and domain of validity, and various rules associated with it. Fig. 1 illustrates some of the definitions below.

Data means ‘what is given’ and refers to what we can perceive, experience or register with our senses or devices.

Concept represents an abstract idea generalized from particular instances.

Model is a set of concepts with defined relationships.

Information is created when data are interpreted based on a set of concepts.

Informational value of data is defined by the ability of data to provide us with useful and reliable information for our decisions and actions.

Data source is a database, a website, a publication, or any other collection of data, which is constructed upon a set of concepts and models.

Data element (DE) is an atomic unit of data collection that is unambiguously defined in the controlled vocabulary of a project. A collection of DEs defines a structure of a dataset.

Authoritative source contains the most reliable values for a specific DE.

Dataset is a collection of data that is produced from the data source at a moment in time based on a defined structure and a set of rules.

Data value is a specific value of a DE. “A body temperature of 101° F” employs DE “body temperature” measured in Fahrenheit and its value “101”.

Data pool is a collection of all datasets gathered for co-processing.

Data transformation is any action performed on data which can be simple selecting all DEs (SELECT *) or any combination of sophisticated statistical and analytical operations.

Project is a temporary endeavor undertaken to create a unique product, service, or result [1].

B. Dimensions of Data Integration

The traditional domain of clinical informatics can be generally characterized by three dimensions (Fig. 2). Dimensions are defined by invariants and variables. Some variables change along dimensions; others are explicitly excluded from consideration.

The intra-facility dimension represents the data acquired within a single health care unit. A variety of health information systems [2,3] has been developed to address the need to store, represent and integrate multimedia, semi-structured and temporal [4] heterogeneous repositories of patient records with radiology, laboratory, pharmacy [2], images [5], and bedside monitors [6] within hospitals. The invariants of this dimension are clinical and data management practices throughout a health care enterprise. Practice variations in time and among facilities are not considered.

The cross-facility dimension represents data coming from multiple health care units that may differ in their treatment practices as well as the use of standards and vocabularies. Seamless sharing of multimedia clinical information [7] is necessary to provide the continuity of care in the modern mobile society where a person may have a number of encounters with diverse health systems over the course of a lifetime. Semantic interoperability is assisted through implementation of clinical data standards such as Health Level 7 (HL7) [2,8], Clinical Data Interchange Standards Consortium (CDISC) [9], and Digital Imaging and Communication in Medicine (DICOM) [2], and development of machine-readable sources, such as Unified Medical Language System (UMLS) [10], that integrate most of the existing medical vocabularies and establish terminological standardization. However, the integration of distributed clinical data produces challenges that go beyond the technical and semantic, because medical practices, as well as baselines and data distributions, are not identical in different parts of the world due to variations in nutritional habits, climate and quality of life [11]. These variables can accounted for in the cross-facility data dimension, which deal with one type of data at the same point in time across multiple facilities. It answer questions like “what differentiates x-rays taken at office Z from x-rays taken at hospital Y?”

The third dimension reflects changes over time, including health conditions of a specific individual [12], advances in knowledge and variations in medical records due to modifications in health care practices [13]. An individual, facility and data types are invariants of this dimension.

Identifying data dimensions provides a frame of reference for characterizing relationships between datasets and for defining metadata required for unique placement of a dataset among others. These metadata serve as coordinates in the data space that can be visualized and studied via geometrical approaches [14]. The process of identifying variables and invariants will eventually lead to defining independent orthogonal dimensions that will structure and reduce the required metadata.

If we represent every collection of data as a point in data space, a complete patient record may be presented as on Fig. 3. Placing all DEs and data sets in one space enables us to identify the path for their integration [14] and to visualize dimensions along which the integration has to occur. While in the physical world a spatio-temporal representation of data comes naturally, the high dimensional biomedical data requires special conceptualization efforts in choosing dimensions for modeling, integration and visualization addressing our limitation to three-dimensional images. Once defined, the dimensions help to build, communicate and validate models and to identify potential errors. This becomes especially important now when advances in bioinformatics and computing continuously create new data dimensions and combine them in a multitude of ways through image processing, telemedicine [15], widely available knowledge bases [10], research information easily accessible to professionals and students [16], and the application of artificial intelligence techniques for obtaining correct interpretations and prescribing appropriate therapies [6]. Advances in biomedicine brought together drug discovery and system biology [17]; enabled integrated studies of gene expression, clinical chemistry and pathology endpoints [18]; initiated integration and analysis of genomic, molecular, cellular, and clinical data, for molecular medicine [19]. This newly created high dimensional biomedical data space requires development of robust mapping and navigation principles with a defined frame of reference.

In the era of ubiquitous computing, clinical decision support depends on the accessibility to and availability of information at the point of care in a comprehensible format such as clinical reports and patient-specific summaries [20] derived from integrated Electronic Patient Record [21]. Merging and interpretation of data from diverse sources includes integration across multiple dimensions and mapping of models of data collection and processing employed at these sources. Models and dimensions become critical to ensure meaningful and testable results in acquisition of clinical information through intelligent clinical information management systems [22]; integration of patient-dependent knowledge and external knowledge bases such as Medline [23] with time-oriented clinical databases [16] in context-sensitive interpretation of these data [24]. Comprehensible conceptual models are needed to engage domain experts in validation of Computer Assisted Diagnosis [25-27] and facilitate patient education [28], e-health services [7], and communication of the available scientific information to the public.

Projects of clinical data integration range from integration of follow-up data for a specific disease [10] or a specific type of data, e.g. laboratory [29], to large scale health information systems [2] and international endeavors [7]. Each of them needs to identify the data dimensions involved to define the path of integration of these diverse data sets.

The overabundance of biomedical data dictates the need to present them in comprehensible formats tailored for specific purposes and audiences. While visualization approaches proved to be highly efficient [30,31] for these purposes, the metrics of a configuration data space have to be properly defined since they differ from the metrics of the Euclidean space used for visualization. Identification of data dimensions is a necessary part of this process.

C. Goal of this paper

In this paper we attempt to identify factors that influence the reliability of information produced from the integrated data. Utilization of standards, controlled vocabularies, ontologies and Semantic Web guarantees neither that the data truly represent the situation at hand nor that integration assumptions are correct. These two factors may lead to misinterpretations of data. Through introducing dimensions of integration and aligning reliability of DEs with business practices, this paper defines a framework for assessing the informational value of data and creating a solid backbone for integration of disparate data sources. The quality of the information produced in intelligent analysis or data mining requires that the data are not redundant or heavily contaminated, and that every entity has a single identity. Considering the influence of business practices on data reliability we complement other comprehensive reviews of common problems in integration of heterogeneous data sources as well as reasons for and methods of data preprocessing [32,33]. Producing reliable information by processing data through pipelines of agents, ontologies, portals, SSL, data warehousing, XML, Bayesian and other techniques, ultimately requires an assessment of quality for every ingredient in preparation of the information pie, regardless of the technology used.

Examples in this paper come from real databases, whose names cannot be released, and illustrate the connections between business practices and quality of DEs that exist in the real world. These examples assist in conceptualizing common challenges, steps and layers in projects of clinical data integration and have no factual value. While the focus of this paper is on the biomedical domain, we emphasize the commonality with other distant domains that might have already developed solutions for some of these problems.

II. Goals of Data Integration

The goal of any project defines the focus, models and tools to be used in achieving it.

In general, integration is pursued in one of two ways. The first way is when we know exactly the questions we want to answer and what data are available. In this case we aim at finding reliable sources and pulling the needed fields into a database designed for the purpose. An example of this type of project could be “identifying a gene expression profile for people with a specific disease.”

The second kind of integration aims at understanding the power of data at large when a number of data sources of varying reliability are available. In projects like this, the goal can be vaguely defined as “find correlations” or “discover new knowledge” via various data mining techniques. Projects of this kind aim at figuring out what exists, how reliable the available data are, and what additional data are required in order to enable answering complex interdisciplinary questions. Health related projects and epidemic outbreak studies are often approached in this way. They bring together data from pharmacies, schools, private practices and major hospitals in a search for indirect indications that something unusual is happening [34].

In both approaches we must be able to analyze the same data with various methods and techniques and present results according to requirements of various user groups. The information pipeline discussed in the next section provides a technology architecture for accomplishing these tasks, while a conceptual data model provides the substance for processing.

Projects of clinical data integration aim to identify possible disease outbreaks; study the causality of diseases; discover the relationship between diseases and risk factors; enable singling out groups of people that are at higher risk for a specific disease due to their occupation, location, medical history and/or possibly dangerous past exposures. These projects employ both modern health information systems and legacy sources, which differ in the quality of records, lists of occupations, disease codes, and business rules. The purpose of the data collection is rarely focused on identifying diseases. Some databases are designed for billing purposes, and others assist in subject tracking. New data are continuously collected and fed into distributed sources, from which they are made available at a central location with varying frequencies of updates. Data feeds overlap thereby producing conflicting data. In this paper we discuss how to assess information quality in a situation when the goals of clinical data integration significantly diverge from the original purposes of data collection.

The users of clinical data usually include medical doctors, researchers, public health specialists and decision making officials.

Medical doctors must be able to recover and present in a comprehensive form all information about their patients. For every encounter they integrate data along the intra-facility dimension and may need to view the time data dimension to create a longitudinal clinical record for a specific patient.

Aiming to identify common factors (age, gender, occupation, location) in groups of people with a certain disease, *researchers* study causality of diseases. Due to the Health Insurance Portability and Accountability Act (HIPAA) compliance datasets have to be de-identified. A multitude of utilized data dimension needs to be defined in this complex endeavor to ensure consistent and testable outcomes.

Public health specialists need to have a reliable way to detect outbreaks in near-real-time; be able to find individuals with identified common risk factors for a disease; and based on results of scientific studies, develop strategies for treatment and prevention for such individuals.

Decision-making officials need access to the results of studies, recommendations for actions and measures, and relevant the factual information.

A. Information Pipeline

Information pipeline architecture depicted in Fig. 4 ensures that the data can be re-analyzed with various methods and presented in a multitude of ways according to specific goals, requirements and user groups. Each part of the information pipeline serves a specific purpose. The Operational Data Store (ODS) collects all required data which are merged, cleansed and validated in Extract-Transfer-Load (ETL) process. The Core Database (Core DB) contains a reliable non-redundant collection of data, parts of which are made available for queries through data marts designed for specific purposes and groups of users.

Datasets from external sources are uploaded into the project-specific ODS, the sole purpose of which is to gather all required data. The data in the ODS can be very redundant and not cleansed or validated.

The information pipeline can be implemented on a logical level, eliminating the necessity to collect distributed data for centralized processing, which is often unrealistic due to the prohibitively high cost of centralized storage, the unavailability of the necessary bandwidth to efficiently transmit the data, or privacy concerns [11]. A chosen architectural solution may introduce some additional steps in data processing, such as synchronization of updates, but it should preserve the main components discussed in this paper and be transparent to the end users [35].

The project data model is implemented in the Core DB. Non-redundant, cleansed and validated data are loaded in a Core DB from the ODS through the Extract-Transfer-Load (ETL) process. While a specific implementation of a well-normalized Core DB may vary (both traditional Entity-Relational or Entity-Attribute-Value design [36,37] are valid), it must be consistent with the conceptual model of the project, ensure data integrity and store both data and business logic. Unfortunately, the process of creating a Core DB is omitted in the majority of modern projects, where data marts or the start-schema-based data warehouses (DWH) are populated directly from an ODS. Enabling fast and effective queries start schema design does not support integration. It represents a denormalized structure focused on a specific subject (e.g. human resources or billing), is practically inverse to hierarchical, and lacks explicitly defined relationships among entities. The normalized Core DB could be inefficient for queries since it serves a purpose of storing data in a consistent way and provides the central source for creating data marts for various subjects, users and business requirements. Normalized models enable the utilization of various abstraction levels that are important in dealing with clinical data [38] and provide efficient structure for data storage. Data marts are created from the Core DB to address the various needs of the end users. They could focus on a specific subject or provide de-identification of data enabling researchers to do data mining at large without challenging the privacy of patients. Goals, users, methods and tools influence the choice of the data mart designs, which could be implemented as a flat table or as a part of the normalized schema or a start schema or employ an object oriented approach for decision support applications [38].

Metaphorically speaking, ODS is similar to a pile of bricks and utility parts that are put together in a building represented by the Core DB, while the data marts distinguish various uses of the house: owners of data, such as doctors, are allowed to see all their possessions, while visitors of various levels are only permitted into the living room or kitchen, where the private information is hidden or only summary reports are available.

B. Conceptual Data Model (CDM)

Assessing the informational value of data is based on a conceptual data model and includes identifying required DEs, finding authoritative sources and creating a strategy for their integration. The same data can be processed in various ways resulting in potentially different information. None of the outcomes, however, provide exhaustive comprehensive knowledge of a reality, and we must always decide and act on incomplete information. Hence, explicit definitions of the surveyed part of the universe and models of data collection and processing enable users to evaluate the information reliability and employ proper models of uncertainty in decision-making.

Despite its importance for assessing the informational value of data, conceptual data modeling only recently started to gain due attention [39]. CDM addresses one of the key issues in intelligent data analysis: the ability to recognize what is important in a problem [32] and how the collected and available data can be used for producing the desired results.

CDM identifies main objects of interest and relationships between them. It defines the variables that are available for the study and those that are outside of it. For example, in the study of occupational risks in the development of certain diseases, we merge medical history and personnel records for various groups of individuals. Personnel records have information on jobs and locations. There are numerous parameters outside of the study scope, such as diets, social habits, or genetic predisposition to diseases that might be thought important but cannot be controlled and recorded. A sketch of a CDM is depicted in Fig. 5.

Identifying variables outside of a CDM helps to evaluate the uncertainty of the results and enables us to employ standards, which may facilitate future integration with the adjacent domains. The influence of these external variables can be included in hypotheses and models, enabling us to enrich data analysis with mean field approximations and boundary condition effects extensively used in computer simulations. Employing these techniques might prove useful in defining, for example, boundary conditions for geographical development of epidemics where some channels, such as water and food supplies or air travel, are not included in the study. Applying modeling methodologies of other disciplines will assist in advancing data and information models to the scientific robustness of engineering and physical models.

Every part of a CDM is associated with a certain set of values such as total population, complete list of occupations, locations, disease groups, etc. ‘Total’ refers to the set that we ask questions about, e.g. the population of the United States, while surveyed defined the set of available data, e.g. 871 patients from 91 health physician offices representing three southern states and less than 2% of cases occurring in USA. Defining the relationship between the total and the surveyed sets allows us to estimate the expected informational value of the results; assists in developing strategy for de-identification of records; enables to create information models and to employ appropriate sampling techniques for meaningful data processing and result interpretations. A well known result of data mining that “everyone who died before 1900 was famous” [40] illustrates the necessity of defining this relationship correctly (obviously we have records only about famous people of the past).

CDM defines the part of the universe under investigation and its relationship to the world at large.

A logical data model created from CDM defines DEs required to accomplish project goals. The next chapter will identify steps to ensure reliability of these DEs.

III. Layers of Integration

Integration occurs in four major layers: data sources, DEs, datasets, and data values. This includes integration of concepts, models, controlled vocabularies, methods of data acquisition, frequencies of updates, as well as units and formats of records.

A. Data Sources

The same DE may exist in multiple sources. Characteristics of data sources shown in Table 1 determine (1) which source can be considered authoritative; (2) what data acquisition methods can be used; and (3) how this source can be meaningfully integrated with other sources.

The most important characteristic of a data source is its purpose. Business processes are usually aligned to support the purpose and to ensure that the critical DEs are captured in the most reliable way. These *focal* DEs are usually mandatory and have the highest quality due to various validation techniques applied to them. The *peripheral* DEs could be optional. In on-line purchases, for example, the credit card number, expiration date, the name of the card holder and the billing zip code represent focal elements, while office phone number and the details of the address are peripheral. Not all mandatory DEs are focal. A custom of requiring DEs, which have no practical importance, creates a source of misleading information. During implementation we always try to minimize the number of optional and free text fields because they create inconsistent data. However, making these DEs mandatory without aligning them with business requirements is even more dangerous because it is more difficult to identify a problem when all entries look absolutely valid. For example, an online form may require you to make an entry in a field, and even limit your choice to the list of values in a drop-down menu. However, *what* you enter in this field may not be important for business operations and, therefore, will not be validated. These entries have no informational value and can be misleading.

Commercially available health information systems primarily focus on administrative tasks and seldom provide additional knowledge based functionality [21]. In clinical databases that support proper billing, the information about insurance, patient identification and the performed procedures defines the basis for billing and, therefore, is captured in the most reliable way. Signs and symptoms that allow a doctor to decide on a diagnosis are not crucial for billing purposes and, therefore, may not be reliable. If captured at all, they can only be used to support the diagnosis (not to derive it from the record), which therefore limits their use in expert systems. Medical treatment facilities that serve the military population often do not file insurance claims. Hence, the information about procedures may not be captured properly. The primary purpose of these DBs is to identify a person, a health care provider, and a date and type of visit. As a result, for example, coding for ‘flu’ might be used for recording both sickness and vaccination and only analysis of the providers reveals the difference between a *diagnosis* defined by a doctor and a *vaccination* administered by a nurse. Having strongly enforced standards in place does not guarantee that the entered information represents correctly the situation at hand. With a few minutes to see a patient, a doctor enters his notes in the largest text field leaving numerous small ones with default or any easy to enter values. Among thousands of ICD10 codes a human mind can memorize a few dozen, and instead of looking for the most appropriate code, a care provider enters the one from his memory that approximately reflects the case. These examples show that unless there are business processes that suffer from the incorrect values of DEs (and, therefore, measures are taking to prevent it), the enforcement of standards alone does not guarantee that the captured data adequately describe a situation such as a clinical encounter.

While quality assessments measure the adherence to standards, identifying business practices and real life situations helps to assess how closely data represent the reality. Every DE should be tested with a number of questions. How would its incorrectness affect the business? How is the incorrectness discovered? What is the practice of correcting errors? What other participants, i.e. customers, patients, suppliers, providers, depend on this DE? What actions are based on it? What process ensures coherence between a DE and the situation it represents?

The remainder of this section will discuss how the stated principles apply to some common types of clinical databases such as an emergency room database (ERDB), a mobile care DB, a laboratory DB, and a hospital DB.

The purpose of an ERDB is to record immediate signs, symptoms and the treatments administered at the ER, along with recommendations on future care. To keep records for a patient that may arrive unconscious, without any documents and medical history, the ER assigns a surrogate social security number (SSN). During rescue operations away from the hospital, ERDB works without relying on connectivity to the server, which prohibits mandating the validation of DEs that, otherwise, could be pulled from the central DB. These DEs are often not populated at all or may contain invalid or default data. A proper maintenance procedure for an ERDB should include a verification of patient identity. However, when the data from the ERDB are received, there is no guarantee that this has been done. Manual entry, surrogate SSNs, and the absence of validation hinder integration of ER data with other data sources. An ERDB provides an example of a primary data source that cannot be considered authoritative for DEs such as name, date of birth, SSN. The most reliable fields here are signs and procedures.

A mobile care DB keeps track of patients who are sent to a larger unit if a case cannot be handled at a local facility. Mobile care is responsible for delivering a patient with a certain ID from one place to another and that proper procedures are performed during the trip. Diagnosis, symptoms, and personal data are not focal in this database.

Requirements for in-patient care and out-patient care differ significantly, which causes differences in records. In both cases, however, the goal is to support proper billing of the correct person. We consider them as one hospital DB that ensures reliability of data about procedures, SSN, date of birth, name and billing address.

A laboratory DB stores results of various tests and images and usually supports billing. Thus, it captures only the summaries and not the details of test results, while storing raw data would open a door for biomedical integration. A patient’s identity and the list of performed tests are essential for the business.

Identification of focal DEs assists in finding authoritative sources. Fig. 6 illustrates possible situations with overlapping DEs. The large circles depict the entire sets of DEs in data sources, whereas the smaller circles represent focal DEs. The data sources presented by two non-overlapping circles cannot be integrated at all because they lack common DEs (e.g. microarray data and library catalogues). Two overlapping circles in the middle create an illusion that they can be integrated. However, the actual overlap happens in the peripheral part. For example, both college records and ERDB record SSN, name and ethnicity. While name and SSN are extremely important for college billing purposes, ERDB cannot always verify a person’s identity. A part of ethnic background reported to the college (e.g. to benefit from an association with a minority group) for mixed ethnicities might not be obvious from first glance in ER. The ethnicity is not essential for an ER business, and there is little effort to record it correctly. In this case, existing in both sources DEs of name, SSN and ethnicity do not provide a solid basis for integration.

Meaningful integration may occur only between sources with a shared pool of focal DEs. For example, real estate listings and geographical information systems can be easily integrated based on addresses essential in both cases.

The discussion of focal DEs illustrates that in order to study correlations in two distant domains, e.g. gene expression profiles and purchasing preferences on eBay®, we may need to build a multi-step integration staircase that links gene expression data with clinical with personnel with credit card DB and only then link it with purchases on eBay (provided all these data are available.) Fig. 7 illustrates a process of identifying authoritative data sources and integration keys.

The importance of metadata, which used to be a “second class citizen” in data projects, has recently gained recognition in integration of tools and data [41] and some discussion of it is necessary here. From the integration standpoint there are two types of metadata. One is used for a unique identification of an instance such as equipment serial numbers, timestamps, etc. Even though these DEs might not directly support the business for which a database is created, they are essential for the data integrity in the database itself and could serve as reliable integration keys. The other type is peripheral data, some of which capture the environment of data acquisition; some are collected because they were thought to be useful or traditional, or planned for future integration. Peripheral DEs do not support any business functions, not validated and, therefore, are not reliable if not misleading. For example, ethnicity captured as a “pick-one” option does not allow for mixed races and thus becomes meaningless for genetic research. Unique identifiers and peripheral metadata should be treated differently in integration endeavors.

B. Data Elements (DE)

1. Controlled Vocabularies—A controlled vocabulary is a set of unambiguously defined terms used within a project. A CDM consists of concepts, while a logical data model deals with DEs, based on which the data are collected. Generally, DE refers to one data field, while concepts could consist of one DE (gender) or multiple DEs (vital signs). Both concepts and DEs are represented by terms. A controlled vocabulary provides quality and consistency in data collection, processing and interpretation within a project. However, the same concept may be named differently in various sources thereby creating “naming differences” [35], or a complex concept in one project can be a simple concept in another one, or the same concept can be represented by different set of DEs. While there are numerous attempts in creating shared ontologies [42] and terminologies [8,9], they would benefit from employing conceptual modeling and defining complex relationships [14].

Table 2 shows names used for the concepts of primary diagnosis and the patient’s date of birth in four different databases. Integrating controlled vocabularies means that we assume a simple one-to-one equality mapping between terms from different sources, i.e. that they mean exactly the same thing. Term mapping also allows us to record temporal changes in terminology, e.g. if what we used to call a “job” we now call an “occupation”. However, exact synonyms are rare and they can usually be defined only in relation to a specific domain, goal and process. Moving between abstraction levels as in substituting concepts of “person” and “organization” with a more abstract concept of “party” requires to map models [43] and define complex relationships [14]. Documenting all mapping assumptions ensures proper interpretation of final results, correction of mistakes, and enables transitions between conceptual models.

2. Integration Keys—For the purpose of integration, all data elements can be split into three categories: integration keys, informative DEs and auxiliary DEs. The informative elements represent the goal of the integration. They contain the information we seek: diagnosis, signs, symptoms, age, and occupation. These elements are necessary, but not sufficient for achieving the goal of integration. In order to link data sources, we need to identify integration keys, which provide the backbone of integration. Auxiliary DEs provide additional information that might be associated with the business rules and allow us to decide whether to include or exclude some records. While integration keys and informative DEs can be chosen only from focal DEs, the auxiliary DEs can be obtained from either focal or peripheral DEs.

Database keys, which uniquely identify every record within a table or a database, usually cannot be used as integration keys because they differ from one DB to another. An integration key is a combination of DEs that identifies exactly the same entity in two sources and is chosen from the overlapping focal DEs. While we generally seek some combination of an SSN, a date of birth and a legal name, to identify a person, these DEs may not be equally reliable in all sources as in the example of an ERDB.

Choosing integration keys is the most crucial part of the integration. A widespread approach [44] “to integrate on all common elements” is highly inefficient. If we assume that all DEs have the same data quality, we will “mow our vegetable garden”. The “weeds” have to be taken out first. Fig. 8 shows that if we include date of birth as a part of the integration key, we may encounter many mismatched records, and they will be lost or require manual handling. If, however, we exclude date of birth, based on the fact that it has many invalid entries, our chances for successful matching will increase. An example follows.

The Internal Revenue Service (IRS) and *Amazon.com* customer data can be easily integrated on a very minimal set of DEs: last name, first name and one of the billing addresses. There is a good chance that one of them is where the person is actually paying taxes from. These DEs are essential for both organizations and are highly reliable. However, including the date of birth in the integration key diminishes the chances for integration. The IRS uses the date of birth to uniquely identify a person and to define when the taxes should be withheld and at what rate. For *Amazon.com*, however, the date of birth is an optional field. It is not essential for running the business, but it helps to promote business by notifying your friends about your upcoming birthday and your existing wish list. So if you want to misinform them about this date that is totally up to you. The reliability of this DE is very low. Therefore, if this integration process is done on a set of DEs that includes date of birth, there is a chance that the number of mismatches will be much higher than without it. Only focal DEs can be used as parts of an integration key.

The choice of integration keys is further influenced by our ability to establish simple procedures for handling invalid values, mismatches and duplicates.

Defining data dimensions assists in identifying integration keys. For example, to create a longitudinal patient record we would seek integration keys among DEs describing a person. When integration involves only time and intra-facility dimensions, a SSN may be a good choice for merging records, however, it becomes an auxiliary element in merging records between multiple health care systems due to variations in policies and data handling practices. While majority of people in the USA have SSNs, foreign nationals may not have them at all or have multiple surrogate ones obtained prior to receiving a permanent SSN. The surrogate SSN cannot be validated and therefore can creates more data entry errors. In international projects a SSN may not be useful at all because the post-World War II laws in Germany, for example, prohibit assigning unique IDs to individuals, and rules in other countries vary.

Auxiliary DEs [45] are associated with business rules and might assist in making decisions on handling exceptions and manual record matching. Identifying a person by her appearance (height and eye color), or circumstances, locations, and dates of events might help in gluing parts of a record together. Sequences of records appearing in various sources, e.g. in ERDB, then in a mobile care DB, then in a hospital DB within a short period of time, can indicate the same encounter even with some parts of integration keys mismatched. However, this part of the integration process cannot be automated and all assumptions and evidence must be properly documented to enable correction of possible errors later. Auxiliary elements could be used for validation of the results of automated record matching based on integration keys. For example, data on sales of over-the-counter drugs in pharmacies and school absences provide effective complementary information in identifying disease outbreaks [32].

Table 3 illustrates integration keys and informative and auxiliary DEs in three DB. The proper categorization of DEs is both data source and project specific.

3. Standards and formats—The standards and formats for data capturing are defined for every DE and in fact should be defined for every concept, too. An example of a concept level standard definition is a HL7 message [8] that defines all DEs and their positions in the message to enable its proper interpretation.

Implicit use of measurement units hinders automatic record merging, especially among international datasets. The weight "157" would be read in pounds in the USA, while in Canada it would be interpreted in kilograms. The body temperature of "96" is unthinkable in centigrade, and the height of "5.7" is equally unthinkable in meters. Dates are usually the most problematic since there are numerous formats to represent them. The date of 8/7/93 can be interpreted as August 7, 1993 in United States, while in Europe it means July 8, 1993. The standard format for SSN XXX-XX-XXXX can be stored without dashes. In some clinical records SSN can also be preceded by two digits identifying a dependant of a health insurance policy holder.

Explicit definitions of DEs that include formats and units of measure enable merging values of mapped DEs from diverse sources. Data type mapping for various database management systems is straightforward, and it is discussed in multiple sources, e.g. Torgue's type map [46].

C. Data Acquisition

Data acquisition refers to the process of acquiring data directly from the subject of study or multiple sources. Methods of data acquisition are largely defined by the characteristics of involved data sources.

Data sources differ by their functional design, availability of data, and the process of data acquisition. Two well known distinct functional designs are represented by On-Line-Transactional-Processing (OLTP) and On-Line-Analytical-Processing (OLAP). Usually data are collected in OLTP and then migrated to OLAP via an ETL process. The goal of OLTP is to keep track of all transactions and record modifications in a database. OLTP does not check for the consistency of meanings – only formats and enforced business rules matter here. It is highly normalized, has numerous tables assisting in implementing business rules, and is used for consistent record keeping only. Along with valid records, OLTP might also contain all invalid records, flagged for auditing purposes. Deep understanding of business logic is required to obtain meaningful results through OLTP queries. Hence, while OLTP has the most up-to-date data, extracting any kind of summary or statistical information from it presents a challenge. OLAP, on the other hand, may not contain the most recent data, but it is designed for querying and reporting and allows the end user to compare different categories and summaries of data.

A data source can be public or proprietary. GenBank®, UniProt, and other repositories of research data are public and available to everybody. Users need to know what to look for and how to extract the necessary information. Extraction of meaningful datasets from these sources requires knowing how these data sources are maintained, whether duplicates are allowed, and what kind of data validation exists. Proprietary data can often be purchased or received for purposes clearly defined in data use agreements (DUA). A DUA defines what data can be exchanged or shared, and how they can be used. Clinical DUAs must be HIPAA compliant. Some companies, such as *Amazon.com*, make their de-identified datasets available for research purposes.

Sources are divided into primary and secondary by the method of data acquisition. A data collection site represents a *primary* data source. Data entry can be manual, e.g. by a receptionist, or automatic, e.g. via barcodes or through a laboratory information management system (LIMS). An ODS as well as a Core DB are both examples of secondary data sources that gather data from other (primary or secondary) data sources and often employ data transformations.

Data gathering agents, data pulls, and flat file updates represent the most common ways of acquiring data from the sources.

Agents are configured to obtain newly-available data at certain time interval – every 5 minutes, on the hour, or daily. They are most useful in obtaining data from primary data sources to report occurrences of single events. However, because reconciling data delivered by agents is difficult, they should be used with caution. Multiple alerts about the same disease encounter may come from multiple sources with a surrogate SSN hindering their integration. Agents are effective for alerts, but not for reporting or producing summarized and integrated information. The real-timeliness of data is significantly altered by the data gathering processes or delays in committing records to the database (e.g. data validation). Understanding what part of a data pool is represented by agents introduces further complexity in the processing of incoming data. Fig. 9 illustrates an example in which agents work on servers updated from multiple offices. The agent usually has no information about whether all participating offices have submitted their data. Therefore, the completeness of the data pool is unknown.

Data pulls represent customized queries performed on external databases. They are created by a person internal to the project, can be automated and are executed on a regular schedule. The query is usually performed on a complete data repository (secondary OLAP with cleansed and validated data) and not on a transactional database, which allows eliminating most of the problems with data inconsistency and incompleteness. Data pulls can be customized for the needs of a project, and can have built-in filters that reduce redundancy and invalid data. However, data pulls do not provide real-time data.

Flat file updates are often used when DUA regulate what data can be made available to the users and when security or HIPAA compliance requirements prevents end users from accessing the source database. In this case, a user requests a dataset to be created at the source site. This type of update is connected with numerous restrictions, both in customization of datasets and in their later use. Public data repositories also employ flat file updates to enable users to run local queries.

The project goals and available data sources define methods of data acquisition. Agents are the most efficient in providing near-real-time alerts if at least limited access to the source is granted, they are equally effective on OLTP and OLAP. Data pulls and flat file updates may involve delays but provide consistent data sets that are easy to analyze and integrate. The project goal defines the choice of technology solutions.

D. Datasets and Data Values

Integration of datasets provides a consistent snapshot of data at a defined point in time. It has two dimensions - time and value.

The fullness of the data pool in time created with the use of all three methods of data acquisition can be characterized via Fig. 10. The data pool consists of quick updates from agents combined with data pulls, which can be done as often as the source repository is updated, and with flat file updates, which could be infrequent or even irregular. These datasets overlap because agents bring data that become available later as a consistent update. This synthesis of datasets that were created at various times and which may overlap [47] is similar to restoring data from incremental backups and to the fusion of asynchronous sensor data [48]. Meaningful summaries can be done up to the date of the update that completes the data pool, even if a part of the data is received in near-real-time.

Identifying an authoritative data source for every DE assists in uniting semantically overlapping datasets, which may supply redundant or conflicting values, such as different diagnoses for the same disease encounter. However, these differences could be valid and simply represent variations in methodologies or opinions. DEs could be similar but not the same [49]. In this case a practice of keeping all redundant data along with the information about the source becomes important. This is also useful in bioinformatics databases, where the quality of DEs may vary from one experiment to another and, therefore, require using different sources for the same DE. This approach may not be realistic in high throughput environments. However, with adequate resources and a clear guidance for dealing with conflicting DEs, keeping all redundant DEs provides a more versatile decision support, allowing users to make decision based on all available evidence.

The relationships between parts of a data pool represented in every source define whether records that do not have a counterpart (e.g. those having a job, but no disease, and visa versa) should be included in the final dataset. Datasets can be complementary or have various complex dependencies, as in [34] the case in which data from pharmacies predict disease outbreak before hospital data. Some overlap in data values can be due to the mobility of people, or the same or similar case can appear in an ERDB, a mobile care DB, and in a hospital DB within a short period of time.

IV. Presenting Integrated Data to Users

After the cleansed and validated data are organized in a Core DB, data marts are created for each distinct group of users. Their purpose is to enable efficient querying of data with user-specific methods and tools. Technically, data marts can be implemented as views, flat files, or a database populated from the Core DB via an ETL process. A traditional star-schema DWH focused on a patient with well designed materialized views [50] could be the most efficient solution for doctors. The research datasets are often presented as flat files that are easy to mine. De-identification, i.e. hiding of personal information while attempting to preserve the informative value of data [51], complicates creating research datasets. The main factors in choosing the solution for data privacy are the level of trust between analysts and the data-owning organization and the probability of person identification in the surveyed population. In collaborations with external (untrusted) analysts, masking values of selected DEs [52] allows to separate the task of data analysis from the interpretation of results. In this case all codes for diseases, occupations codes, and other identifiers are replaced with some fake coding which is recorded and stored at the data owner site. Which DEs could be masked without loosing the informational value of data depends on the dimensions of integration. In studies of event consequences or correlations with social and environmental factors, masking or distorting the dates of encounters will significantly decrease the value of the results. In comparing longitudinal records, i.e. medical histories and patterns of disease development, preserving exact dates may not be as important as keeping the time intervals intact. Usually DEs with limited numbers of values are masked, while dates and other DEs of high granularity are distorted. Masking tables store mappings between all distinct original values and created substitute keys, i.e. masks, (Table 4) and are accessible only to internal users. The dimensions of a data pool define the risk of disclosing quasi-identifiers that enable to identify a person uniquely by a combination of non-key DEs. For examples, a few cases per geographical location or health care facility in a specified time frame may compromise privacy of patients. In studies of rare diseases where a person can be easily re-identified by a set of seemingly non-informative characteristics, more sophisticated methods of de-identification including controllable distortion of data [53] are used. The increasing integration of patient-specific genomic data into clinical practice and research raises other serious privacy concerns [54].

Table 5 provides an example of masked data. Since the masked sets processed by mathematicians and computer scientists only gain the informational value when unmasked, a meaningful grouping of diseases, occupations, locations and other variables optimizes data analysis. However, the meaningful grouping is defined by the project data dimensions and presents a classification problem, which is a part of a general challenge in selecting relevant features for the target concept [55]. Depending on the goal, grouping all infectious diseases might not be the best approach, but grouping flu and cold-like illnesses together or reducing locations to the level of city or county might prove useful. The Center for Disease Control (CDC) [56] provides some practical guidance for disease grouping.

The process of integration is not yet completed when all the data are made available to the users. In order to communicate the meaning of the data, the underlying models of the sources should be mapped and presented in a comprehensible form. Disease recording and reporting differs among health care facilities (cross-facility dimension). While the concept of primary diagnosis is usually defined unambiguously, even with the use of different names as illustrated in Table 2, the details recorded along with a diagnosis differ significantly from one database to another (Table 6). Their integration requires mapping of the basic concepts, disease models and their implementation in DB schemas [57].

Schema matching aims to map semantically corresponding to each other elements of two schemas [58] and to enable co-processing of data collected against different models [45]. Semantic integration [59] of models is challenged by multiple ways to interpret relationships between entities [60] and by different levels of granularity in presentation of models. Semantic modeling of a schema includes ontological categories, properties and contextual constraints [58], followed by developing measures of category properties. The ability to consider data at abstraction levels higher than the levels at which they are stored [38] enables mapping between diverse models.

Schema matching often involves database re-engineering, schema transformations [61] and middleware data models [62]. While it mostly remains a manual, labor-intensive and tedious process, some approaches have been developed to automate it [63]. Semantic networks assist in identifying equivalent concepts in disparate databases and in producing candidate concept mappings [64]. Statistical, heuristics, or Bayesian [65] approaches support the process. With the rare exception of some large integrated healthcare systems utilizing a common data model for integrating information from multiple facilities [66], the vast majority of existing clinical information is stored in heterogeneous databases using different names and different data models. Therefore, direct system-to-system mapping of data elements remains the most frequently used method for data exchange [64]. All mapping steps have to be documented to enable re-analysis of data on different sets of assumption.

A *full description* of a disease encounter may include diagnoses, symptoms, complaints, prescribed medication, recommendations on follow-up lab tests, etc. It presents an implicit model of an existing clinical practice and contains all DEs that were found useful for a defined purpose. The variations in clinical practices among facilities cause differences in the structure of clinical records illustrated in Table 6. The ability of a patient to afford the service, his discipline as well as other specifics of his case define which DEs are populated with values. This difference between the structure of intended protocol and its actual implementation, i.e. actually populated DEs, has to be accounted for in the process of integration [14].

Identifying confirmed cases of diseases also requires building and mapping models. If one visit is sufficient to establish a diagnosis of flu or an injury such as a broken arm, the diagnoses of other diseases can be modeled in data dimensions. A time dimension is involved when a certain number of similar visits occurring within a specified period of time identifies a disease. For example, for a heartburn more than one visit within a 6-week period can be treated as a confirmed case [56]. More complex two-dimensional models describe cases of diabetes and cancer, in which sequences of heterogeneous records, such as a visit followed by a set of procedures, lab-tests with certain results and treatments establish a confirmed case. Interpretation of these patterns by domain specialists does not present a problem for a few cases – the challenge emerges in integrating thousands of disparate records. These diverse models of confirmed disease cases need to be presented to the end users, as illustrated by Fig. 11, to enable interpretation of disparate data. All standards including those for data collection are based on implicit models that need to be mapped in order to unite data collected based on them.

Integrating medical data about diverse social groups and countries advances us to another level of sophistication [11,67] that should include models of social behavior. For example, because military personnel have to report to the doctor's office if they do not report to duty, a clinical record exists for every disease occurrence. An employed civilian may take a sick day or get an over-the-counter drug to deal with an ailment without documentation. A majority of disease cases among children and unemployed or self-employed individuals remain undocumented, unless a transaction at pharmacy can be directly linked to these individuals. In other cultures, sick people can go to a monastic retreat or perform other activities that are believed to help. These examples demonstrate that the global integration of clinical records, and especially early diagnostics of disease outbreaks, presents a significant challenge due to culture-dependant responses to diseases. Cross-culture concept mapping is well illustrated through comparing recommendation letters [68], which shows how cultural differences are reflected in both the choice of important concepts and the ways to communicate them. Comparative account of cultures, while rarely discussed in health service literature, presents a mandatory component for global integration of health-related data [69].

Integration with other knowledge domains, including genomic, proteomic and images (e.g. MRI, CAT scan), while enriching possible discoveries tenfold, further complicates the data integration process. Problems of database heterogeneity equally apply to clinical data describing individual patients and biological data characterizing our genome [35], while high-throughput data are in poor agreement, even between experiments of the same type [70].

V. Conclusion

The paper introduces a framework of data dimensions, authoritative data sources, focal data elements and essentials of presenting data to the users. While the mainstream integration approaches enable efficient merging of heterogeneous data, the presented framework helps to assess the informational value of data, which is largely defined by the relationship between the data and the reality they represent. Standards, ontology, reference models, semantic relationships and other integration technologies [71,72] are necessary but not sufficient for obtaining reliable information. They provide guidance on putting together building blocks of data, i.e. DEs, while the quality of these blocks must be assessed as well. Metaphorically speaking, if these blocks are made of sand, the integrated structure will not stand.

We process data collected by people with many human characteristics imprinted in them. To produce reliable information we need to associate data with real life processes and practices.

Identifying factors that influence the informational value of data enables us to create strategies for extracting information from incomplete or partially unreliable data. These strategies gain special importance with the emergence of collaborative technologies such as wikis and blogs, where many people of varied expertise and intentions have an equal opportunity to contribute.

These collections of unverified information are, nevertheless, useful, popular and easily accessed by powerful search engines like Google™. Moreover, the online encyclopedia Wikipedia is about as accurate in covering scientific topics as written by 4000 experts Encyclopedia Britannica [73]. A further study of information quality, accuracy and completeness distributions among articles may help us to develop methods for evaluating the reliability of data sources and information models that account for incompleteness and partial unreliability of data.

Informational value of data depends on the purpose. If the goal of data processing and interpretation differs significantly from the original goal of data collection, the value can be quite low. It can be assessed based on the overlap of focal DEs for each of these cases. In a situation when collection supports billing and when processing aims at identifying risk factors for a disease group, the overlap of DEs may not be sufficient to support the latter.

Another application of the principles developed in this paper is the harmonization of business practices through identifying data quality dips. Every DE that misrepresents reality is associated with an inadequate business process where the requirements for data collections are not balanced with the incentive to do it right. If a doctor has 15 minutes to see you, would you like him to spend all this time with you or would you rather prefer that he spends 5 minutes with you and 10 minutes on filling out forms? Modern technologies, e.g. voice recognition in this case, may assist greatly in harmonization of business practices, while low quality data help us to identify them. Implementation of standards should also be leveraged against human factors and business practices. The situations where there are too many codes to remember, too little time to find the correct one and too little incentive to do it right will feed many generations of data miners in making sense of these misleading data. Ensuring consistency of DEs can be done based on a simple automated procedure (an on-line registration is not complete unless a confirmation email is received from the registered email address) or a user interest to have it right (an employee paid through a direct deposit will make sure that the bank account is recorded correctly because he is interested in receiving his paycheck).

Our attempts to create universally accepted definitions impede employing of controlled vocabularies in our daily data processing. Developing project specific vocabularies and conceptual models, which define the relationship between a project and the rest of the world, will facilitate integration of data from diverse sources. MAGE-OM [74] in support of MIAME standards presents an attempt to employ this approach.

While mathematics and physical sciences developed robust conceptual models, biology and medicine still remain largely fragmented and phenomenological. This situation can be attributed to several factors such as complex structures and high dimensionality of biomedical data; evolving concepts and insufficient data modeling practices; limited applicability of mechanistic models in biomedical domain [75]. One of the most important factors is that while in the physical world, the spatio-temporal presentation of data is natural, a special conceptualization effort is required in choosing dimensions for representation and modeling of high-dimensional biomedical data. Conceptual modeling employs abstraction, aggregations and idealization [76] of objects to identify simple structures or behaviors and enables extracting of basic and invariant features out of the overall complexity. A CDM allows us to simplify presentation to the level of detail appropriate for a specific audience and purpose. A CDM with identified data dimensions can assist in resolving many challenges of clinical data integration from heterogeneous systems that include partitioning or grouping data for different purposes, e.g. data management or de-identification; the necessity to distinguish between clinically approved and research data, where the use on the later could be illegitimate for prescribing treatments; the need to compare results of studies that follow different protocols; creating rules on disclosing combinations of data elements that preserve privacy of patients; capturing the difference between a study design and how it is actually conducted [14] and many others [35,77].

Data dimensions identify a space where we can apply mathematical approaches [14] to the process of data integration. With the defined metrics, the data space can be treated as a topological space. Integration of topographical data [72], for example, requires "the data sets to be integrated are from the same geographic space and represent the same 'snapshot' in time." Similar requirements should be developed for the multidimensional space of biomedical data, which means we must define and understand dimensions involved in every integration endeavor.

In order to take advantage of the abundant biomedical data, we need to develop information architecture that enables us to assemble an information jigsaw puzzle from the clean and reliable pieces. We need to appraise the internals of data processing: the origin of data, purposes of data collection, business practices, and to identify data dimensions, authoritative data sources, focal data elements, integration keys and underlying models.

The authors hope that understanding of the anatomy of data integration will assist project managers, informaticians, data architects, data warehouse and health information system designers in producing reliable information from the abundance of biomedical data.

Acknowledgements

We would like to thank Devin Satz, Wesley Watkins, Stavros Kalamatianos, and Corey Ritchie for productive collaboration; and express our gratitude to Steve Thornton, Ksenia and Konstantin Brazhnik for editing help.

References

1. A Guide to the Project Management Body of Knowledge. Project Management Institute, Inc.; 2004.
2. Pietka E. Large-Scale Hospital Information System in clinical practice. International Congress Series 2003;1256:843.
3. Giuse DA, Kuhn KA. Health information systems challenges: the Heidelberg conference and the future. International Journal of Medical Informatics 2003;69(23):105. [PubMed: 12810116]
4. Combi C, Oliboni B, Rossato R. Merging multimedia presentations and semistructured temporal data: a graph-based model and its application to clinical information. Artificial Intelligence in Medicine 34 (2):89. [PubMed: 15894175]
5. Grosu A-L, et al. Validation of a method for automatic image fusion (BrainLAB System) of CT data and 11C-methionine-PET data for stereotactic radiotherapy using a LINAC: first clinical experience. International Journal of Radiation Oncology*Biology*Physics 2003;56(5):1450.
6. Moret-Bonillo V, Cabrero-Canosa MJ, Hernandez-Pereira EM. Integration of data, information and knowledge in intelligent patient monitoring. Expert Systems with Applications 1998;15(2):155.
7. Tsiknakis M, Katehakis DG, Orphanoudakis SC. An open, component-based information infrastructure for integrated health information networks. International Journal of Medical Informatics 2002;68(13): 3. [PubMed: 12467787]
8. HL7 Standard: RIM data model. http://www.hl7.org/library/data-model/index.cfm. cited; Available from: http://www.hl7.org/library/data-model/index.cfm
9. CDISC Standard. http://www.cdisc.org/. cited; Available from: http://www.cdisc.org/
10. Ingenerf J, Reiner J, Seik B. Standardized terminological services enabling semantic interoperability between distributed and heterogeneous systems. International Journal of Medical Informatics 2001;64(23):223. [PubMed: 11734388]
11. Tsoumakas G, Angelis L, Vlahavas I. Clustering classifiers for knowledge discovery from physically distributed databases. Data & Knowledge Engineering 2004;49(3):223.
12. Coplan JD, et al. Nocturnal growth hormone secretion studies in adolescents with or without major depression re-examined: integration of adult clinical follow-up data. Biological Psychiatry 2000;47 (7):594. [PubMed: 10745051]

13. Hannan TJ. Variation in health care--the roles of the electronic medical record. International Journal of Medical Informatics 1999;54(2):127. [PubMed: 10219952]

14. Brazhnik O. Databases and the geometry of knowledge. Data & Knowledge Engineering. In Press, Corrected Proof

15. Glykas M, Chytas P. Next generation of methods and tools for team work based care in speech and language therapy. Telematics and Informatics 2005;22(3):135.

16. Mendonca EA, et al. Accessing Heterogeneous Sources of Evidence to Answer Clinical Questions. Journal of Biomedical Informatics 2001;34(2):85. [PubMed: 11515415]

17. Davidov E, et al. Advancing drug discovery through systems biology. Drug Discovery Today 2003;8 (4):175. [PubMed: 12581712]

18. Hamadeh HK, et al. Integration of clinical and gene expression endpoints to explore furan-mediated hepatotoxicity. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 2004;549(12):169.

19. Molidor R, et al. New trends in bioinformatics: from genome sequence to personalized medicine. Experimental Gerontology 2003;38(10):1031. [PubMed: 14580855]

20. Dugas M, et al. Complexity of biomedical data models in cardiology: the Intranet-based AF registry. Computer Methods and Programs in Biomedicine 2002;68(1):49. [PubMed: 11886702]

21. Muller ML, et al. Towards integration of clinical decision support in commercial hospital information systems using distributed, reusable software and knowledge components. International Journal of Medical Informatics 2001;64(23):369. [PubMed: 11734398]

22. Kalogeropoulos DA, Carson ER, Collinson PO. Towards knowledge-based systems in clinical practice: Development of an integrated clinical information and knowledge management support system. Computer Methods and Programs in Biomedicine 2003;72(1):65. [PubMed: 12850298]

23. Borst F, et al. Happy birthday DIOGENE: a hospital information system born 20 years ago. International Journal of Medical Informatics 1999;54(3):157. [PubMed: 10405876]

24. Boaz D, Shahar Y. A framework for distributed mediation of temporal-abstraction queries to clinical databases. Artificial Intelligence in Medicine 2005;34(1):3–24. [PubMed: 15885563]

25. Pietka E, Gertych A, Witko K. Informatics infrastructure of CAD system. Computerized Medical Imaging and Graphics 2005;29(23):157. [PubMed: 15755535]

26. Adlassnig K-P. The Section on Medical Expert and Knowledge-Based Systems at the Department of Medical Computer Sciences of the University of Vienna Medical School. Artificial Intelligence in Medicine 2001;21(13):139. [PubMed: 11154878]

27. Scott JA, et al. Integration of clinical and imaging data to predict the presence of coronary artery disease using neural networks. Journal of Nuclear Cardiology 2004;11(4):S26.

28. Lovell NH, Celler BG. Information technology in primary health care. International Journal of Medical Informatics 1999;55(1):9. [PubMed: 10471237]

29. Kuroki T, et al. Data Management and Integration of Clinical Laboratory Information: Current Status and Future Perspectives of Physiological Function Examination Systems in Japan. Journal of the Association for Laboratory Automation 2000;5(5):53.

30. van Bemmel JH, et al. Databases for knowledge discovery: Examples from biomedicine and health care. International Journal of Medical Informatics. In Press, Corrected Proof

31. Korn, F.; Shneiderman, B. Navigating Terminology Hierarchies to Access a Digital Library of Medical Images HCIL-96-01. University of Maryland; 1996.

32. Famili A, S W-M, Weber Richard, Simoudis Evangelos. Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis 1997;1:3–23.

33. Hand DJ. Intelligent Data Analysis: Issues and Opportunities. Intelligent Data Analysis 1998;2:67–79.

34. Espino, et al. Removing a Barrier to Computer-Based Outbreak and Disease Surveillance: The RODS Open Source Project. Morbidity and Mortality Weekly Report 2004;53(Supplement 1):32–39. [PubMed: 15714624]

35. Sujansky W. Heterogeneous Database Integration in Biomedicine. Journal of Biomedical Informatics 2001;34(4):285. [PubMed: 11977810]

36. He Q, Ling TW. An ontology based approach to the integration of entity-relationship schemas. Data & Knowledge Engineering. In Press, Corrected Proof

37. Dinu V, Nadkarni P, Brandt C. Pivoting approaches for bulk extraction of Entity-Attribute-Value data. Computer Methods and Programs in Biomedicine 2006;82(1):38. [PubMed: 16556470]

38. Combi C, Chittaro L. Abstraction on clinical data sequences: an object-oriented data model and a query language based on the event calculus. Artificial Intelligence in Medicine 1999;17(3):271. [PubMed: 10564844]

39. Schewe K-D, Thalheim B. Conceptual modelling of web information systems. Data & Knowledge Engineering 2005;54(2):147–188.

40. Elder, J. The Top Ten Data Mining Mistakes -- and How to Avoid Them. The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003; Washington, DC.

41. Sen A. Metadata management: past, present and future. Decision Support Systems 2004;37(1):151.

42. Gene Ontology. http://www.geneontology.org/. cited; Available from: http://www.geneontology.org/

43. Wang C-B, et al. Design of a Meta Model for integrating enterprise systems. Computers in Industry 2005;56(3):305.

44. Reasner, DS. e-Clinical Interchange. Arlington, VA: 2004. Standardization and Integration of Data in the Small Company Environment.

45. Saake G, Sattler K-U, Conrad S. Rule-based schema matching for ontology-based mediators. Journal of Applied Logic 2005;3(2):253–270.

46. Torque Map. http://www.devaki.org/nextobjects/torque-type-map.html. cited; Available from: http://www.devaki.org/nextobjects/torque-type-map.html

47. Buneman P, Sanjeev K, Keishi Tajima, Wang-Chiew Tan. Archiving Scientific Data. ACM Transactions on Database Systems 2004;29(1):2–42.

48. Wang YF, Wang JF. On 3D Model Construction by Fusing Heterogeneous Sensor Data. CVGIP: Image Understanding 1994;60(2):210.

49. Schallehn E, Sattler K-U, Saake G. Efficient similarity-based operations for data integration. Data & Knowledge Engineering 2004;48(3):361.

50. Theodoratos D, Ligoudistianos S, Sellis T. View selection for designing the global data warehouse. Data & Knowledge Engineering 2001;39(3):219.

51. Truta, TM.; F, F.; Barth-Jones, D. Disclosure Risk Measures for Microdata; 15th International Conference on Scientific and Statistical Database Management; 2003; Cambridge, MA, USA.

52. Domingo-Ferrer J, Torra V. On the connections between statistical disclosure control for microdata and some artificial intelligence tools. Information Sciences 2003;151:153.

53. Palta JR, Frouhar VA, Dempsey JF. Web-based submission, archive, and review of radiotherapy data for clinical quality assurance: A new paradigm. International Journal of Radiation Oncology*Biology*Physics 2003;57(5):1427.

54. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics 2004;37(3):179. [PubMed: 15196482]

55. Dash M, H L. Feature Selection for Classification. Intelligent Data Analysis. 1997

56. CDC. http://www.cdc.gov/node.do/id/0900f3ec8000e035. cited; Available from: http://www.cdc.gov/node.do/id/0900f3ec8000e035

57. Bayer B, Schneider R, Marquardt W. Integration of data models for process design -- first steps and experiences. Computers & Chemical Engineering 2000;24(27):599.

58. Yi S, Huang B, Tat Chan W. XML application schema matching using similarity measure and relaxation labeling. Information Sciences 2005;169(12):27.

59. Wen-Syan, Chris. SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. Data & Knowledge Engineering 2000;33(1):49.

60. Knapp JL. ER isomorphisms and uniqueness conditions. Data & Knowledge Engineering 1998;26:271–290.

61. Kwan I, Li Q. A hybrid approach to convert relational schema to object-oriented schema. Information Sciences 1999;117(34):201.

62. Abiteboul S, Cluet S, Milo T. Correspondence and translation for heterogeneous data. Theoretical Computer Science 2002;275(12):179.
63. Bernstein PA, E R. A survey of approaches to automatic schema matching. The VLDB Journal 2001;10(4):334–350.
64. Sun Y. Methods for automated concept mapping between medical databases. Journal of Biomedical Informatics 2004;37(3):162. [PubMed: 15196481]
65. Diamond GA, Kaul S. Prior convictions: bayesian approaches to the analysis and interpretation of clinical megatrials. Journal of the American College of Cardiology 2004;43(11):1929. [PubMed: 15172393]
66. B H, Bowes WA III, Holston FT, Gundersen MLCPDNSPHSMPTAHPJLTMSERSR. Building a comprehensive clinical information system from components: The approach at Intermountain Health Care. Methods of Information in Medicine 2003;42(1):1–7. [PubMed: 12695790]
67. Kimball, AM. New Horizons in Molecular Sciences and Systems: An Integrated Approach. Okinawa, Japan: 2003. Surveillance and Information Management of Infectious Disease Outbreaks.
68. Precht K. A cross-cultural comparison of letters of recommendation. English for Specific Purposes 1998;17(3):241.
69. Braithwaite J, et al. A tale of two hospitals: assessing cultural landscapes and compositions. Social Science & Medicine 2005;60(5):1149. [PubMed: 15589681]
70. Jensen LJ, Steinmetz LM. Re-analysis of data and its integration. FEBS Letters 2005;579(8):1802. [PubMed: 15763555]
71. Simon J, et al. Formal ontology for natural language processing and the integration of biomedical databases. International Journal of Medical Informatics 2006;75(34):224. [PubMed: 16153885]
72. Uitermark HT, et al. Ontology-based integration of topographic data sets. International Journal of Applied Earth Observation and Geoinformation 2005;7(2):97.
73. Giles J. Internet encyclopaedias go head to head. Nature 2005;438(7070):900. [PubMed: 16355180]
74. MAGE. http://www.mged.org/Workgroups/MAGE/mage-om.html. cited; Available from: http://www.mged.org/Workgroups/MAGE/mage-om.html
75. Rutishauser R, Moline P. Evo-devo and the search for homology ("sameness") in biological systems. Theory in Biosciences 2005;124(2):213. [PubMed: 17046357]
76. Carson, ER.; Cobelli, C.; Finkelstein, L. The Mathematical Modeling of Metabolic and Endocrine Systems. New York, NY: John Wiley & Sons; 1983.
77. Famili A, et al. Data preprocessing and intelligent data analysis. Intelligent Data Analysis 1997;1(14):3.

APPENDIX B. List of abbreviations

DB	database
ER	emergency room
OLTP	online transaction processing
OLAP	online analytical processing
DE	data element
SSN	social security number
DUA	

data use agreement

HIPAA

the Health Insurance Portability and Accountability Act of 1996

ODS

operational data store

DWH

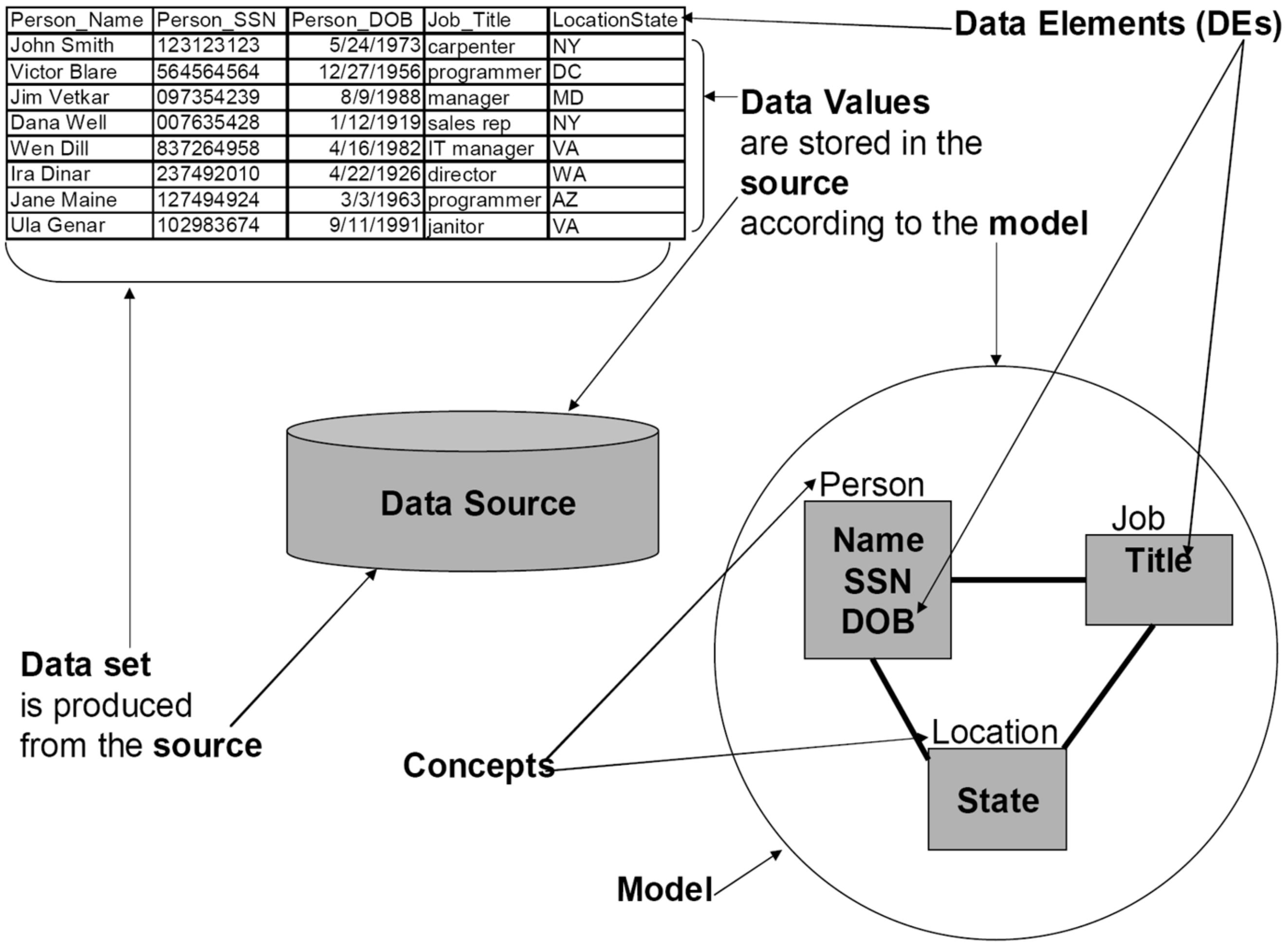
data warehouse

ID

identifier

ETL

extract – transfer – load



Person_Name	Person_SSN	Person_DOB	Job_Title	LocationState
John Smith	123123123	5/24/1973	carpenter	NY
Victor Blare	564564564	12/27/1956	programmer	DC
Jim Vetkar	097354239	8/9/1988	manager	MD
Dana Well	007635428	1/12/1919	sales rep	NY
Wen Dill	837264958	4/16/1982	IT manager	VA
Ira Dinar	237492010	4/22/1926	director	WA
Jane Maine	127494924	3/3/1963	programmer	AZ
Ula Genar	102983674	9/11/1991	janitor	VA

Data Elements (DEs)
Data Values are stored in the source according to the model
Data Source
Data set is produced from the source
Concepts
Model
Person
Name
SSN
DOB
Job
Title
Location
State

Fig. 1.
Illustration of main modalities of data integration.

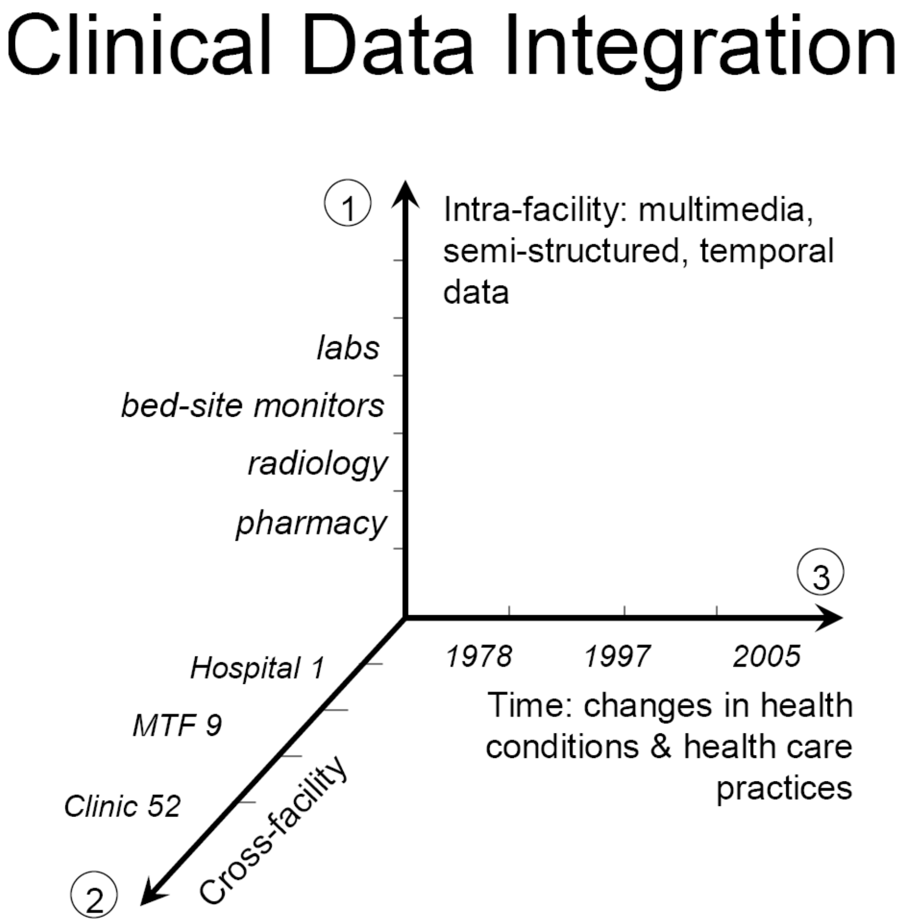

Clinical Data Integration
1
Intra-facility: multimedia, semi-structured, temporal data
labs
bed-site monitors
radiology
pharmacy
3
1978
1997
2005
Time: changes in health conditions & health care practices
Hospital 1
MTF 9
Clinic 52
Cross-facility
2

Fig. 2.
Dimensions of clinical data

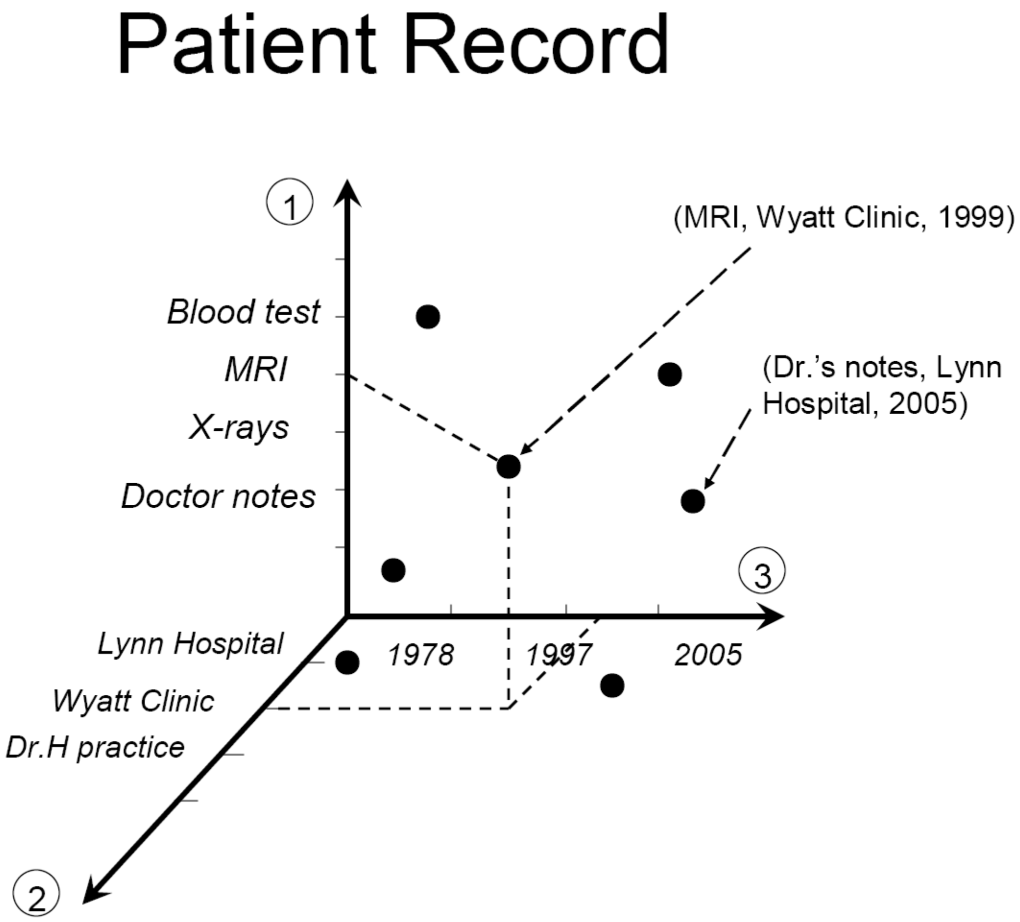


Fig. 3.
3D patient record

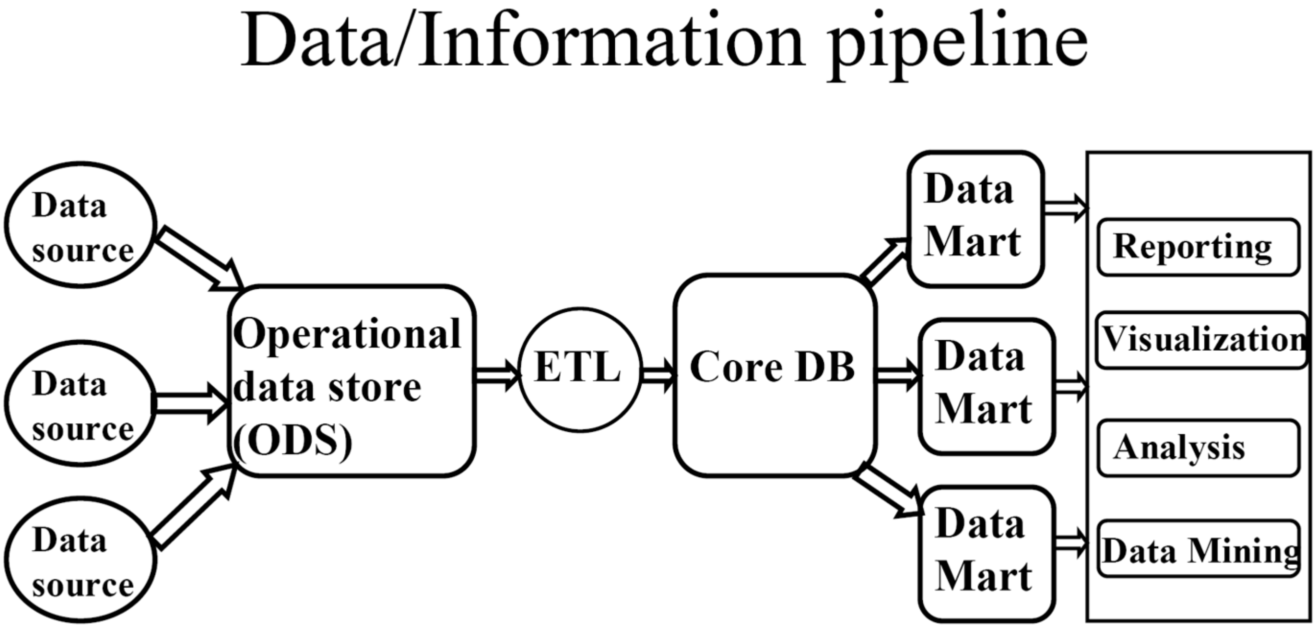


Fig. 4.
Information pipeline

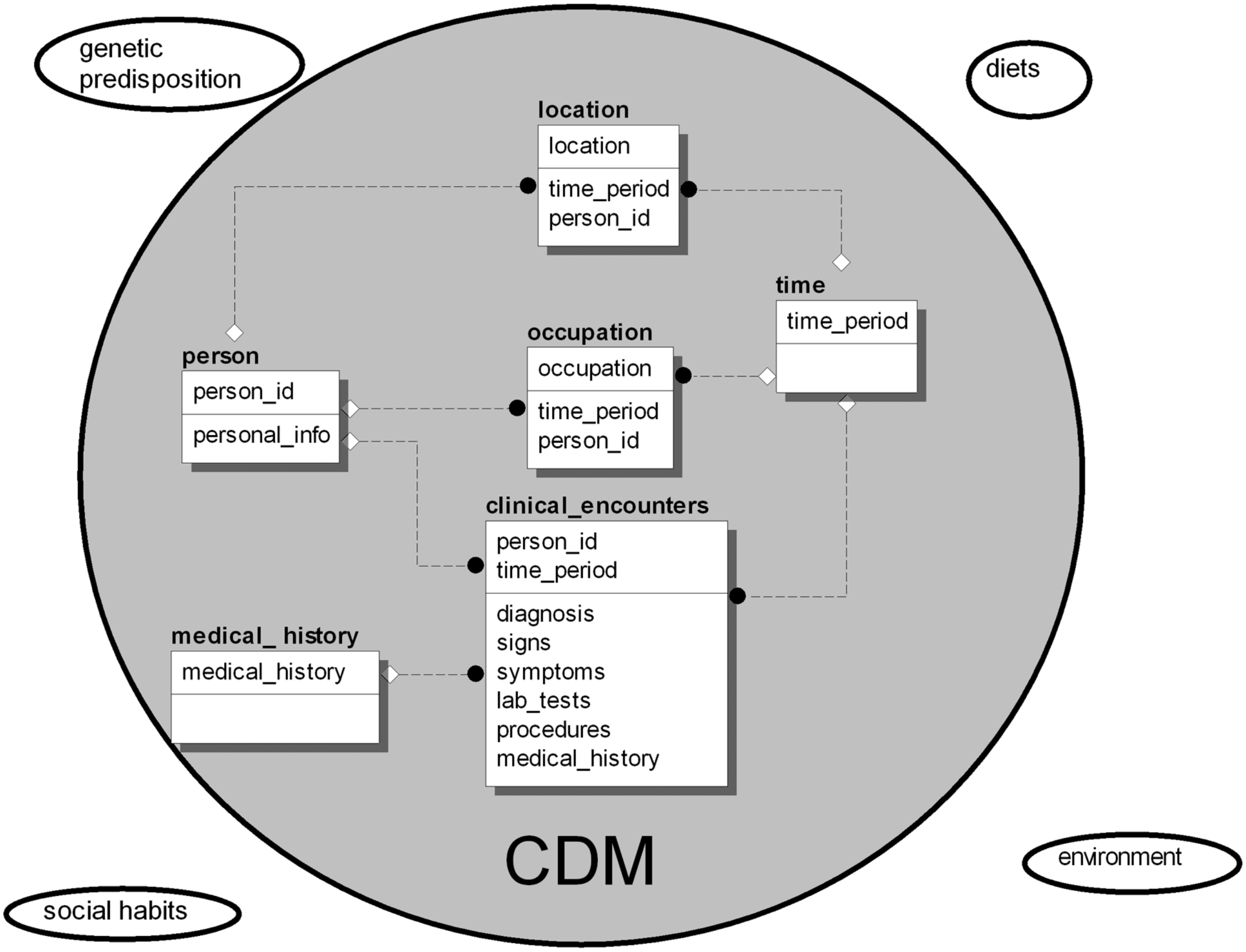


Fig. 5.
A conceptual data model (CDM) defines the main concepts included into and excluded from the study.

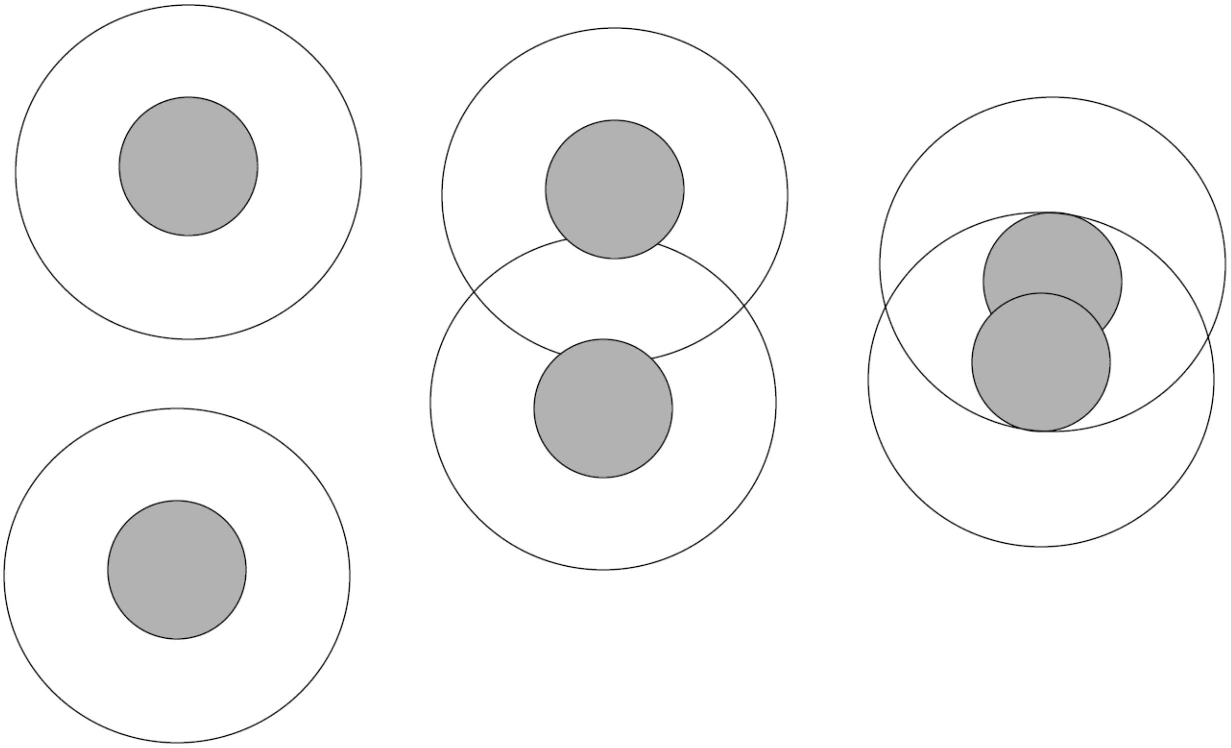


Fig. 6.
Meaningful integration is possible only between data sources with overlapping focal DE.

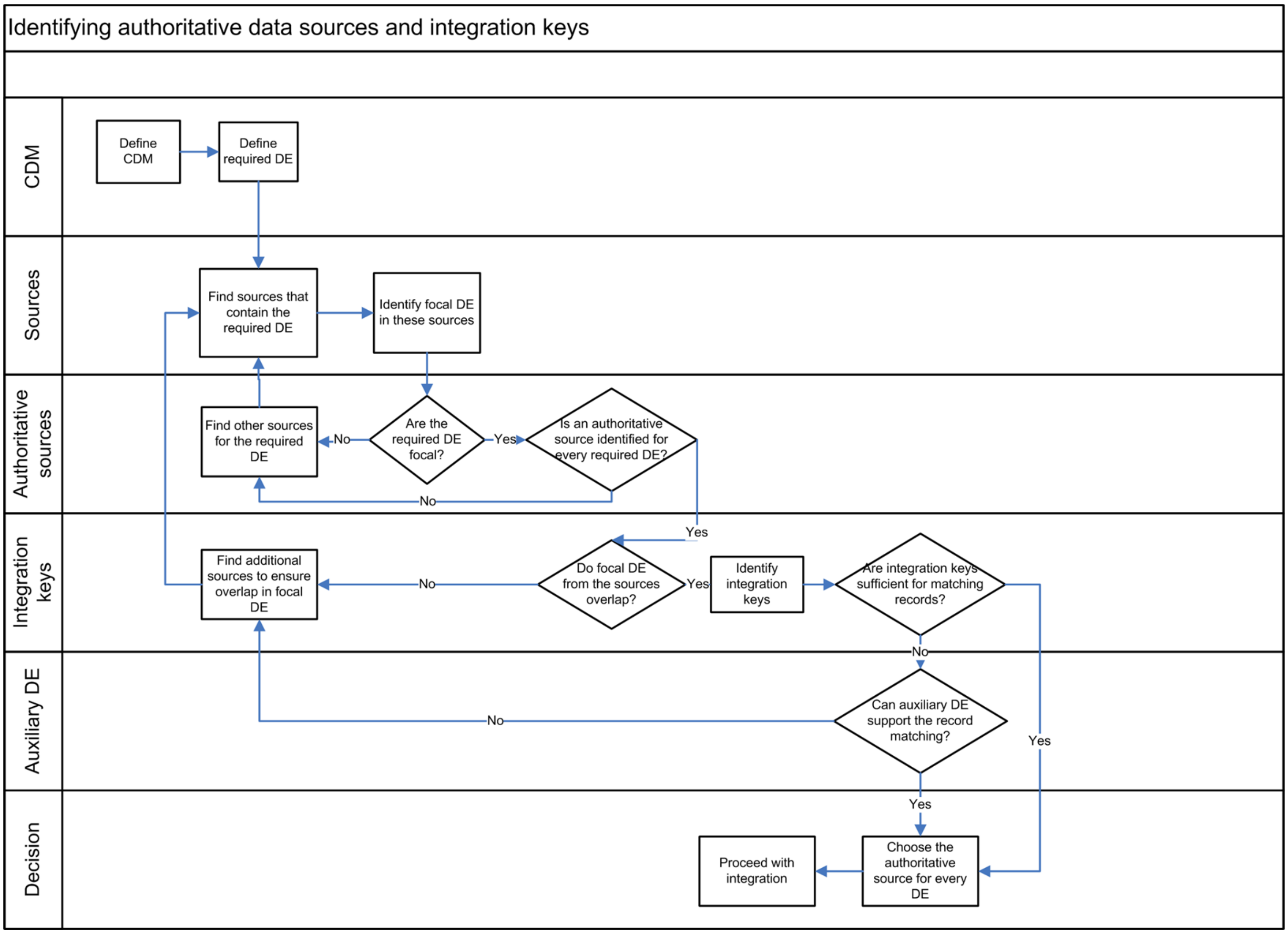
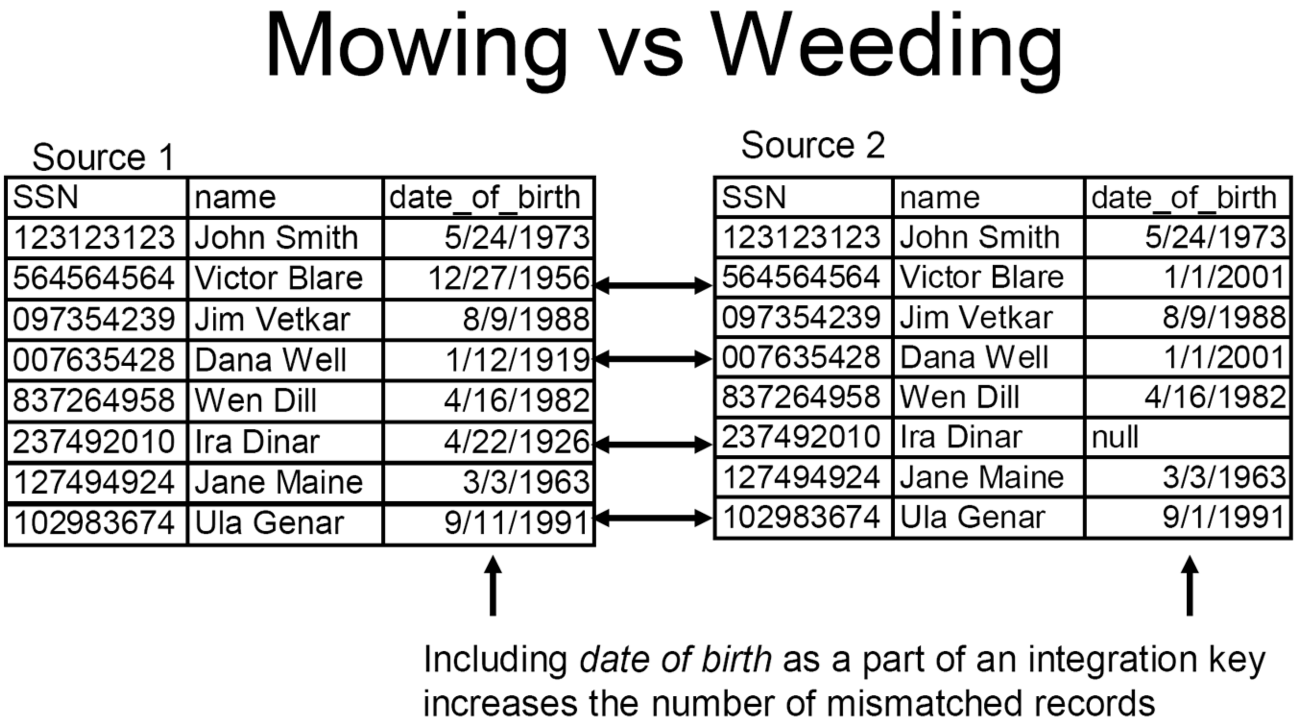

Identifying authoritative data sources and integration keys
CDM
Sources
Authoritative sources
Integration keys
Auxiliary DE
Decision
Define CDM
Define required DE
Find sources that contain the required DE
Identify focal DE in these sources
Find other sources for the required DE
No
Are the required DE focal?
Yes
Is an authoritative source identified for every required DE?
No
Yes
Find additional sources to ensure overlap in focal DE
No
Do focal DE from the sources overlap?
Yes
Identify integration keys
Are integration keys sufficient for matching records?
No
Yes
No
Can auxiliary DE support the record matching?
Yes
Proceed with integration
Choose the authoritative source for every DE

Fig. 7.
The process of identifying authoritative data sources and focal DEs



Mowing vs Weeding

Source 1

SSN	name	date_of_birth
123123123	John Smith	5/24/1973
564564564	Victor Blare	12/27/1956
097354239	Jim Vetkar	8/9/1988
007635428	Dana Well	1/12/1919
837264958	Wen Dill	4/16/1982
237492010	Ira Dinar	4/22/1926
127494924	Jane Maine	3/3/1963
102983674	Ula Genar	9/11/1991

Source 2

SSN	name	date_of_birth
123123123	John Smith	5/24/1973
564564564	Victor Blare	1/1/2001
097354239	Jim Vetkar	8/9/1988
007635428	Dana Well	1/1/2001
837264958	Wen Dill	4/16/1982
237492010	Ira Dinar	null
127494924	Jane Maine	3/3/1963
102983674	Ula Genar	9/1/1991

Including *date of birth* as a part of an integration key increases the number of mismatched records

Fig. 8.

DOB in this example has many invalid entries. Excluding DOB will increase the chance for successful integration.

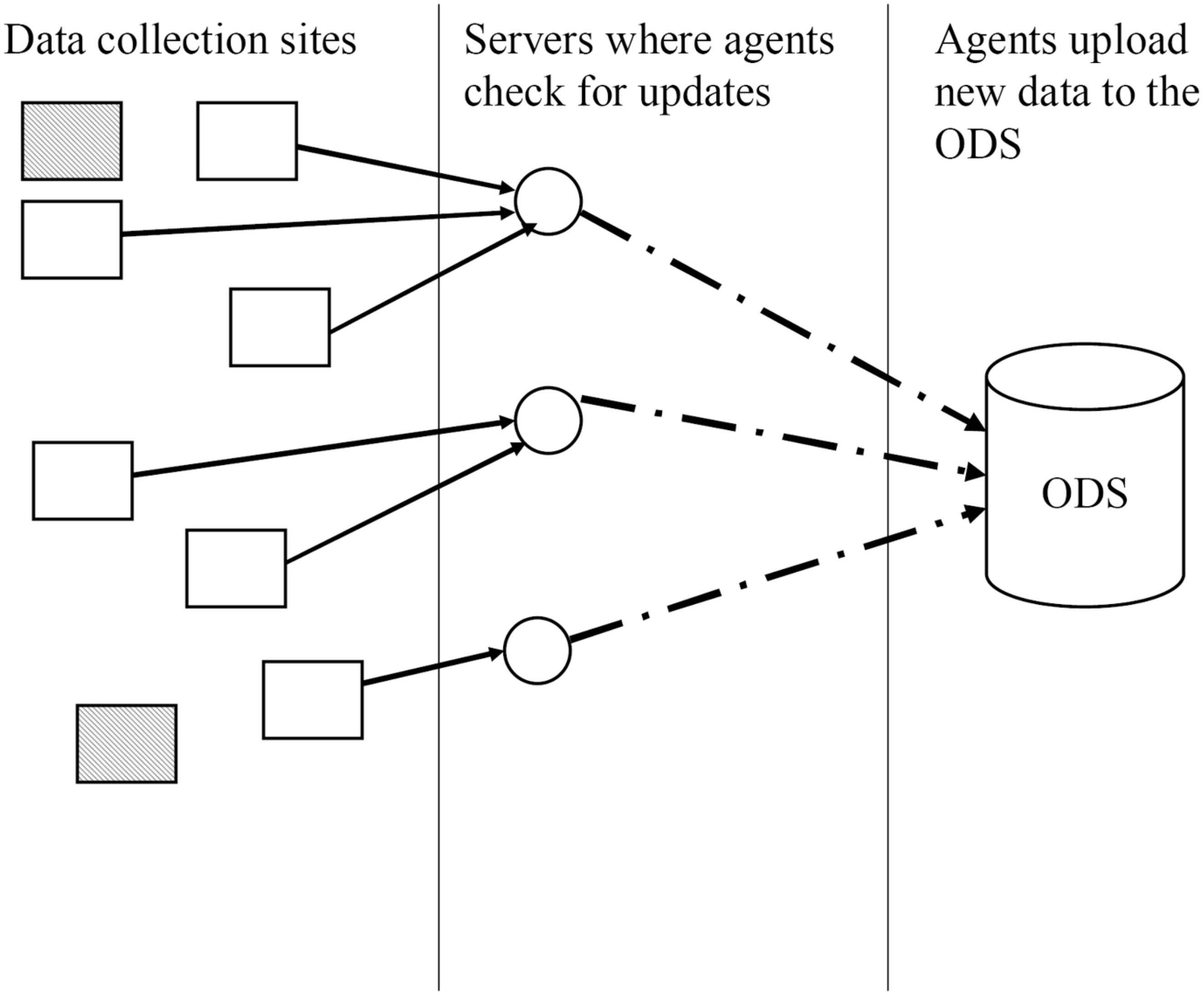

Data collection sites
Servers where agents check for updates
Agents upload new data to the ODS
ODS

Fig. 9.
Data gathering agents provide only a part of the data pool. If agents act on servers that receive updates from multiple sites there is often no information on which sites provided and which ones did not provide the updates.

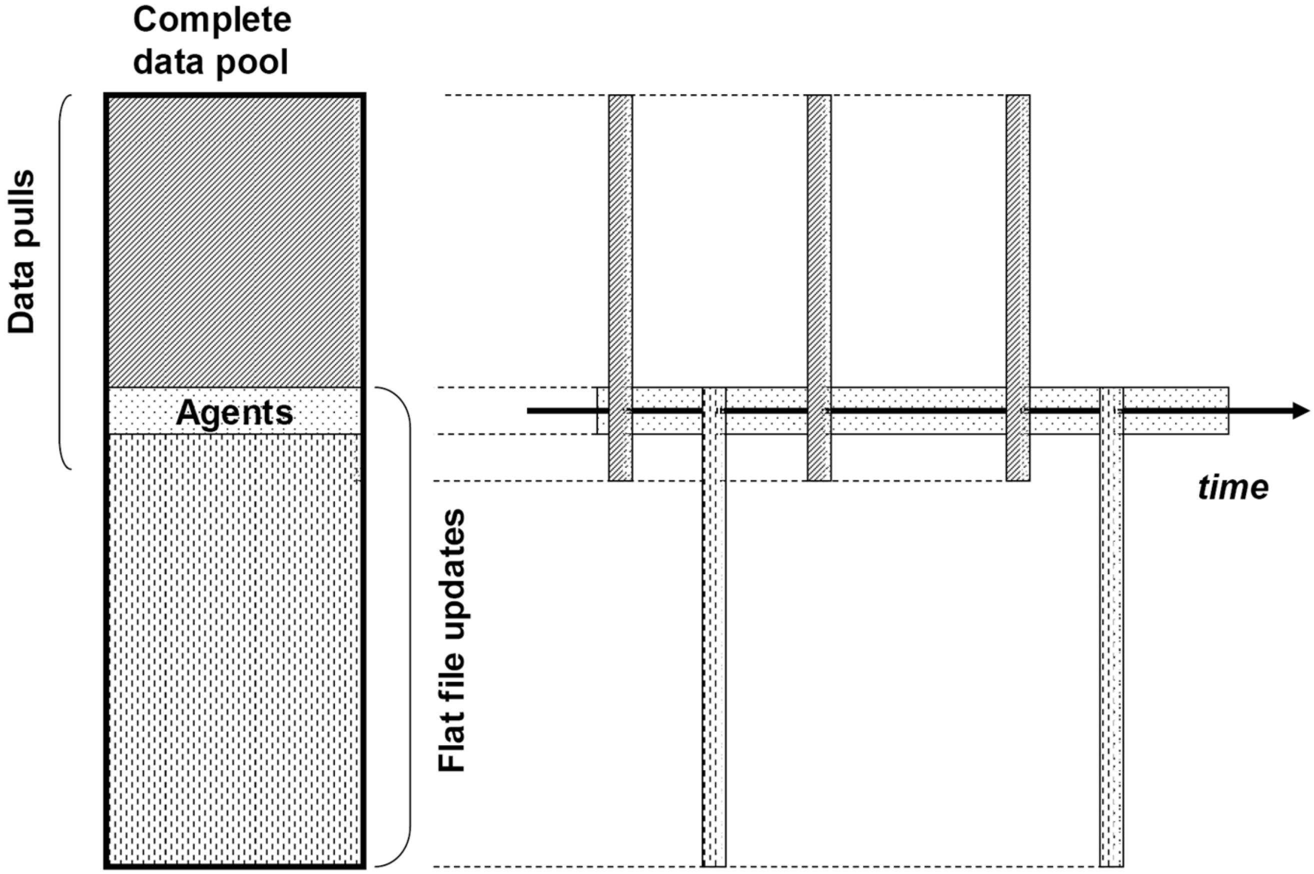


Fig. 10.

While agents provide near-real-time updates, the complete data pool is defined only at the time of the update which brought the pool to its fullness. For example, if flat file updates are received on the first of each month, data pulls are performed on Mondays, and agents deliver data in near real-time. Then on Friday the 27th, a consistent summary can be produced as of the first of the month, in this case 26 days ago, when the data pool was complete.

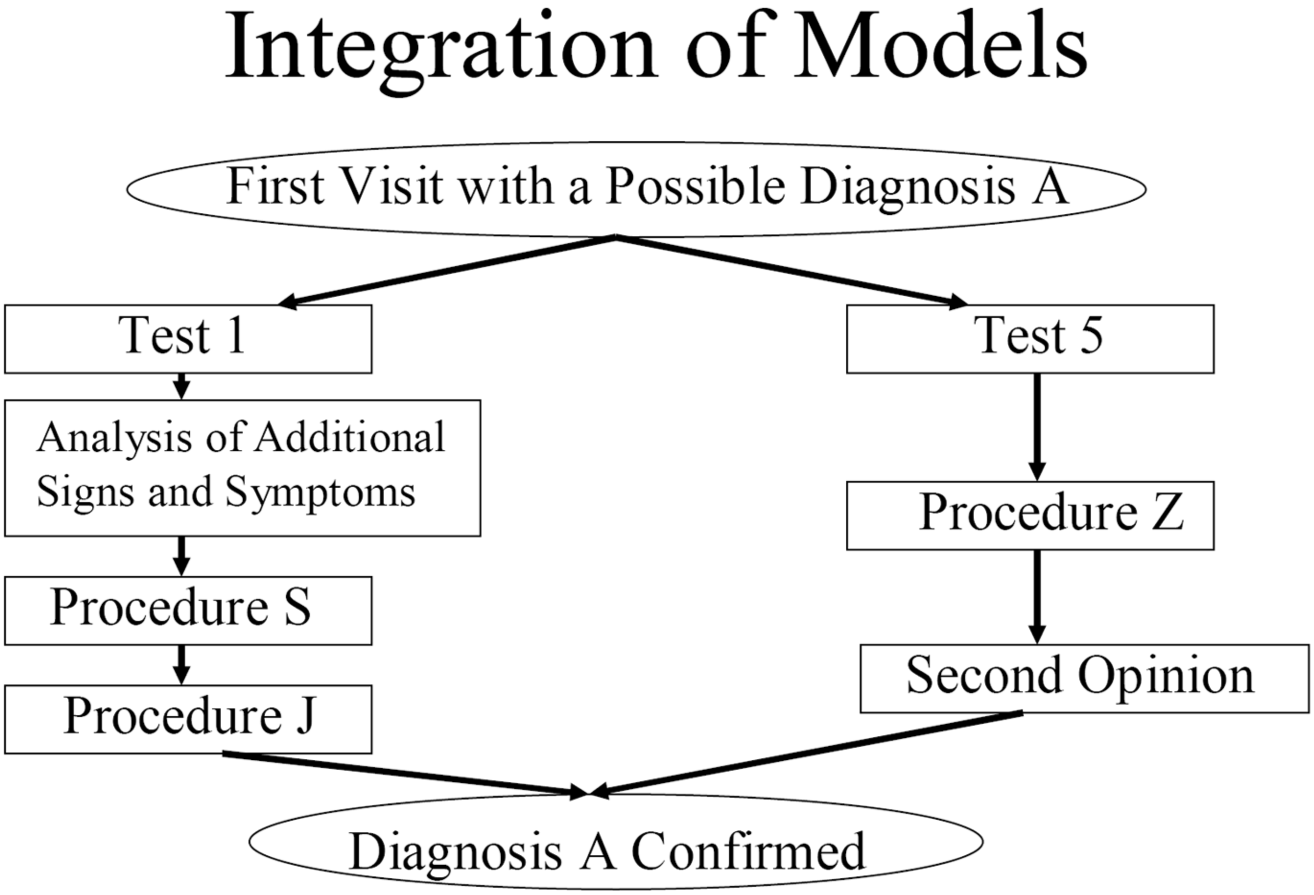


Fig. 11.
Semantic equivalency should be established between confirmed cases in diverse disease models.

TABLE 1

Characteristics of Data Sources that influence Project decisions

Project needs	**Characteristics of data sources**
What DEs are reliable?	business processes supported by data source
	the purpose of data collection and processing
	primary or secondary
	business rules for data collection & processing
	dependencies between DEs
How to get data from the source?	availability of data (public or proprietary)
	the type of the data source: OLTP or OLAP
	frequency of updates
	database management systems and other enabling technologies employed
How to meaningfully integrate with other sources?	controlled vocabularies
	underlying models, concepts, ontology
	frequency of updates
	part of the world represented, i.e. subsets of population, occupations, disease code, etc., captured in the data source

TABLE 2

Integrating Controlled Vocabularies

DB name	Primary Diagnosis	Patient’s date of birth
In-patient DB	Primary_Diagnosis_Code	Patient_DOB
Hospital DB	Diagnosis_1	Date_of_Birth
ERDB	Primary Diagnosis	DateofBirth
Mobile care DB	Primary_Diagnosis_Code	DOB

TABLE 3

Integrative, Informative and Auxiliary DE

PCDB	**Personnel**	**ER**	
hiph_ssn	**ssn**	**ssn_or_surrogate_id**	**Integration Keys**
first_name	**first_name**	**first_name**	
last_name	**last_name**	**last_name**	
gender	**sex**	**patient_sex**	
date_of_birth	**DOB**		
appointment_date	*job title*	*vital_signs*	*Informative data elements*
primary_diagnosis	*job location*	*symptoms*	
secondary_diagnosis	*job_start*	*diagnosis*	
symptom	*job_end*	*procedures*	
ethnicity	race marital status	race birthdate	Auxiliary DE

Integration Keys are shown in **Bold**, informative DE are in *Italic*, auxiliary ones are in regular print.

TABLE 4

Masking Table: Gender

DB name	Patient’s date of birth
Male	gA
Female	gB

An example of masking table: genders.

TABLE 5

Masked Data

hiph_ssn	appointment_date	date_of_birth	race	gender	primary_diagnosis_group	primary_diagnosis	secondary_diagnosis
ssn144314	3/11/98	01/14/56	rD	gB	diagn_gr_1	diagnosis_16678	diagnosis_26496
			rX	gA			
ssn264545	5/25/03	04/12/60			diagn_gr_1	diagnosis_15039	
ssn155980	1/31/04	12/12/25	*rD*	gA	diagn_gr_1	diagnosis_16678	diagnosis_21286
			rX	gB			
ssn249262	7/7/01	05/17/96			diagn_gr_1	diagnosis_133	diagnosis_26496
ssn241178	12/23/01	03/15/34	*rA*	gB	diagn_gr_1	diagnosis_16678	diagnosis_2227

TABLE 6

Differences in the Structure of Clinical Records

	Diagn_2	**Diagn_3**	**Chief Complaint**	**Symptom**	**Sign**	**case**
ERDB	x	x	x	x	x	o
Mobile care DB	o	o	x	o	x	o
Hospital DB	x	x	o	x	x	x
LabDB	x	o	o	x	o	x

DEs accompanying the primary diagnosis differ from one database to another. **X** shows existing DEs, **o** identifies the absent DEs.