Information Exchange between Medical Databases through Automated Identification of Concept Equivalence

by

Yao Sun

A.B. Human Biology
Stanford University, 1984

M.D.
UCLA School of Medicine, 1989

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2002

Signature of Author: _____
Department of Electrical Engineering and Computer Science
October 18, 2001

Certified by: _____
Peter Szolovits
Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by: _____
Arthur C. Smith
Chairman, Committee for Graduate Students

Information Exchange between Medical Databases through Automated Identification of Concept Equivalence

by

Yao Sun

Submitted to the Department of Electrical Engineering and Computer Science
on October 18, 2001 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Computer Science

ABSTRACT

The difficulty of exchanging information between heterogeneous medical databases remains one of the chief obstacles in achieving a unified patient medical record. Although methods have been developed to address differences in data formats, system software, and communication protocols, automated data exchange between disparate systems still remains an elusive goal.

The Medical Information Acquisition and Transmission Enabler (MEDIATE) system identifies semantically equivalent concepts between databases to facilitate information exchange. MEDIATE employs a semantic network representation to model underlying native databases and to serve as an interface for database queries. This representation generates a semantic context for data concepts that can subsequently be exploited to perform automated concept matching between disparate databases.

To test the feasibility of this system, medical laboratory databases from two different institutions were represented within MEDIATE and automated concept matching was performed. The experimental results show that concepts that existed in both laboratory databases were always correctly recognized as candidate matches. In addition, concepts which existed in only one database could often be matched with more "generalized" concepts in the other database that could still provide useful information.

The architecture of MEDIATE offers advantages in system scalability and robustness. Since concept matching is performed automatically, the only work required to enable data exchange is construction of the semantic network representation. No pre-negotiation is required between institutions to identify data that is compatible for exchange, and there is no additional overhead to add more databases to the exchange network. Because the concept matching occurs dynamically at the time of information exchange, the system is robust to modifications in the underlying native databases as long as the semantic network representations are appropriately updated.

Thesis Supervisor: Peter Szolovits
Title: Professor of Computer Science and Engineering

Acknowledgments

My deepest thanks go to the following people and institutions for their support.

Professor Peter Szolovits, for his wisdom and teaching in matters academic and professional, and his constant encouragement and patience throughout my doctoral endeavor.

Dr. Isaac Kohane, for his research leadership, inspirational role-modeling, and sage career counseling.

Dr. Merton Bernfield, for his mentorship during multiple stages of my academic career, and for his encouragement to pursue further training in the field of computer science.

Dr. Nachman Asch, for sharing programming insights which facilitated the completion of this project.

My colleagues at the Medical Decision Making Group, for their availability and knowledgeable advice.

My colleagues at the Children's Hospital Informatics Program, for providing a friendly environment energized by intellectual stimulation and the open exchange of ideas.

The National Library of Medicine Medical Informatics Training Program, for providing the opportunity to further my education and pursue an expanded horizon of research.

Dr. Kristin Sun, for her tireless enthusiasm and loving support, and my children for providing a wider perspective on life.

TABLE OF CONTENTS

1 INTRODUCTION

As electronic storage of patient medical information increases, the potential for rapid access to the entirety of a patient's medical record offers tantalizing possibilities for improving clinical care and supporting medical research. Patients rarely, however, receive all their medical care from a single provider or facility. Consequently, the electronic medical information for any given patient is commonly scattered across multiple heterogeneous information systems.

The effort to combine or enable access to all these disparate sources of medical information has many obstacles. Techniques have been developed to address basic hardware and software incompatibility issues, but it remains difficult to resolve inconsistencies and conflicts at the semantic level. Subtle distinctions arise even when the same vocabulary is used to describe the same concept. For example, a "thyroid function test panel" (TFTs) at one institution might include a "reverse T3 level", whereas TFTs at a different institution may not.

This investigation demonstrates a new method to combine medical information from disparate electronic sources. The Medical Information Acquisition and Transmission Enabler (MEDIATE) system automatically determines semantic equivalencies between concepts from different databases and enables the retrieval and exchange of data with greater fidelity to the semantic content of the information. Using the previous example, MEDIATE enables the automatic identification of TFTs from any medical laboratory database, and at the same time preserves the unique composition of the test panel for each database.

Fundamentally, MEDIATE facilitates data integration by matching semantically equivalent concepts between medical databases. It performs this task by utilizing a semantic network data structure to represent the elements of a medical database. During information exchange, MEDIATE transmits the semantic network database representations between systems for analysis. By operating on characteristics of the semantic network representations, medical concepts within one information system are automatically linked with concepts from a disparate system through concept matching algorithms.

This process allows a user to retrieve data from multiple information systems without regard to how that data is actually stored within each system. In addition, the information exchange occurs without the need to pre-negotiate the list of data elements to be exchanged, since data equivalencies between the databases are revealed automatically.

MEDIATE's approach to data exchange contrasts with the two most common approaches to sharing medical data: construction of a common data model, and manual system-to-system mapping of data elements.

The use of a common data model works well if the data model is comprehensive (as in small knowledge domains) and requires infrequent modification. Under these circumstances, the work required to exchange data between N databases is order(N) for the mapping between each database and the common data model. In the medical record domain, however, repeated attempts at creating comprehensive data models have failed to gain widespread acceptance. In fact, one of the most ambitious collaborative efforts to create such a model, the Health Level 7 Reference Information Model [1, 2], has completely changed directions to produce a modeling framework instead of an actual data model.

There are other drawbacks to common data models. Modifications to the common model entail modifications to the data mapping process for every database involved in data exchange. This tends to be most problematic when new databases are added, and deleteriously affects the scalability of such systems. In addition, the data mapping process itself may cause the loss of information as data concepts are force-fit to the common model. This affects the semantic fidelity of information transmitted through these systems.

The other common approach to data exchange, direct system-to-system mapping of data elements, is perhaps the method that is most frequently chosen. This occurs because of expediency and the lack of accepted common data models. One disadvantage to this approach is the lack of scalability. This is an issue because each database must be mapped to every other database with which it exchanges data, which makes the amount of work approximately

order(N^2). This approach is also sensitive to modifications in the participating databases, since changes in the data elements may break the mapping links and prevent data exchange.

In comparison, MEDIATE utilizes a dynamic model of data exchange in which semantically equivalent data elements are identified at the time of data transfer. This allows the participating databases to be modified freely, without creating additional work or overhead for eventual data exchange. Adding a new database to the data exchange group only requires creating the semantic network representation for that database.
These functional qualities make MEDIATE easily scalable and robust to changes in the underlying databases, and ease the task of data integration across heterogeneous information systems.

1.1 Problem Motivation

MEDIATE's capability to retrieve and combine all of a patient's medical information offers many potential advantages. It promotes continuity of care by potentially providing a single source of medical information to clinicians, and minimizes the risk that important aspects of the past medical history, such as allergies, previous surgery, or recent diagnoses, may be overlooked. It can also provide the data to populate a longitudinal record to perform clinical and research investigations over time, on an individual or population basis. This longitudinal information forms the ideal substrate for continuous analysis processes, such as trend detection or alerts and warnings.

The following sections list some of the situations in which the ability to integrate medical data from many sources can have an impact.

1.1.1 Clinical Scenarios

Emergency care. A 69 year-old relative who is visiting from another state is found to be lethargic and confused one morning. During evaluation in the local emergency room, the host family can only state that the patient is known to have had recent medical problems. Using a hospital identification card found in the patient's wallet, the treating physician obtains emergency access to the patient's hospital record. Through MEDIATE, the physician is able to locate a set of laboratory tests performed just a week ago that indicate borderline renal function, but normal

hematological and thyroid function. This information allows the physician to focus the diagnostic workup and determine that the patient is suffering from acute renal failure, with a consequent need for emergency dialysis.

Continuity of care. A 2 year-old male with multiple congenital anomalies including structural heart disease, tracheo-esophageal fistula, vertebral anomalies, and renal problems (i.e. VATER syndrome), has an appointment to be seen by his new pediatrician. In order to familiarize herself with the patient's problems and past treatment, the pediatrician uses MEDIATE to retrieve the medical history from several sources: the cardiology foundation computer, the pediatric hospital main computer, and the previous pediatrician's office. The pediatrician locates and reviews the last "progress note" from each of the systems. The information gleaned from these notes enables the pediatrician to establish an efficient agenda for the initial visit without duplicating evaluations that have been performed at the other facilities.

1.1.2 **Research Benefits**

Data collection. Research studies that rely on clinical data often collate information from multiple sources. For example, a recent study of jaundice in young infants seen at Children's Hospital, Boston required maternal and infant data from several different hospitals in which the infants were born [3]. In this situation, the medical information for any single patient is available from a single source, but the research study design requires information from many sources.

Population studies. Large scale population based studies require data collection schemes that often encompass multiple institutions and geographic sites. The Framingham Heart Study, for example, has followed thousands of men through decades of life in a multi-factorial study of heart disease. [4-8] The study subjects have received their medical care in a variety of settings and facilities, and obtaining data about their health status continues to be a major undertaking.

Time series studies. Supporting investigations into the evolution and natural history of medical processes requires a longitudinal medical record that contains observations over time. Due to the peripatetic nature of health care, completely and efficiently populating such a longitudinal record typically necessitates the retrieval of information from many different sources.

1.1.3 **Decision support platform**

The application of clinical support tools has been one of the central promises of an electronic medical record. Trend analysis, automated guidelines, decision support programs, automated alerts, and expert systems for diagnosis and therapy are just a few of the applications that have been created which depend upon complete and accurate data for optimal function. In the vast majority of cases, computerized support tools improve in performance if more data is available for input. Again, integrating all the available sources of medical data would have a beneficial effect on the function of these tools.

1.2 **Obstacles to Data Integration**

As expressed by McDonald, "Each island system [within a healthcare facility] contains different data, different structures, and differing levels of granularity, and each uses a different code system to identify similar clinical concepts. The external islands differ even more than those within an institution. They each tend to use different patient, provider, and location identifiers, and the numbers of such independent systems are legion." [9]

This inconsistency between systems that store medical information presents the main obstacle to integration of medical information. Unfortunately, the inconsistencies exist on multiple levels, each of which may require its own solution. Examples of these levels include different hardware platforms, different types of databases and data models, different communication protocols, and different vocabularies, in addition to the differences listed by McDonald.

The level of inconsistency addressed in this investigation is that of "semantic inconsistency". At this level, many of the inconsistencies listed previously may be resolved, but accurate retrieval of data may still be difficult due to differences in the "meaning" of the medical concepts that are represented within an information system. A "complete blood count" (CBC), for example, may vary in composition from institution to institution, despite the fact that all clinicians would agree that the test consists of an analysis of the cellular elements of blood. This problem of semantic inconsistency has been recognized as a critical obstacle to data integration. [10-26]

Ambiguities inherent in medical terminology and definitions of concepts help create these semantic inconsistencies. It is tempting to hypothesize that a standardized data model that rigorously defines all medical concepts would be sufficient to eliminate semantic inconsistencies. Unfortunately, the majority of medical information existing today does not conform to a standard data model of any kind, and would be difficult to fit into a new global data model. Regardless of the enormity of such a task, however, there are other considerations that make such an undertaking impractical.

One problem is that the semantic meaning of a medical concept is not just determined by the "definition" of that concept, but also by the usage of the concept as determined by the local clinical environment. For example, "sputum cultures" are a standard way to test for respiratory infections. But for hospitals in areas where tuberculosis is endemic, sputum cultures often include tests for the tuberculosis bacterium (such as staining of the sputum for microscopic examination) that would not be run in other settings. The local clinical environment thus determines the meaning and interpretation of the "sputum culture" concept.

This leads to another phenomenon that confounds the use of a global data model: the creation of new semantic meanings. As medical concepts are used and modified for a particular clinical setting, novel semantic meanings are created and assigned. This may even result in a situation like the one at Children's Hospital, Boston where more than a dozen types of "serum sodium" laboratory tests exist.

Although the resolution of semantic inconsistencies is not the only factor in data integration, it is an essential part of the solution and is one of the core principles upon which MEDIATE is based.

1.3 Goals of MEDIATE

The semantic network representation system and concept matching algorithms used in MEDIATE were derived from functional goals delineated during the design stage of the system. In turn, many of the functional goals were generated to preserve the semantic meaning of data as it is transmitted between different information systems. These functional goals are:

1) *Reduce the semantic ambiguity of data transmitted between electronic databases.* The semantic network data representation system accomplishes this goal in several ways. First, nodes of the semantic network contain associated information about data elements such as concept definitions and formats (detailed further in section 3.1.1). Secondly, the network structure allows the representation of conceptual relationships between data elements that may otherwise be hidden. Finally, the semantic network itself provides a form of "context" for each data element. This context, formed by neighboring nodes and the relationships between them, provides a much richer basis of data interpretation and supports the concept matching algorithms used to find semantic equivalencies.
2) *Represent the structure and granularity of native databases.* Many databases have an inherent structure that reflects the logical organization of data and the manner in which it is used. The data itself may be represented at various levels of granularity, which is also a reflection of the local information environment. The semantic networks can capture this structure and granularity, which can make transmission and interpretation of data more efficient. [27]
3) *Provide support for automated exchange of data between databases.* One of the main goals of MEDIATE is to automate the process of data exchange as much as possible. The concept matching algorithms enable the discovery of semantically equivalent concepts between databases in a dynamic fashion, without a pre-negotiated static list of concepts and meanings. This means that any two databases that utilize MEDIATE can exchange data without the need to establish a common data model through previous human intervention.
4) *Facilitate retrieval of useful information in the absence of exact data correlation between databases.* If an attempt to retrieve a data element fails because the target database does not contain the element, it is sometimes useful to retrieve more "generalized" data, or other data elements that are somehow associated with the desired data. [28-30] The structure of the semantic network allows exploration of these alternative data elements, although the actual utility of the alternatives is a judgment left to the human user.

1.4 MEDIATE Overview

To achieve the functional goals delineated in the previous section, MEDIATE offers two tools to facilitate data exchange: a data representation system utilizing semantic networks, and algorithms to match semantic concepts between networks. Additional functionality is layered upon this representation and processing framework to capture all the elements required for data exchange. These elements include an interface to create and modify the data representation, a method to link the representation with native databases, a process for matching information between databases, and a method for retrieving and displaying the desired medical data.

MEDIATE attempts to capture some of the richness in medical information by explicitly representing some of the conceptual relationships that exist within a medical record system. These conceptual relationships form the links of a semantic network representation, and the data elements themselves form the network nodes. Several of the defined relationships are hierarchical in nature. This permits the representation of complex medical concepts as higher-level nodes with sub-nodes that are lower in the hierarchy. For example, the "composed-of/component-of" relationship can be used to state that a "complete blood count" node is composed of "white blood cell count", "hemoglobin level", "hematocrit", and "platelet count" nodes.

This semantic network representation provides an abstraction layer that is the key element to the data exchange process. Any system that implements the MEDIATE interface acquires the capability to exchange data with other systems that implement this abstraction layer. The MEDIATE system as a whole acts as a kind of "interpreter" for native database systems, identifying the semantically equivalent concepts between databases.

Functionally, there are three major components in the system: representation construction, concept matching, and query processing.

The representation constructor enables users to build semantic network representations of the medical record system using system-defined conceptual relationships. This representation acts as a model of the information database, and is stored with the medical record system. The original

record system and its MEDIATE representation are packaged with an associated query processor, thus forming an information source that can process queries from any requesting MEDIATE system.

The concept matching process utilizes the characteristics of the semantic network representations to match medical concepts between any two databases. Both networks are matched in an iterative process that produces a table of semantic equivalencies between databases. These equivalencies are then used in the data query process.

To initiate a query, the requesting database system utilizes the MEDIATE interface to find the semantic equivalents of the data elements that are to be retrieved. The request for these semantically equivalent data elements is then sent to the target MEDIATE system, which controls the actual retrieval of information from the native database. For example, if a user at Hospital A wishes to retrieve "Thyroid Function Tests" from Hospital B, the query processor would identify the equivalent concept "Endocrine Panel, Thyroid" from the semantic equivalency table and request this information from Hospital B. The query processor for Hospital B then cooperates with the native database to retrieve the desired information and transmit it back to Hospital A.

The system supports two methods of retrieving data from remote databases. The first method retrieves the matching nodes from the target database. For example, if "nodeA" in Hospital A is matched with "node1" in Hospital B, then when Hospital A's system makes a data request for "nodeA", Hospital B's database will return the data elements for "node1". The second method retrieves the matching leaf sub-nodes from the target database. Using the same example, if "nodeA" has leaf sub-nodes "nodeB", "nodeC", "nodeD", then a data request for "nodeA" will return nodes in Hospital B's database which match "nodeB", nodeC", and "nodeD" (i.e. not "node1"). The two match types are illustrated in Figure 1. For the remainder of this report, the former retrieval method will be called a "concept match", whereas the latter retrieval method is a "leaf match".

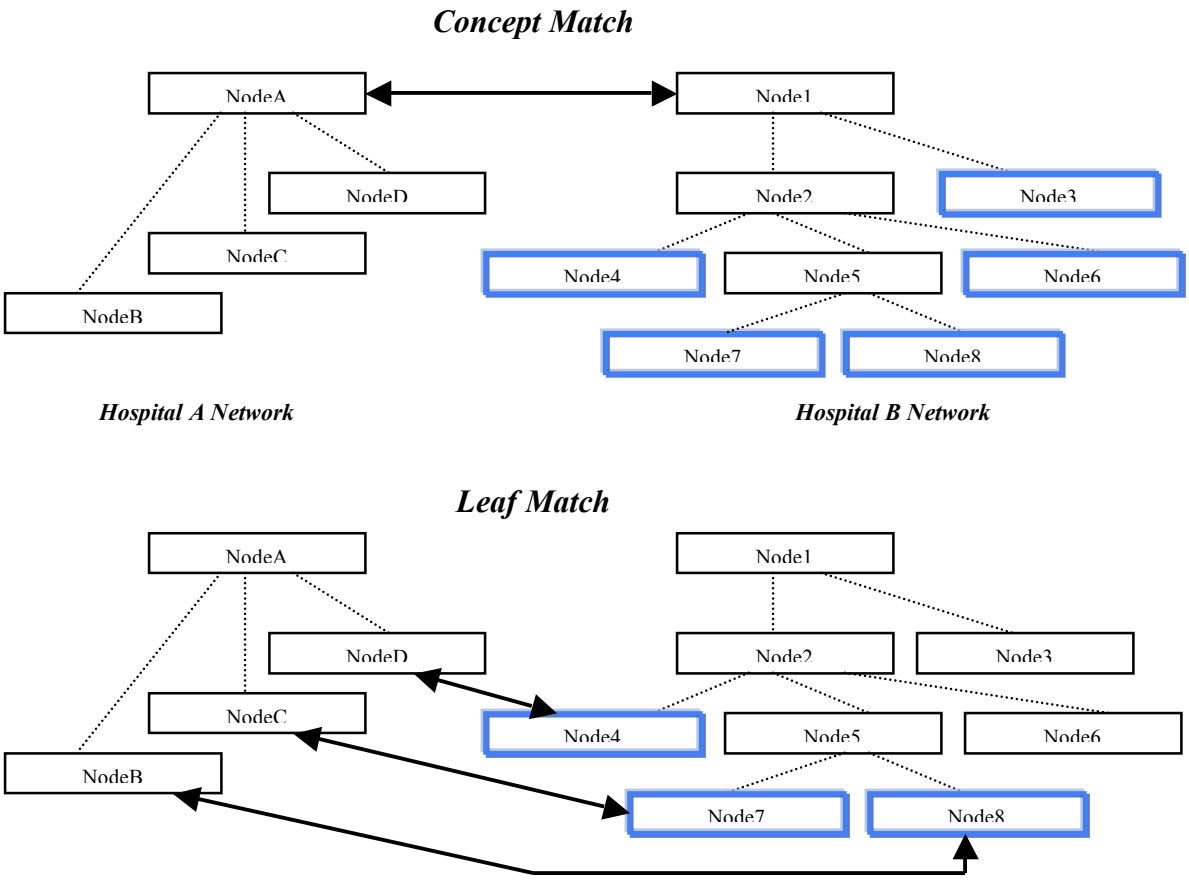


Figure 1. Concept vs. Leaf Match. In the concept match shown at the top of the figure, NodeA has been matched to Node1 (as denoted by double arrows). Subsequently, a data query will return five component nodes for Node1, namely Node2, Node4, Node6, Node7, and Node8 (bold outlined nodes). In the leaf match at the bottom of the figure, the three leaf nodes for NodeA have matched to nodes Node4, Node7, and Node8, which would be returned as the results of a data query.

Although the semantic network representation provides the data abstraction layer to support information exchange, the complementary process of concept matching provides the computational functionality that actually powers MEDIATE. Together, these components provide the foundation for the process of data exchange between heterogeneous medical databases.

1.5 System Benefits

The characteristics of MEDIATE offer many benefits in terms of scalability, robustness, and functional operation.

To exchange information with other MEDIATE systems, all that is required of a new database is the construction of a semantic network representation. This work is linear in the size of the native database if we make the reasonable assumption that there is a limit to the connectedness of the network, i.e. the maximum number of connections for any given node is a constant. More importantly, this work only needs to be performed for the new database, and no additional overhead or work is required to change existing MEDIATE-enabled databases to accommodate the new database. The work needed to enable the integration of N databases is thus order (N), or linear in the number of databases to connect. These qualities make MEDIATE a highly scalable system.

The dynamic nature of data exchange in this system confers desirable traits of stability and robustness. Since concept matching occurs at the time of data exchange, each database is isolated from the effects of changing or modifying other databases. (The ultimate case is the addition or deletion of a database to the data exchange group). Thus, MEDIATE provides an avenue for the underlying databases to evolve over time yet continue to exchange data with other MEDIATE-enabled systems.

In cases where a query request does not find the desired data in the target system, MEDIATE fails in a graceful manner by offering "generalized" concept matches that may still prove useful. Alternatively, the user may choose to execute a leaf match query if the requested data is a higher-level concept with subcomponents.

Since MEDIATE functions as an abstraction layer between databases, it facilitates the efficient use of legacy database systems. No changes need to be made to the operation of a database or its schema to accommodate data exchange through MEDIATE. As an added benefit, the semantic network representation helps to preserve and communicate the semantics and granularity of data elements, and reflects the way they are used within the legacy system.

The semantic network database representation presents data in a manner that is intuitively comprehended by most people. This satisfies one the four requirements for data integration software proposed by Rector, namely, "understandability". [31] This requirement states that that

information can only be maintained if people can understand its structure, despite any formalization for software use.

In addition to understandability, the semantic network offers the user a method of searching for information that is more intuitive than direct inspection of a database. Especially with the advent of the Internet, user interfaces that navigate through information by following "links" have become a well-known paradigm.

1.6 Scope of Investigation

This investigation is a proof-of-concept for the MEDIATE system. Instead of a large empirical data gathering effort, these initial experiments are targeted at characterizing the obstacles and possible solutions (within MEDIATE's representation and inference framework) to the problem of data exchange.

The initial test bed for MEDIATE involves two real world medical laboratory databases. Semantic network representations of both databases are constructed, and concept matching is demonstrated. Testing the ideas of MEDIATE within this restricted domain allows a more focused investigation, with the goal of generalizing the findings to other portions of the electronic medical record.

1.7 Thesis Outline

The remainder of this report is organized as follows. Section 2 will review previous approaches to data integration and explore the significance of MEDIATE. Section 3 delineates the details of MEDIATE, including system components, processes, and functionality. Section 4 explains the experimental setup that utilizes medical laboratory test results from two different hospitals, and section 5 presents the results of these experiments. Section 6 presents the analysis and discussion of this entire investigation, and concluding remarks are presented in section 7.

2 BACKGROUND

Investigators have tried many different techniques to access information from heterogeneous information sources. Since there is extensive research in this area, this section is intended as a brief digest rather than an exhaustive review of all possible methodologies and issues. Selected examples of major approaches and systems are presented, and the significance of using MEDIATE is discussed in relation to this work.

2.1 Common Data Models

One method to address the problem with database heterogeneity is to specify a common data model which would ensure compatibility if it is utilized. [32-39] For example, the W3-EMRS system by Kohane et al. specifies a Common Medical Record (CMR) structure into which information from remote sites must be mapped. [40] The CMR, however, is an abbreviated collection of medical information, such as problem lists, medications, allergies, and visit notes. It is not a rich semantic model and does not capture many data elements and informational relationships. In addition, each time the CMR definition is changed, the manual process of mapping remote information into the CMR structure is repeated. Any approach that specifies a common model suffers from this problem; if the model changes, then the transformations that map the remote information into the model must also change.

As discussed previously in section 1.2, common data models also have problems dealing with semantic inconsistencies that are due to the influence of the local clinical environment. The assignment of new semantics to existing medical concepts entails changes to the common model or to the mapping transformation between the local databases and the common model.

An additional problem to achieving uniform medical information access by this method is the proliferation of medical data models, each of which addresses some issue that would make a computerized patient record more effective. The large number of data models and system architectures, along with the generally slow process of arriving at consensus standards, means there is little likelihood of solving system incompatibilities via this method.

Well-known examples of common data models in the medical domain include the Reference Information Model (RIM) and Clinical Document Architecture (CDA) efforts by the Health Level 7 (HL7) organization. [1, 2, 41-43]

The RIM started as an attempt to create an encompassing data model for healthcare, but has subsequently become a generically descriptive model in which to frame processes within the healthcare system. RIM has 6 high-level "stereotype" classes that are designed to subsume all the elements of healthcare. These classes are: Entity, Role, Role_relationship, Participation, Act, and Act_relationship. The underlying "vocabularies" which define how concepts are encoded within these classes are still in evolution. Although it is certainly possible to represent data within the RIM, semantic inconsistencies can still exist because the RIM does not explicitly specify the nature of all data elements.

At this point in time, the CDA has not been specified in enough detail to describe the specific contents of a clinical document. A generic document header description exists, and work continues on descriptions for the document content.

In Europe, the GALEN project represents a multi-year effort to create a rich information model, the GALEN Common Reference Model, which can be utilized in a variety of medical information settings. [26, 44-53] In the view of the system designers, this model represents a "clinical terminology" which supports multiple perspectives on medical information encoded using the model. One of the chief benefits of this model is that concept relationships and inferences about those relationships are explicitly supported, partly due to the formal characteristics offered by the GALEN Representation And Integration Language (GRAIL). [25, 48, 54, 55]

Like all central models, however, the GALEN system requires mapping of local concepts to the central model (although the GRAIL formalism could be utilized at a local level with subsequent linking to the central model). Thus, the problem of resolving semantic differences between heterogeneous systems remains.

Another variation of the common data model is the use of a central ontology that specifies the conceptualization of the knowledge domain. [50, 56, 57] Since research in this area often originates from the knowledge representation field, ontologies are frequently designed from the start to deal with semantic issues. Despite this advantage, central ontologies can present significant mapping problems. Since central ontologies are designed to be encompassing, the formal specifications of such systems are often complex, and may utilize dense logical inferences that are difficult to understand without in-depth study. This complicates the mapping of local database concepts to the central ontology.

2.2 Federated Database Systems

Information management of heterogeneous database systems has led to the development of federated database architectures. [58-65] In contrast to a centralized "composite database" of integrated data, a federated system attempts to support local database operational autonomy within a design that allows sharing of information among interconnected databases. The goal of a federated system is to present a common interface for queries and transactions which are ultimately executed by the local databases.

To create the common interface, the designers of a federated system must integrate or reconcile the database schemas of its component databases. This integration may require a multi-level architecture as shown in Figure 2. This figure reflects the amount of effort that may be required to support a common interface. Schemas at various levels of abstraction (e.g. local, component, export, etc.) need to be integrated despite diversity from many sources, including different user perspectives, differing granularity in the model constructs, and incompatible design specifications.

Systems that implement some features of a federated database architecture include: ADDS (Amoco Distributed Database System), DATAPLEX (General Motors Corporation), IMDAS (National Institutes of Standards and Technology, U. Florida), Ingres (Ingres Corporation), Mermaid (Data Integration, Inc.), and Multibase (Xerox Advanced information Technology). [66]

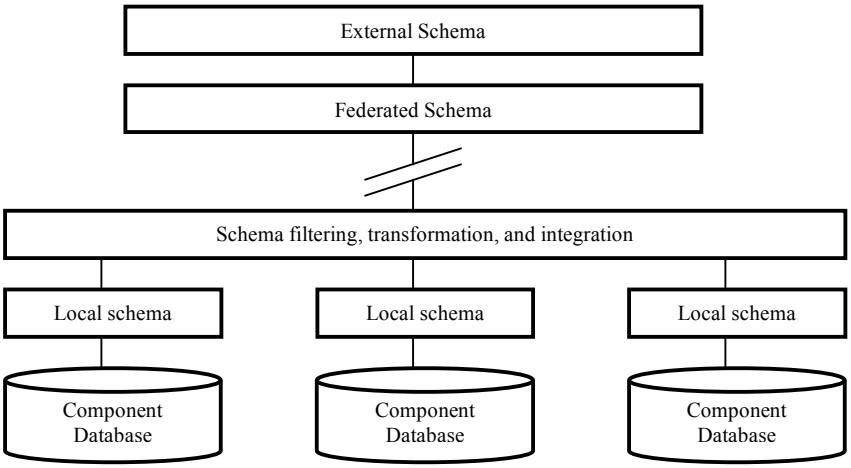


Figure 2. Federated Database Architecture. Local database schemas are processed, sometimes through multiple intermediate steps and transition schemas (indicated by the interrupted link below the federated schema), and eventually integrated into an overarching federated schema.

In all of these systems, manual transformations of database schema must be performed to match a "common" model at some level of the system architecture. Each time a new database is added, schemas must be integrated, often at multiple levels. If the new database offers unique information that must be available to all users, all levels of the federated architecture will be affected because of the schema dependencies. Thus, scalability becomes a significant issue if numerous databases might be added during future expansion of the system.

The SIMS project (Services and Information Management for decision Systems) is a variation that implements a semantic model of the problem domain to integrate various information sources. [67] The domain model represents all the information available in the sources within the system. SIMS uses the domain model in conjunction with models for each information source to execute a query.

The information source models can be created independently, which decreases the overhead of adding new sources. SIMS is also dependent, however, upon the comprehensiveness and integrity of the domain model, which must be incrementally enlarged as new sources are added. The authors of the SIMS system argue that since SIMS is designed to handle one domain at a time, this modeling effort will eventually reach closure.

Nevertheless, the central domain model has some of the characteristics of a central data model, and must be maintained to reflect changes in the sources. The need for continuing modifications to the model to capture new sources may affect scalability.

Although federated systems and variants such as SIMS also rely on a central framework, the approach differs slightly from fitting new information sources to a static central model. In a federated structure, the central framework expands and is adapted to utilize new information sources as they are added to the system. The main drawback to this approach is that additional effort is required to modify the central framework when new sources are added, and thus scalability remains an issue.

2.3 Mediators and Wrappers

The Context Interchange (COIN) project aims to make heterogeneous information sources more usable and accessible by establishing a structure for context management. [15, 18, 21, 68] Within the COIN system, data receivers as well as data sources have an associated "context" within which all information transfer is interpreted. Contexts are representations of the assumptions underlying the way that data is used within a system (e.g. all prices within a particular monetary database are in US dollars). In particular, the semantic meaning of data that is expected by the system (either for import or export) can be made explicit within a context.

COIN relies upon a "mediator" architecture, where the mediator acts to reconcile semantic conflicts between receivers and sources (Figure 3). By creating a common context mechanism for each data receiver or source, the need for static schema integration is transformed to a process of dynamic context mediation at the time that data is requested and transferred. The semantics of the data are captured in a dispersed manner, improving scalability and stability under system evolution.

Other systems that implement mediators to access heterogeneous information sources include Cobase and TSIMMIS.

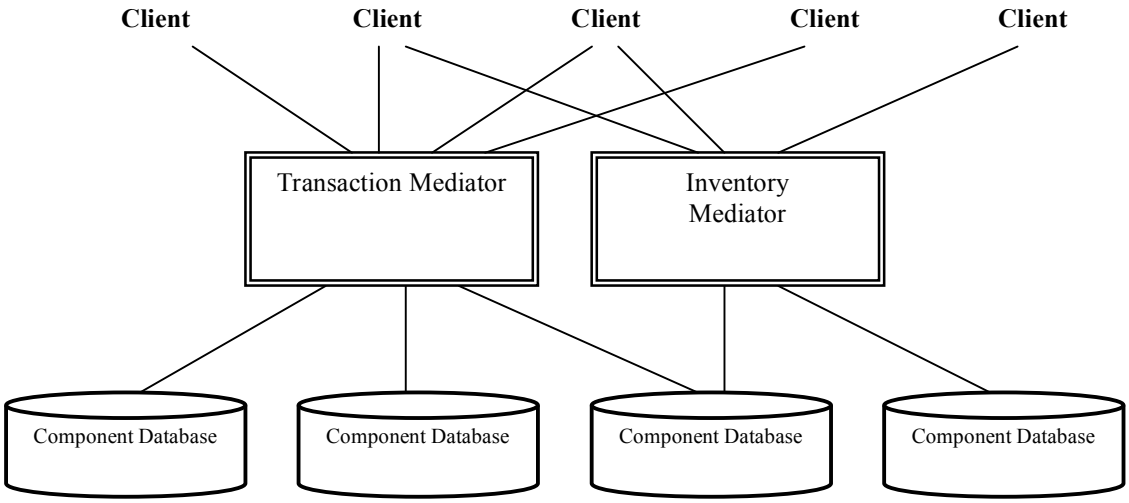


Figure 3. Example Mediator Architecture. Mediator modules centralize the processing of data queries and responses. Each mediator implements one or more functional processes, mapping transformations, inference engines, etc. depending upon the overall system design.

In Cobase, an integrated knowledge base provides representation of the data semantics in the form of "type abstraction hierarchies" (TAH). [28-30] This knowledge representation allows the system to "cooperatively" provide query answers by automatically generalizing or specializing the query when a specific answer does not exist. For example, if the query asks for a list of long-range runways in southwest Tunisia, Cobase may "relax" the query to obtain the list of all runways in Tunisia. In order to perform these cooperative functions, TAHs must be created for each information source and integrated into the overall system. An overall TAH directory stores the characteristics of all the TAHs.

To facilitate the integration of new information sources, Cobase relies upon a Mediator level that coordinates all information flow. The mediators are functional in nature (e.g. Relaxation mediator, Association mediator, TAH mediator, etc.) and may utilize other mediators to accomplish their function.

The mediator architecture is also a central component of The Stanford-IBM Manager of Multiple Information Sources (TSIMMIS). [69, 70] Other features in TSIMMIS include the use of "wrappers" to create uniform interfaces to information sources, and the use of an object model

called the Object-Exchange Model (OEM) to transfer information between components. TSIMMIS is similar to MEDIATE in some respects, because the goal of the OEM is to allow data representation to be "self-describing", or parsed without reference to an external schema, and the wrappers provide an abstraction layer that isolates the details of the underlying databases.

The Garlic system also features information source wrappers and is described as a "middleware" system between users and information sources. [71] Wrappers are used to model the contents of information sources as Garlic objects. This allows the Garlic system to invoke methods on the objects and retrieve their attributes. Similar to federated systems, Garlic maintains a global store of "metadata" that describes the unified schema of Garlic objects available from source systems.

2.4 **Information Translation**

The idea of "translating" information from one system to another system is appealing in its elegance and linguistic essence. In reality, however, the sheer variety of information systems makes direct translation unfeasible except on a limited basis. Instead, an "interlingua" or intermediate representation is often used. Information from one database is translated to the interlingua, and then translated from the interlingua to a form that can be utilized by a disparate database. [50, 51, 53, 72]

The Ontolingua system is representative of techniques that aim to increase the efficiency of sharing knowledge bases. [73] Using Ontolingua, the user can create "portable" ontologies of knowledge that can be translated into other knowledge representation systems. This method of knowledge sharing presumes that a domain representation with a high level of ontological commitment can be translated between different systems. The complexity of this task makes the utility of this system an open question in anything other than a research environment.

2.5 Other Information Encoding Systems

2.5.1 HL7 and XML

At a lower level of information encoding, HL7 provides a standard communications protocol for medical information messages. In its current form, however, HL7 is under-specified and does not provide the semantics to describe conceptual relationships within a medical record.

An ongoing effort to encode HL7 messages as Extensible Markup Language (XML) documents attempts to leverage XML's descriptive abilities to create a better representation of medical data. Some medical concept relationships are captured intrinsically through "containment" between XML data tags. Like all efforts to standardize on XML messaging, however, the HL7 endeavor still depends upon the creation of a central data model to use when interpreting the meaning of the XML field tags (i.e. the XML document type definition, or XML schema).

2.5.2 LOINC

The Logical Observation Identifiers Names and Codes (LOINC) system is a specific effort to encode laboratory test results in a standard structure that can be used to represent and communicate the contents of any laboratory database. A "fully specified" six-part name for the laboratory test forms the basis for this standard, and associated LOINC codes are assigned to each fully specified name. The six parameters for a fully specified name are: 1) analyte, 2) property of measurement (e.g. mass or concentration), 3) time aspect (e.g. point measurement or collection over time), 4) type of sample (e.g. urine, serum), 5) scale of measurement (e.g. qualitative vs. quantitative), and 6) method of measurement. The overall goal of LOINC is to encode all existing laboratory tests using fully specified names and associated code numbers.

LOINC shares the advantages and drawbacks of all common data models (as discussed in previous sections). Although LOINC has enjoyed wider implementation than many efforts, it still has problems that can impede data exchange. In its current form, there is no support for test panels since the fully specified names can only encode atomic laboratory tests. LOINC lacks the general structure to support multiple types of conceptual relationship between lab tests. Additionally, there is no mechanism for automatically mapping test codes between systems.

Consequently, the choice of LOINC codes for local data is non-trivial, and ambiguity in the choices can lead to failure of test matching as shown by Baorto (i.e. only exact matches between LOINC codes can be identified). [74]

2.5.3 **UMLS and other Clinical Terminologies**

On a terminology level, the Unified Medical Language System has collected many independent medical vocabularies under the umbrella of the Metathesaurus. The medical concepts catalogued through the Metathesaurus form a fairly comprehensive subset of concepts that are in current clinical use. Although the Metathesaurus is not intended to be a common data model per se, the collection of medical concepts from many sources allows it to function as a grounding point for mapping between vocabularies. MEDIATE utilizes the Metathesaurus for this very purpose (discussed in section 3.1.1.1).

The Metathesaurus, however, was not designed to be a data representation system, and therefore is not sufficient by itself to be used as a vehicle for data exchange. Similar to LOINC, there is no support for aggregating concepts, and little support for representing relationships between concepts (although there is some support for synonymy). Again, there is no mechanism for automatic mapping of concepts between information systems.

Similar problems exist when attempting to use other clinical terminologies as data representation systems. For example, the SNOMED and Read Codes nomenclatures are widely used, but neither these systems nor the UMLS Metathesaurus were found to be completely adequate for encoding clinical concepts (although SNOMED does support composition of concepts). [39]

The UMLS does provide concept relationships in another of its components, the Semantic Network. This system contains (as its name suggests) a semantic network of types and relationships. Furthermore, the goal of this system is to provide a broad framework for representing medical information rather than to provide an actual data model. The “semantic types” (network nodes) are broad categories such as “nucleotide sequence”, “sign or symptom”, and “clinical attribute”. Examples of the semantic relationships (network links) include “isa”, “surrounds”, “branch-of”, and “complicates”.

In total, the Semantic Network is a fairly rich representation framework and in some ways encompasses the nature of the semantic network representations within MEDIATE. The two systems differ in that the current relationships within MEDIATE are not fully supported by UMLS, and the UMLS Semantic Network only supports limited computations that do not extend to concept matching.

2.5.4 **KIF, KL-ONE, NIKL, and other Languages**

The Knowledge Interchange Format (KIF) is the standard language in which ontologies are defined within Ontolingua. [75] As a general language that supports first order predicate calculus, KIF could be used to fully specify the semantics and conceptual relationships within a medical record. The drawback of using KIF as a medical representation language is the amount of work that needs to be done to describe each system. MEDIATE aims to be a simpler system that provides constructs for common medical concepts and relationships, making it easier to describe a medical record. MEDIATE itself could be encoded in KIF or any other language general enough to express semantic relationships and operations on those relationships.

Other knowledge representation languages, such as KL-ONE, NIKL, and KOLA, have been studied in terms of their capability to encode general medical knowledge. [76-78] Although these languages have known deficiencies for representing general medical knowledge, the scope of their capabilities is much greater then the representation scheme for MEDIATE. Unlike the general knowledge representation languages, MEDIATE has a restricted and relatively simple structure with the goal of representing database concepts rather than general medical knowledge. This limited goal provides advantages in terms of understandability and efficiency. As with KIF, these other knowledge representation languages form a superset of the representation system used in MEDIATE.

One of the advantages of the restricted representation implemented within MEDIATE is that all the implemented inferences are decidable and non-exponential. This contrasts with some of the reasoning mechanisms of more general representation systems, in which certain problems may be undecidable.

2.6 Resolving Semantic Ambiguity

The vast majority of investigators in database integration advocate some form of central model to address the issue of semantic ambiguity, although the form of the central model ranges from data models, to schemas, to terminologies, to ontologies, to representation languages. Rossi Mori performed a survey of these approaches. [79] Approaches that do not utilize a central data model, however, do exist.

2.6.1 Extensional Definitions

Zollo and Huff have demonstrated a system where derived data can be used to characterize a laboratory test concept. [80] These "extensional definitions" of a concept are extracted from a representative data set for the pertinent concept, and may include parameters such as the mean, standard deviation, and units of measure for the concept. In essence, the extensional definitions provide additional semantic fields by which to identify the concept. Concept matching proceeds through matching of these extensional definitions.

Like many of the other systems, this approach lacks the ability to represent the relationship between different concepts. Consequently, it is not apparent how aggregate concepts are amenable to extensional definitions. In addition, semantic ambiguity is more of a problem when similar concepts have similar measurements (e.g. various forms of serum glucose measurements).

Interestingly, these investigators also implement a very crude context measure by including a "co-occurrences" field as one of the extensional definitions. The co-occurrences field list the 14 tests most frequently ordered in conjunction with the pertinent concept.

2.6.2 Taxonomic Reasoning and Graph-based Semantic Inferences

In a formal taxonomy of concepts, "classification" of a concept to determine its place in the taxonomy is a fundamental reasoning task. Bergamaschi argues that the taxonomic inference is a powerful technique to support conceptual schema design, recognize data instances, and validate queries. [19] Although she does not propose automated concept matching in her work, it

requires minimal extension of her thoughts to arrive at potential mechanisms to accomplish this task.

Many similarities exist between MEDIATE and the graph-based system proposed by Palopoli named DIPE (database interscheme property extractor). [23, 24] In DIPE, concepts from different database schemes are compared automatically to produce four output "dictionaries": the Synonymy, Homonymy, Type Conflict, and Object Cluster Similarity dictionaries. The synonymy dictionary is analogous to MEDIATE's concept matching, and the dictionaries are derived from a form of context comparison that is similar in philosophy to MEDIATE's concept comparisons.

Unlike DIPE, however, MEDIATE does not require initial human judgment and assertion of synonymy and "inclusion" (subclass) properties between schemes to start the inference process. DIPE also utilizes natural language processing to facilitate the synonymy/homonymy inference process, which may work with well-formed words and phrases but is unlikely to perform well with abbreviated and arcane medical terminology.

Although the overall approach to inter-scheme concept comparison is similar, DIPE is less automated and is therefore more sensitive to choices made during manual input. In addition, the definition of context used in DIPE is purely structural and relies upon delineation of relation attributes and keys. Database schemas that are constructed along functional lines (e.g. optimizing for the most frequently retrieved and updated data) may thus detrimentally impair the synonymy inferences. In contrast, MEDIATE utilizes the conceptual context as denoted by neighboring nodes in the semantic network representation, which is less sensitive to structural choices in the database design.

An example of how differences in context definition affect concept matching is that DIPE would seem likely to infer that all laboratory test results are synonymous, since they all share the same relational attributes (in the databases tested in this investigation).

2.7 Significance of MEDIATE

In contrast centralized data models, MEDIATE provides a uniform representation and processing model that allows information exchange without the need for Procrustean fitting to a static model. The ability to describe and quantify the amount of information transmitted via MEDIATE also differs from the unknown amount of information that is lost when fitting data to a centralized model. The fragility of common data models in the face of modifications and semantic change is avoided by the dynamic processing that occurs when MEDIATE executes its concept matching.

Compared to federated databases systems, MEDIATE does not enforce a central schema framework, which means that no additional overhead is needed to add each new information source. Scaling to virtually any number of sources thus has a linear amount of related work that involves creating the semantic network representations of the native databases.

The translation approach used by the interlingua systems approximates the goals of MEDIATE. In particular, the semantic network representation used in MEDIATE can be construed as an interlingua to which all native databases must be mapped. Unlike many of the systems, however, MEDIATE requires minimal ontological commitment because the representation system only requires a "fuzzy" form of mapping atomic data elements to medical terminology.

"Mediator" and "wrapper" systems are the most architecturally similar to MEDIATE. The MEDIATE semantic network that is associated with each information source is an implementation of a wrapper, although the specific functions differ from TSIMMIS and Garlic wrappers. MEDIATE classes are, however, structurally similar to TSIMMIS object-exchange models.

MEDIATE differs from these other systems in that the semantic network "wrapper" is designed not as a common database interface that abstracts away details, but as a way to actually reflect the structure and complexity of the underlying databases. Also, the functional process of identifying semantically equivalent data elements is not supported by the reviewed systems.

Similar to COIN and DIPE, one of the goals of MEDIATE is to explicitly represent and use data context to facilitate information exchange. The context implementations differ greatly between the systems, and ultimately, only empirical testing can provide evidence of practical efficacy.

Unlike the use of extensional definitions to resolve semantic ambiguity, MEDIATE easily supports the representation of aggregate concepts, and also provides a representation that clearly delineates the differences between similar concepts.

Current efforts to optimize information exchange in the healthcare field have provided many beneficial standards that aid communication of medical information. The HL7 communication protocol is widely used and implemented, and the UMLS Metathesaurus is utilized within MEDIATE. Most of the data representation efforts in this area, however, are attempts to construct common data models (e.g. the RIM, CDA, XML document standards, and the Galen Common Reference Model), and thus suffer the drawbacks of all such models. The redirection of the RIM effort is testimony to the difficulty inherent in this approach.

LOINC and the UMLS Metathesaurus offer different approaches to standardized vocabularies. Although common terminology is required at some level in order to define semantic equivalence, these systems lack the flexibility and power to represent complex aggregate medical concepts, and so cannot easily address problems with semantic ambiguity in such concepts. MEDIATE exploits the benefits of a standardized vocabulary, but also provides a richer representation scheme and a computational method to automated the identification of equivalent concepts.

The UMLS Semantic Network uses the same representation formalism for medical information, but the details of the system are not designed for facile data exchange. MEDIATE uses a different set of semantic relationships, and also employs a different level of computational power to achieve automated matching of concepts between database systems.

General knowledge representation systems or languages such as KIF and KL-ONE can be viewed as a superset of the representation scheme and functionality offered by MEDIATE. However, having a machine shop available to build any tool you desire is not the same as having

a specific tool on hand to perform a specific task well. MEDIATE implements a specific type of data representation and performs a specific set of computations that are targeted towards the goal of data exchange. It is not merely a reduction of a general representation system, but instead embodies a set of choices designed to meet specified goals.

In summary, MEDIATE provides the following contributions in its approach to integrating disparate sources of medical data.

1) It provides a way to represent and communicate the semantic context of database elements, and ameliorates the problem of semantic ambiguity.
2) The database representation reflects the way an information source is structured and organized, which allows an assessment of the granularity of transmitted information.
3) The task of identifying semantically equivalent data elements is automated.
4) The work needed to add new databases for data exchange is order(N), with no additional overhead or need to modify existing databases in the exchange group.
5) By avoiding central data models, it provides better scalability and protects the functionality of the system against evolving data element semantics.
6) The dynamic process of concept matching at the time of data exchange allows the system to be robust with respect to modifications in the databases, and even with respect to the addition of new databases.

Through the combined use of semantic network representations and concept matching algorithms, MEDIATE achieves the goal of data exchange with desirable characteristics that differentiate it from other systems of data integration.

3 MEDIATE SYSTEM DESIGN

The overall architecture of the system is illustrated in Figure 4. As described previously, there are three main components to the system. The constructor enables a user to build a semantic network representation of the native database. The concept matcher takes two semantic networks as input, and produces a table of concept equivalencies between the networks. Finally, the query processor uses the semantic equivalencies and network representations to retrieve data from the native databases. These elements are described further in the following sections.

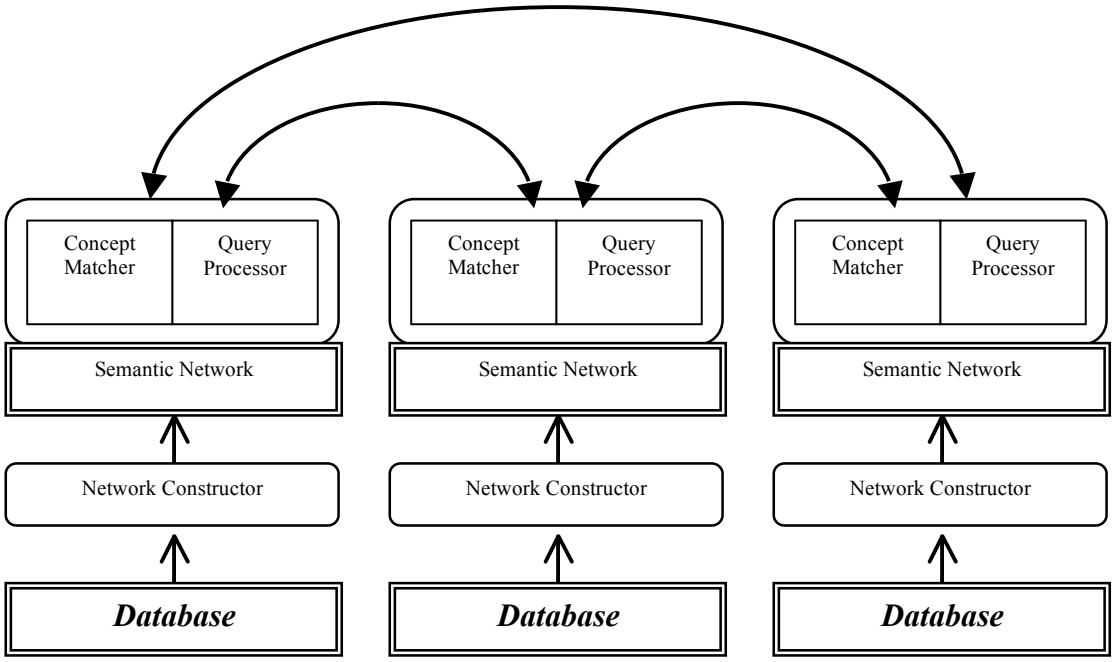


Figure 4. MEDIATE Architecture. The MEDIATE network construction routines are used to create the semantic network representation for each native database. The semantic network, along with an associated concept matcher and query processor, forms the interface to route communication with other databases. Concept matching occurs every time data is communicated if the semantic network representations (of the participating databases) have been modified since the last data exchange.

3.1 Semantic Network Components

Like any semantic network, the MEDIATE representation of native databases is composed of nodes and links. The nodes represent medical concepts, and the links represent defined relationships between those concepts.

The overall goal of the semantic network representation is to capture a conceptual view of a medical database, which includes "higher-level" concepts as well as the atomic data elements. In a medical laboratory database, for example, this would include concepts which denote the normal organization of laboratory test types, e.g. hematology, microbiology, pathology, chemistry, etc. These higher-level concepts may or may not be encoded as data elements within the native database. Along with the information represented by the relationship links, the "meta-data" contained by these higher-level concepts and the network topology enable MEDIATE to perform the computations to determine semantic equivalence between concepts.

3.1.1 **Semantic Network Nodes**

The network node represents a single medical concept, and contains all the information for that concept including the relationships to other concepts. The node contains other data structures that specify concept identifying information, relationship links, data formats, and database hooks.

3.1.1.1 ***Node Identification***

Identifying information is necessary to uniquely classify a node. Identification of a node is unique to the database system that the node represents: it is not intended to be a universal identifier that carries across database systems. The identification fields include the following:

1) *Name:* a human readable label that corresponds to the medical concept.
2) *Unique ID:* a unique identifier (perhaps randomly generated) for the node that will never be reused.
3) *UMLS link:* a link to a standardized vocabulary to associate the node with known terms.
4) *Definition:* a plain-text "definition" of the concept embodied within the node. The definition is another method for directly representing semantic information about the medical concept of interest.

The UMLS link is used to associate the MEDIATE medical concept with concepts contained in the Metathesaurus. Although this appears to force MEDIATE to conform to a common data model, the UMLS link itself is not a rigid association between the node and a Metathesaurus concept. Instead, the link is represented by a list of Metathesaurus concepts with semantics that are compatible with the node. This is an important distinction because the semantics of the Metathesaurus concepts are often open to interpretation. Therefore, instead of forcing a single

semantic association, the UMLS link represents a "fuzzy" set of possible associations. This increases the flexibility of the system compared to rigidly conforming to a central data model.

3.1.1.2 *Format*

Format information is divided into two components, the type of information being transmitted, and the encoding of the information. The type describes the semantic type of the information being represented (e.g. number, text, image, sound, aggregate concept, etc). The encoding specifies how the information is actually stored. The encoding for the information may differ from the type. For example, a platelet count should be interpreted semantically as type "number", but the value may be encoded as a text string in the source medical record system. Also, a variety of encodings may be available for the same type, e.g. type: "image", encoding: JPEG vs. PICT vs. PDF, etc. The explicit representation of encoding information allows the usage of standardized routines to display the data or allow conversion between encodings.

This form of format representation contains both semantic (type) and syntactic (encoding) information about the data concept.

3.1.1.3 *Database Link*

In order to retrieve data from the native database, there must be a link between nodes and atomic data elements. This database link represents a call to the native database system to retrieve the actual data item of interest. Currently, the data structure and functionality of the database link has been optimized for relational databases, which are the most prevalent type of databases in use. (Linking nodes to relational databases is further discussed in section 3.4).

The database link currently contains the following components:

1) *Table:* the database table that contains the data element of interest.
2) *Column:* the table column that contains the data element of interest.
3) *Next link:* the next database link to use when executing some forms of multi-part queries.
4) *Previous link:* the previous link in some forms of multi-part queries.
5) *Query type:* the method used to retrieve information from the database. The query types currently reflect usage within a relational database, and include:
    a. *Column value:* retrieve data by specifying the name of a column.

b. *Column domain:* retrieve data by specifying a value within the column domain (i.e. the values of data elements within the column).
c. *Column pointer:* the data value within the column is a pointer to another table or column.

6) *Aggregate:* the data element is actually composed of lower level data elements. Therefore, the database links for the lower level data elements are to be used, possibly in a recursive fashion, to retrieve the information for the higher-level data element.

7) *Attributes:* parameters associated with the node concept that must be retrieved whenever the concept data is retrieved, and that will be inherited by all subclasses (specialization relationship) of the node. For "laboratory results", attributes might include the result units, a time-stamp for when the result was reported, and an order accession number.

   It is difficult to assign a strict definition to an "attribute", since the core idea of a parameter that is always "related to" the main concept is not quantifiable. In a relational database, an attribute is most likely to be other columns within the same table. Thus the laboratory results table would contain columns for result units, time stamp, etc.

   The choice of attributes directly relates to the design choices that are made for inheritance in an object-oriented system. There are no strict criteria to follow when deciding on inheritable parameters for an object, but many such choices are relatively straightforward.

8) *Constraints:* a set of Boolean expressions that constrain the data values to retrieve.

Using these defined database links, MEDIATE directly generates SQL queries that are executed by the native database system. This function is part of the query processor, and needs to be customized for different types of databases.

The SQL statements generated by the query processor are generic in form in order to be compatible with the broadest range of relational databases. One corollary disadvantage is that

these statements are not optimized, and therefore may not produce the best performance in terms of retrieval speed.

3.1.1.4 *Relationships*

The data structure for relationships contains the information specifying how the node is related to other nodes. The relationships are directional, so each node directly specifies its relationship with the target of that relationship. For example, if "time stamp" is an attribute of "Lab Result", then "time stamp" contains the relationship "attribute-of" "Lab Result", and "Lab Result" contains the relationship "has-attribute" "time stamp". More information about relationships is contained in the following section.

3.1.2 Network Links

Links within the semantic network represent conceptual relationships between medical concepts. The network itself is defined to be a directed acyclic graph, in order to facilitate the function of the concept matching algorithms.

3.1.2.1 *Relationship semantics*

3.1.2.1.1 Identity: *same-as*

This relationship states that two medical concepts are synonymous. In particular, all the components of the node data structure are identical except for the name and Unique ID fields in the Identification data structure.

3.1.2.1.2 Specialization: *subclass-of, superclass-of*

This relationship follows the semantics of traditional object-oriented class specialization, where subclasses inherit attributes and functionality (or "methods") of their superclasses. Subclasses are restricted to modifications that preserve the attributes (i.e. may add more attributes) and retain the method call forms (i.e. may change the function of the method but preserve the call and parameter list, or may add a new method) of the superclass.

3.1.2.1.3 Composition: *component-of, composed-of*

The composition relationship states that the semantic content of the higher-level node (the "construct") is built from the semantic content of the lower-level nodes (the "components"). In

addition, all the components must be present in order for the construct to be a valid entity. The components are necessary and sufficient parts to define the higher-level node, and the addition or elimination of a component creates a different construct. For example, if a “bleeding screen” is composed-of the prothrombin time (PT), the partial thromboplastin time (PTT), and a fibrinogen level, then ordering the PT and PTT without the fibrinogen level does not constitute a “bleeding screen”.

This relationship is analogous to the “part-whole” relationship discussed in the linguistics and knowledge representation fields. [31]

3.1.2.1.4 Aggregation: *element-of, collection-of*

In contrast to composition, aggregation does not require all of the lower-level nodes (the “sub-elements”) to be present in order to define the higher-level node (the “aggregate”). The semantic content of the aggregate is defined by the content of the sub-elements, whatever those sub-elements might be. This relationship enables the representation of lists with variable size (e.g. a medication list) and aggregates of data that may have variable membership (e.g. the aggregate symptoms required for the diagnosis of Rheumatic fever).

3.1.2.1.5 Set relationships: *subset-of, superset-of*

This relationship follows the standard mathematical definition, with set elements defined by lower-level nodes.

3.1.2.1.6 Attribution: *attribute-of, has-attribute*

Attributes are lower level nodes that are associated with a higher-level node (the “foundation”) through the property of inheritance. Attributes are the characteristic bits of information that are inherited by subclasses of the foundation. As illustrated previously, a “Lab Result” may have attributes of “result units”, a “time stamp” for when the result was reported, and an “accession number”. These attributes are inherited by all subclasses of “Lab Result".

The attribution relationship must be included on an engineering basis in order to facilitate the proper retrieval of data with related properties (e.g. the “Lab Result” discussed above). In

particular, the structure of relational databases confers a practical definition in terms of the associated (single table) columns that are retrieved during a query.

Since the definition of an attribute is not fully specified, MEDIATE treats this relationship as orthogonal to the other relationships. Attribution is the only relationship included in the database link, but it is not included within any of the search algorithms used in the concept matching process.

3.1.2.2 *Relationship properties*

Properties of the relationship links are shown in Table 1.

	Commutative	Transitive	Hierarchy	Inheritance	Dependence	Overlap
Identity	Yes	Yes	No	No	No	Yes
Specialization	No	Yes	Yes	Yes	No	Yes
Composition	No	Yes	Yes	No	Yes	No
Aggregation	No	Yes	Yes	No	No	No
Set relations	No	Yes	Yes	No	No	Yes
Attribution	No	Yes	Yes	No	No	No

Table 1. Relationship properties. For a given relationship * (or its inverse), the properties have the following meaning. *Commutative*: a * b implies b * a. *Transitive*: a * b and b * c implies a * c. *Hierarchy*: a * b implies a is a "higher-level" class and b is a "lower level" class. Hierarchy has transitive closure. *Inheritance*: a * b implies b inherits attributes from a. *Dependence*: a * b implies the semantic meaning of a is dependent upon b. *Overlap*: a * b implies there are overlapping properties or elements between a and b.

3.2 Network Construction

Constructing the semantic network representation of a native database constitutes the primary work required to implement MEDIATE. This work is only performed for the local database, without regard to the nature or number of other databases with which information exchange will occur. Modifications to the semantic network are required only to reflect changes in the local database, and do not need to reflect changes in remote databases.

As a representation of the native database, MEDIATE provides functionality that correlates directly with the accuracy and completeness of the representation. Thus, time and energy spent during the representation construction phase will have a direct payoff in terms of later

functionality. Unfortunately, the corollary is also true, that inaccurate or incomplete representations may hide underlying information or actually mislead users about the contents of the legacy database.

3.2.1 User Interface

A graphical user interface was designed to facilitate the construction of the semantic network. A screen shot of main interface window is shown in Figure 5. The semantic network itself is shown graphically in a sub-window that allows navigation through a point-and-click interface. This allows users to easily visualize the node nodes and relationship links as they are created or modified.

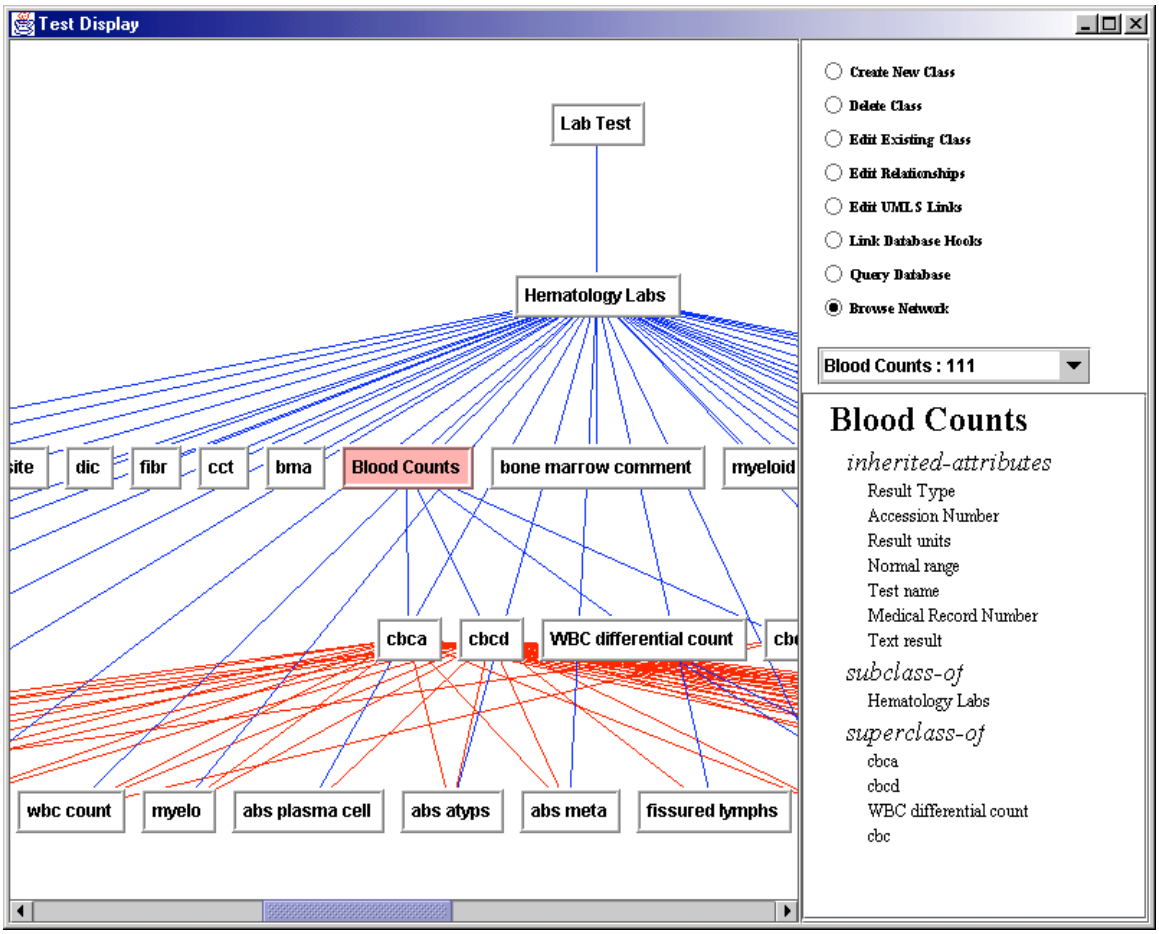


Figure 5. Semantic Network user interface main window. The "Browse Network" view has been selected, and relationships for the highlighted class "Blood Counts" are displayed in the sub-window on the right. Attributes are not displayed in the network view sub-window on the left.

All the functionality required to construct the semantic network is supported within the interface, including node creation, modification, and deletion. Other functions of MEDIATE are also accessed through this interface, and those aspects will be discussed in following sections.

One way of facilitating the construction of the semantic network is to use external programs to read information from the native database and convert that information to MEDIATE system nodes and relationships. This approach can be used to initially populate the network, with further refinement performed by utilizing the graphical interface. (This was done to help construct one of the semantic networks used in the experimental phase of this investigation). The design and finalization of many of the relationship links, however, must be performed within the MEDIATE interface since the relationship semantics are seldom (if ever) directly extractable from the native databases.

3.2.2 **Node Identification**

Most data elements within a native database can be represented by a node that uses the data element "name" for the node name. When the data element names are cryptic, an expanded node name using basic medical terminology is desirable but not always possible if the original data naming convention is too obscure to interpret. The node unique ID can be assigned in any manner that ensures non-duplication of the field within the semantic network. (The MEDIATE interface does not allow entry of duplicated unique ID fields).

Implementing a unique ID field allows the reuse of node names if the underlying data element changes but the semantics of the concept remain the same.

3.2.3 **UMLS concept assignment**

One of the most important tasks in constructing the semantic network is linking a node with UMLS Metathesaurus concepts. The "standardized" vocabulary embodied in the Metathesaurus provides fundamental support for concept matching. The user interface window for accomplishing this task is shown in Figure 6.

The UMLS link is constructed by creating a list of Metathesaurus concepts that are semantically equivalent to the node. Ideally, semantic equivalence should imply semantic identity, but this is not possible for several reasons.

Even in a standardized vocabulary, semantic ambiguity exists. For example, "sodium level" and "sodium in sample" are listed as two non-synonymous concepts in the UMLS. Yet any medical professional would most likely interpret the two concepts to mean the same thing.

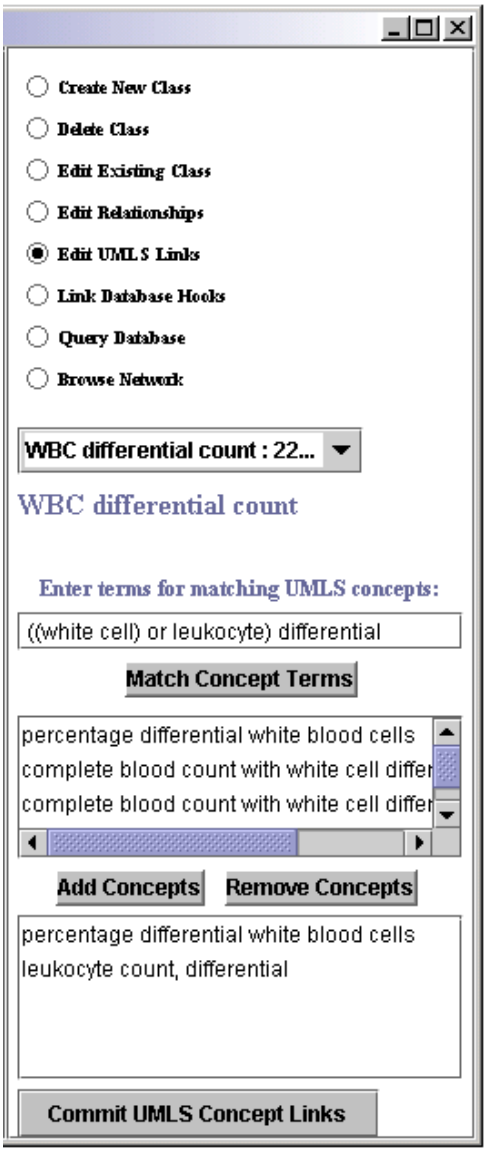


Figure 6. UMLS Link sub-window. This window is used to link a list of UMLS concepts to the selected node.

The Metathesaurus also lacks the semantic richness to describe the type of relationships that are integral to MEDIATE. Thus, "serum sodium level" is a more specialized concept than "sodium level", but this relationship is not possible to represent within the Metathesaurus.

As previously discussed, the semantics for a node in any given database are highly contingent upon the usage of that concept within the local clinical environment. Therefore, there is no guarantee that any Metathesaurus concept will be "identical" to a node.

To address these semantic obstacles, MEDIATE constructs the UMLS link by allowing the user to choose from a list of concepts. Since the individual users may differ in their judgment of "semantically equivalent" terms, the UMLS link is not a precise or rigorous parameter. Instead, it functions as a "possibility set" of semantic states that the node might attain.

To create the UMLS link, the user specifies a list of terms that are used in a matching algorithm to retrieve locally

stored Metathesaurus concepts. Several features are implemented within the matching algorithm to optimize the presentation of candidate Metathesaurus concepts.

One feature is a parser that allows the search terms to be entered as a boolean expression. Another feature is an automatic plural form generator that produces the plural forms of match terms using standard rules of English. When the match term “cell” is entered, the plural form “cells” is automatically generated, and when “fungus” is entered, “fungi” is automatically generated.

Finally, Metathesaurus concepts that contain the node match terms are assessed using a metric that takes into account the number of matched node terms as well as the position of those terms within the concept phrase. Concepts with the highest score are placed at the top of the candidate list so that the user is presented with the most likely matches first.

Once the user has chosen any number of equivalent Metathesaurus concepts (from zero to n), he or she confirms these concepts and they are placed in the node UMLS Link.

3.2.4 **Relationship assignments**

Once a node has been created, it can be linked to other existing nodes using the predefined relationships. These relationships are then displayed within the user interface as network links between the participating nodes. Relationships cannot be created between non-existent nodes.

3.2.5 **Network structure**

As previously stated, the semantic network representation is restricted to a directed acyclic graph topology for any given relationship link. In practice, the networks are more likely to resemble trees because of the hierarchical property of many of the relationship links. The terminal nodes, or “leaves” of these networks often correlate with atomic data elements within the native database.

The overall structure of the semantic network is not explicitly represented. Instead, each node describes its own local network using its relationship links, and the sum total of all the node relationships gives rise to the whole network. The basic granularity of the network representation

thus resides at the node level, which makes it more robust to change and easier to scale (add more nodes). Similarly, all the network traversal and matching algorithms operate at the node level and do not depend upon knowledge of the overall network topology.

3.3 Concept matching

The central functionality of MEDIATE resides in the algorithms that match concepts between semantic network representations. As discussed previously, the ability to reduce semantic ambiguity and discover semantic equivalencies forms the fundamental basis for integrating heterogeneous databases within this system. Given semantic network representations of two databases, this problem reduces to finding "matching" concepts between the semantic networks.

Matching cannot occur between more than two databases simultaneously because finding the semantic equivalent of multiple concepts simultaneously is not a well-defined problem within this system. MEDIATE "views" information exchange from the perspective of a single database, and data integration takes place with respect to that database. Multi-network matching is more akin to finding a "common" semantic model that satisfies all the networks, and this perspective is not supported by MEDIATE.

3.3.1 Matching Algorithms

3.3.1.1 *Overall matching process*

The general process of concept matching utilizes an algorithm that has three phases.

In the first phase, each of the two networks is enumerated on a node-by-node basis and matches are attempted using multiple algorithms (detailed in section 3.3.1.2). The majority of node matches will be found during this phase.

In the second phase, an iterative matching process is used for unmatched nodes from the first phase. Some of the algorithms depend upon matches between neighboring nodes in order to match the target node, and thus may fail during the first matching phase but succeed in subsequent iterations. The iterations in the second phase continue until the total number of matched nodes remains static.

Throughout the first two matching phases, all the identified concept matches are stored in a hash table for later referral. This improves the efficiency of the matching algorithms which rely on finding similarities between concept contexts, since multiple neighboring nodes may also need to be matched.

In the third phase, the remaining unmatched nodes are put through an iterative "generalize and match" process. During this process, the system generalizes a node by finding its superclass, using the subclass-of relationship links. If the subclass-of relationship does not exist for the pertinent node, the subset-of, component-of, and element-of hierarchical relationships are tested successively until a higher-level class is found. The higher-level class is then matched if possible. The generalization and match process is recursively iterated until the superclass is matched, or no superclass is found. The theory for this phase is derived from the query "relaxation" function provided by Cobase systems (discussed previously in section 2.3). This theory postulates that even if a semantic equivalent is not found, information of a generalized form may still prove useful.

3.3.1.2 *Specific matching algorithms*

There are currently six specific matching algorithms employed during the three phase matching process, and one algorithm which may be employed in a discretionary fashion after the automated concept matching process. A node is matched if at least one of the six basic algorithms returns a matching node from the remote network. If multiple matching nodes are returned, each node is displayed by the system with an associated "match quality" metric (discussed further in section 3.3.2). This quality metric can be a guide for users to choose the best match from the candidate matches, or it can be used to automate the choice of matches.

The matching algorithms can be categorized in the following manner:

1) *Terminological match*. This algorithm matches concepts using links to the UMLS Metathesaurus.
2) *Context match.* These algorithms execute matching by examining the context (network neighborhood) of the target node.

a) *Subcomponent context*. Use the context represented by subcomponents (leaves) of the target node.
b) *Nearest neighbors context.* Use the context represented by all the neighbors of the target node.
c) *Sibling context.* Use the context represented by sibling nodes.

3) *Leaf match.* Match as many of the subcomponents as possible.

The specific matching algorithms are described in the following sections, with some illustrative pseudo-code.

3.3.1.2.1 Terminological Match by UMLS link

This algorithm uses the UMLS links to find matching nodes. Nodes from the two semantic networks match if they have any common elements in their UMLS links. Due to the indeterminate content of the UMLS links, there is no guarantee that matches can be found, or that they will be unique.

In contrast to the other algorithms, the local "neighborhood" of a node is not considered in this algorithm. In situations where a node has sparse relationship links (e.g. in leaf nodes), this algorithm may be the main determinant of the matching outcome.

```
For each target-node in the local network
      target-UMLS-list <= UMLS list of target-node
      For each remote-node in the remote network
            remote-UMLS-list <= UMLS list of remote-node
            For each target-item in the target-UMLS-list
                  For each remote-item in the remote-UMLS-list
                        If (target-item equals remote-item) then
                              Add remote-node to matching-nodes
Return matching-nodes
```

3.3.1.2.2 Subcomponent context match: finding the "lowest common superclass"

To match a given "NodeA" in the local network, the algorithm starts by finding any leaf nodes that are in NodeA's sub-hierarchy. These leaf nodes are then matched to nodes in the remote network. Within the remote network, a search process is started from each of the matching nodes. The search proceeds in a breadth-first (BFS) fashion "up" the network hierarchy from each of the remote matching nodes. The "lowest common superclass" is the lowest node with the greatest number of search "hits" from the remote matching nodes.

```
For each leaf-node of the target-node
      Retrieve remote-matching-node from matching hash table
      While termination condition is false
            For each remote-matching-node in the remote network
                  Perform BFS up the remote network hierarchy
                  Mark each node traversed with a unique "hit" label
            Count hits for each node traversed
            If ((maximum hit count remains static) or
               (no more nodes to Search)) then
                  Terminate condition for While loop is true
Return remote node with maximum hit count
```

3.3.1.2.3 Subcomponent context match: variation on lowest common superclass

Specialization links contain hierarchical information about the semantic network. These links, however, are much less constraining than the other hierarchical relationships. To narrow the search space, this algorithm implements a variation of the lowest common superclass algorithm that excludes specialization links from any network traversal operation (e.g. while finding leaves or during BFS).

This algorithm and the previous algorithm are somewhat complementary. The previous algorithm uses the broadest search space available, which is useful when the semantic network is sparse. By narrowing the search space, this algorithm returns more accurate results when the network is denser.

3.3.1.2.4 Nearest neighbor context match: match by ripples

The intuition for this algorithm originates from the ripples that result when pebbles are cast into a calm body of water. As the ripples spread from each pebble's impact, they intersect in various patterns. The points of greatest ripple intersection are the "centroids" of interaction between the original pebble impacts.

In this algorithm, a BFS is executed within the local network to find the nodes closest to the target "NodeA". These neighboring nodes are then matched in the remote network. The remote matching nodes are analogous to the cast pebbles, and performing BFS from these nodes is analogous to creating ripples. The remote network node(s) with the greatest number of hits from the intersecting BFS pathways are returned as the overall match for NodeA.

```
Local-neighbors <= perform BFS for 1 link distance from target node
Remote-neighbors <= retrieve match for each Local-neighbor from
                    matching hash table
While termination condition is false
      For each Remote-neighbor
            Perform BFS in remote network
            Mark each node traversed with a unique "hit" label
      Count hits for each node traversed
      If ((maximum hit count remains static) or
          (no more nodes to Search)) then
            Terminate condition for While loop is true
Return remote node with maximum hit count
```

3.3.1.2.5 Nearest neighbors context variation

This process essentially duplicates the ripples algorithm, but the surrounding BFS nodes in the local network are also matched in the remote network, and these matched nodes are then excluded from the final result.

3.3.1.2.6 Sibling context match: neighbor exclusion

To perform a match using this algorithm, the parent node and "sibling" nodes are matched in the remote network, then excluded as candidate matches. For example, assume there exists parent NodeA and children NodeB, NodeC, and NodeD. When attempting to match NodeB, the parent

NodeA is found and matched in the remote network to find NodeARemote. The children of NodeARemote are then found. NodeC and NodeD are then matched in the remote network, and the matching NodeCRemote and NodeDRemote are excluded from consideration by eliminating them from the children of NodeARemote. The remaining children of NodeARemote are returned as candidate matches for NodeB.

3.3.1.2.7 Leaf match

After the three-phase general concept matching process is performed, the user can choose one more algorithm if the previous match results are unsatisfactory. For nodes that have subcomponents, the user may execute this algorithm to match the leaves of the sub-hierarchy instead of matching the target node itself. The purpose of this algorithm is utilitarian: it does not attempt to find the semantic equivalent of the target node, but instead tries to match all the data elements that make up the sub-hierarchy of the target node.

In some circumstances, this may be preferable to using the semantically equivalent match to retrieve information from a remote database. For example, if the sub-hierarchy for the target node in the local network is larger than the equivalent sub-hierarchy in the remote network, more information may be retrieved using this algorithm than by using the semantically equivalent match to the target node.

3.3.2 **Match Quality Metric**

Once the concept matching process is completed, a method to assess the quality of node matches can assist the user in evaluating the efficacy of the matching process. In particular, if a local node is matched to more than one node in the remote network, the quality metric can be used to judge the relative "fit" of the matches.

Several parameters are used within the quality metric to capture different aspects of the match. These parameters include:

1) *Overall quality.* A match between two nodes is called a "perfect" match if the all subcomponents of both nodes also match. Otherwise, the match is a "partial" match.

2) *Coverage.* A match has "full set coverage" with respect to the local target node if all the subcomponents of the local target node are matched and contained in the subcomponents of the remote node. Otherwise the match has "partial set coverage".
3) *Score.* The score is calculated by taking the number of matching subcomponents (intersection between the subcomponents) divided by the total number of unique subcomponents (union of the subcomponents), multiplied by 100. This produces a range from 0 to 100. Using the subcomponent context (nodes in the sub-hierarchies) is a more specific measure of concept similarity than using the more general context, which includes all neighboring nodes.

If more than one candidate matching node is found in the remote database, the system can calculate a "best match" based on the highest quality score. In the case where two or more candidate matches have the same quality score, the node with the smallest sub-hierarchy is returned as the most "specific" node (i.e. least generalized).

3.3.3 **Match Types**

Match types are differentiated by the method used to establish the match. The differentiation is necessary because different network traversal routines and variations of the quality metric algorithms are required for the different types. From the concept matching process described previously, the match types are:

1) *Direct match.* The match is made during the initial concept matching process.
2) *Generalized match.* The match is made during the "generalize and match" process because the node was previously unmatched.
3) *Leaf match.* The user manually directs the system to perform a leaf match.
4) *Validated match.* During review of the concept matches, the user manually confirms that a match is semantically equivalent and should be used for all future data integration purposes. A validated match is always preferentially used regardless of the quality metric.

3.3.4 **User Interface**

To assist the user in evaluating the semantic concept matches, a graphical user interface was designed to display the two networks along with user-selected node matches. This interface is illustrated in Figure 7.

The graphical interface displays the network environments within which the matches are made. The quality metric for each node match is also displayed. This allows the user to better judge the suitability of the automated matches and decide which matches to validate.

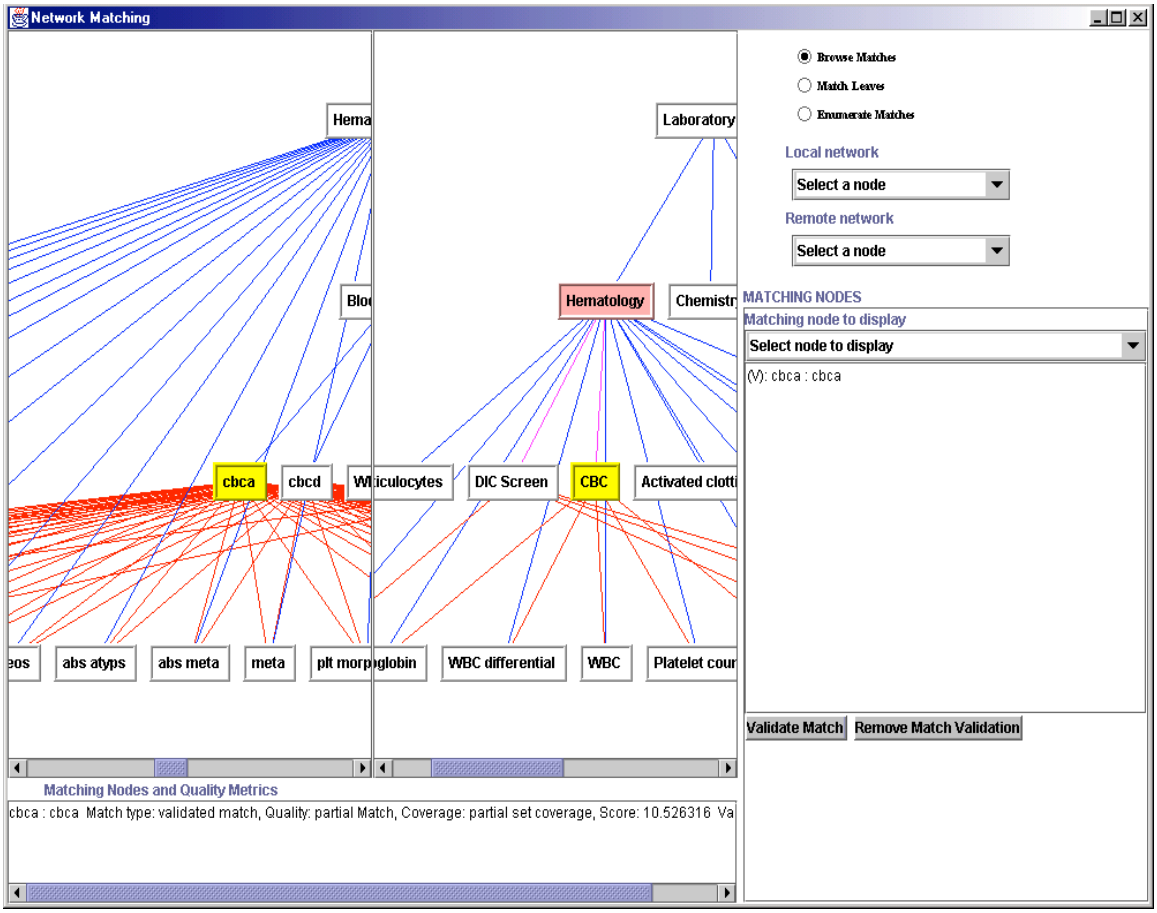


Figure 7. Concept matching review window. The left and middle panels display the semantic networks and allow the user to select node matches for review by clicking on the target node within either window. Below the network windows is a display for the quality metric of the current match. The right panel allows the user to choose various functions, including validation of matches.

3.4 **Database Linkage**

Currently, MEDIATE has a user interface which enables linkage between node nodes and database elements within a relational database. A sample window is shown in Figure 8.

Four different query types are currently recognized within the node database link. It is important to correctly delineate the query type in order to process the retrieved data elements. These types are:

1) *Column value.* The information content for the node is directly contained within the table column. For example, the node for "serum sodium" would have its primary link to the column "serum sodium" within the table "serum electrolyte values".

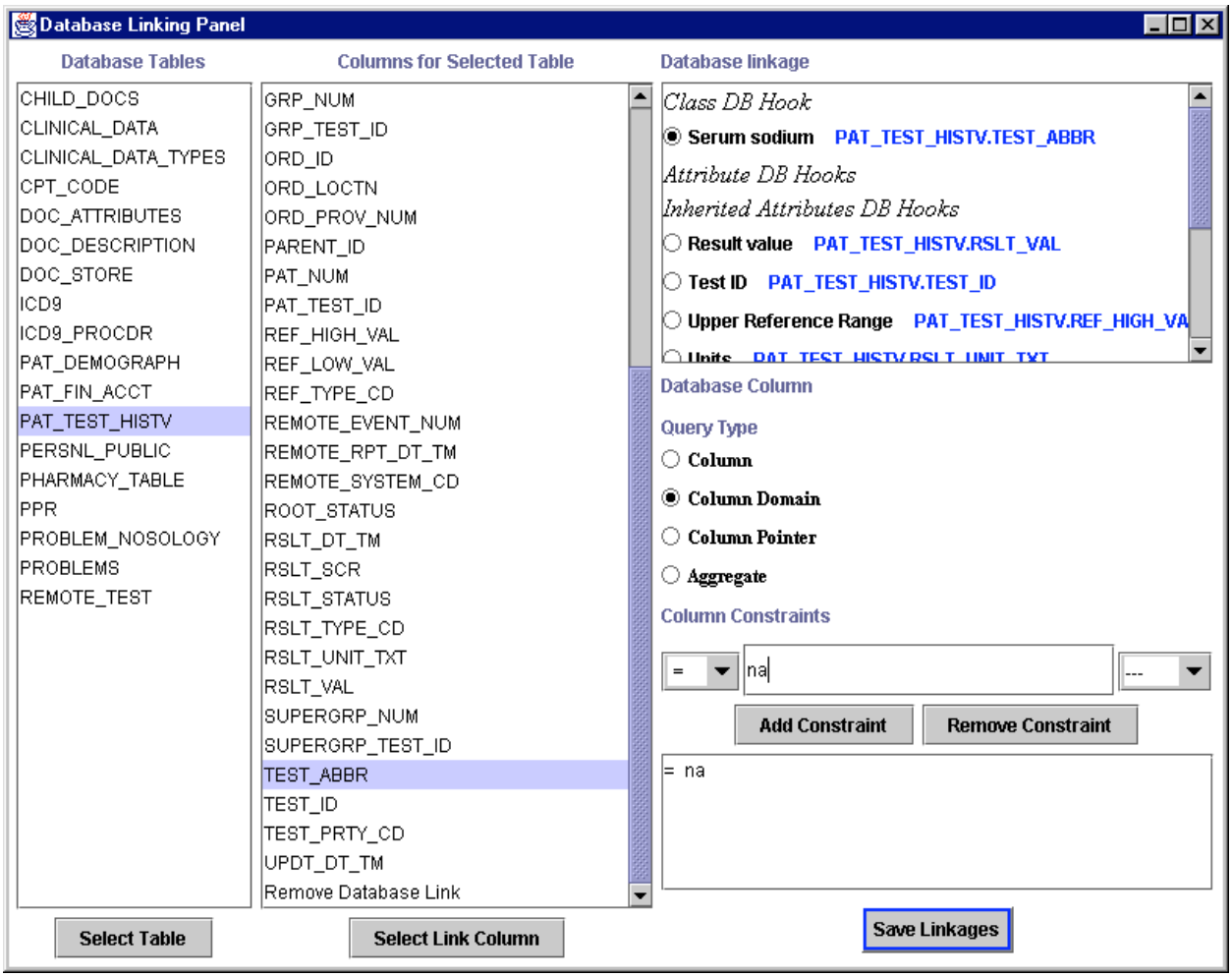


Figure 8. User interface for linking a node to a relational database. The user selects the table and column to link with each element of the node database link, including the main concept (serum sodium in this example) and attributes (e.g. Result value, Test ID, etc.)

2) *Column domain.* This is the example given in Figure 8, where the node main concept is in the domain of the column, i.e. one of the possible values of the column. In the majority of cases, the column contains a label that is equivalent to the node identity and the actual data elements are contained within other columns.

3) *Column pointer.* The column does not contain data directly related with the MIC, but instead contains a pointer to another column, possibly in a different table.
4) *Aggregate.* As discussed previously, this storage type indicates that the node is not directly linked to the database, but derives its information from nodes within its sub-hierarchy.

Database links also contain information linking attributes of the node to their respective data elements. In many relational databases, all the data elements for a node are contained within one table. This makes the linking process relatively straightforward.

3.5 Query Processing

Once the semantic equivalencies between networks have been identified through the matching process, queries are executed by retrieving the matching nodes from remote networks. To retrieve a thyroid function panel, for example, the system identifies the semantically equivalent concept in the remote network by looking up the node match. The information contained in the remote nodes database link is then used to retrieve the data directly from the remote database.

To facilitate the retrieval and formatting of data, a graphical interface for query processing has been designed to enable basic organization and sorting of the query results. An example of this interface is shown in Figure 9.

3.6 Platform considerations

MEDIATE was developed using the Java programming language to utilize Java's portability and its network and database capabilities. Support for Java applications is almost ubiquitous among operating systems, and standard Java classes can implement many network operations. The many drivers that are available to support Java database calls facilitate linking semantic networks to native databases.

To benefit from emerging data interchange standards, nodes are encoded using tagged fields that can easily be exported as XML documents. Since the nodes contain all the information needed to

construct the semantic network, information exchange between MEDIATE systems can piggyback on standard communications protocols such as HL7 in a simple fashion.

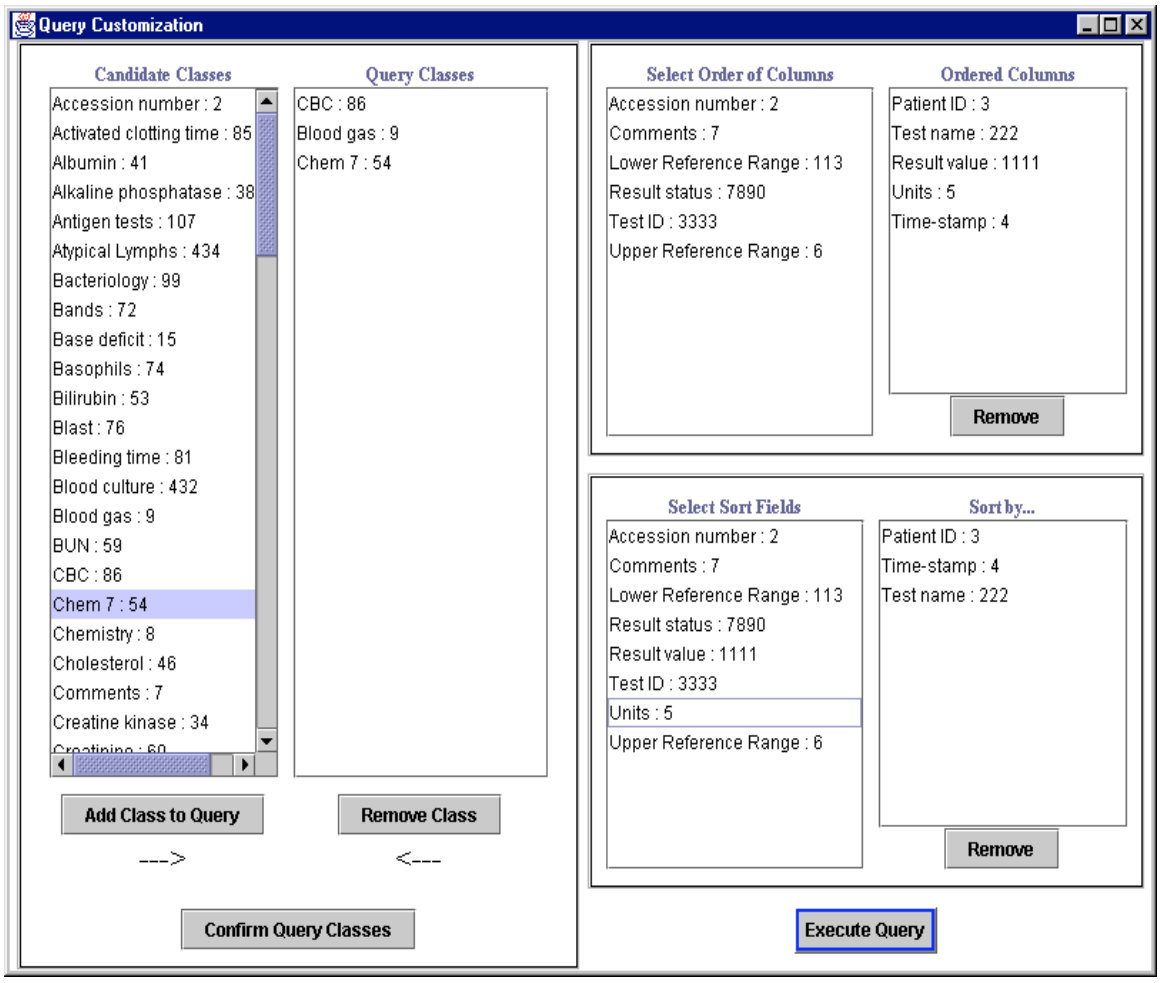


Figure 9. User interface for customizing a query. Once the query classes have been selected, either manually or automatically by the system, the presentation of the results can be organized in the panels on the right. The order of data presentation as well as the manner in which the data is sorted can be specified.

4 EXPERIMENTAL DESIGN

To evaluate the ideas and functionality of MEDIATE, two laboratory databases were represented and tested. Laboratory databases have several characteristics that make them attractive as an initial test platform.

1) *Clinical importance.* Laboratory information is critical to clinical decision making, as illustrated by the scenarios in section 1.1.1. Healthcare providers depend upon accurate delivery of test results on a daily basis, and communication of these results between providers also has an extremely high priority.
2) *Ubiquitous implementation.* Within large health care facilities, laboratory test results are often the first type of medical information to be made accessible electronically. Some form of laboratory database is available in virtually all hospitals and in many large clinics.
3) *Organized data structure*. Laboratory tests have traditionally been divided into categories that are used to support effective communication between health care providers. Although these categories are not always reflected within the database structure, they help inform the structure of the semantic network representations within MEDIATE. Examples of some of the top level categories include: hematology, chemistry, microbiology, pathology, and radiology. These categories, along with others, will be represented as nodes within the semantic networks.
4) *Ability to leverage other healthcare standards.* Because of the clinical importance of laboratory information, there are efforts in many areas to improve communication of test results. Some of these systems can be utilized by the MEDIATE platform. For example, the medical vocabulary contained in the UMLS Metathesaurus is specified in great detail for the laboratory domain, which enhances its utility in MEDIATE. And communications between medical information systems can utilize HL7 as a standard messaging protocol that is widely implemented.

The experimental setup is designed to provide initial insight into the feasibility of this system and serves as a proof-of-concept rather than as an investigation to gather data for empirical analysis. Although the original intent was to implement and evaluate MEDIATE on multiple

laboratory databases, a variety of factors precluded this possibility. Nevertheless, the two laboratory databases which were included in this experiment proved disparate enough to provide a rich testing environment for MEDIATE.

4.1 Databases

4.1.1 Pediatric Hospital

The first database is a test database of laboratory results from a large academic pediatric hospital (Hospital A). This test database contains actual laboratory results, but identifying patient information has been altered. A "scrubbed" database such as this one can be use for testing without the need to consider issues of informed consent or patient confidentiality.

The database itself is a relational database that contains the vast majority of laboratory results in a single table named "Pat_Test_HistV". Although all the database tables are available through the user interface, all the laboratory tests that were represented within MEDIATE are actually from Pat_Test_HistV.

The table structure for Pat_Test_HistV stores test results as column domains, where the columns are attributes of a laboratory test such as Test_ID , Test_Abbr and Rslt_Val, and the test results themselves are possible values for each column. Thus, a specific test result is obtained not by addressing a specific column, but by using a test attribute as a constraint on a column within the table. Although this is the most space efficient way to store these results, elucidating the nature of the tests has a higher-level of complexity because of inadequate documentation.

A "data dictionary" relating test names with other identifiers (i.e. the Test_ID and Test_Abbr) does not exist. Linking a node representation of a laboratory test to the database then becomes an exercise in decoding cryptic test abbreviations contained in the Test_Abbr field of Pat_Test_HistV. In addition, the evolution of the database over time has led to variations in some of the test abbreviations that make it difficult to discern the true meaning of the abbreviation. The lack of documentation for the test abbreviations leads to some of the semantic ambiguity discussed in previous sections.

The semantic ambiguity is not only a problem for testing MEDIATE. In discussions with Information Systems personnel at Hospital A and other research investigators, the difficulty in interpreting results from Pat_Test_HistV has had a negative impact on many projects. Ideally, the entire database should have appropriate documentation to address some of these problems. With the lack of such documentation, the semantic network representations in MEDIATE could actually serve as ad hoc documentation in many situations. The difficulty, of course, lies in the creation of the network representations in the first place.

4.1.2 **Oncology Institute**

The second database is represented by table information from the laboratory database at a large academic oncology institute (Hospital B). No scrubbed information or test database was available from this institution, so no actual patient data was used.

The Hospital B database is also a relational database, and all the laboratory test results are contained in a single table, named (appropriately enough) Lab_Results. Similar to the database for Hospital A, the table structure for Hospital B stores test results as column domains.

This database also had tables that relate test orders to test abbreviations. This is useful because a test "order" may consist of one or more actual tests. For example, a serum sodium order is linked directly to a serum sodium test, whereas a white cell differential count order is linked to multiple tests each representing a different type of white cell. All of this structure is captured in the semantic network representation.

Unfortunately, the Hospital B database is similar to that of Hospital A in that there is no table or "data dictionary" which relates test abbreviations to clinical test names. Since the database tables utilize test abbreviations, this leads to similar problems of name interpretation as discussed previously for Hospital A.

4.1.3 **Other Databases**

A similar theme of semantic ambiguity was a deterrent to the utilization of a third database from an academic general hospital (Hospital C). The motivation to include this database in the

investigation is that the database is based on the MUMPS file system, which is a hierarchical system rather than relational.

The semantic ambiguity in the Hospital C laboratory system is even greater than in the two previous systems. All laboratory tests are referenced by alphanumeric designations that have no correspondence with the clinical names. These designations, such as “a1” and “b2”, actually correspond to “print fields” within hard-coded report forms which are used to display test results.

Although data dictionaries exist to relate the print fields with test names, these dictionaries are scattered and have not necessarily been updated to reflect the current use of the system. During attempts to represent the Hospital C database within MEDIATE, it was difficult for the database administrators to produce a collated list of test orders, names, and print fields. Because of this difficulty, the database was eventually excluded from testing because the semantic structure of the database was not possible to ascertain at a detailed level in time for the completion of this investigation.

A fourth relational database from another academic general hospital was excluded because of similar difficulties obtaining detailed documentation about the relationships between test orders, test names, and database fields.

4.2 Semantic Network Representation

The semantic network representing the laboratory database from Hospital A was constructed in a top-down fashion. Higher-level concepts were added to the network first (e.g. Hematology, Chemistry, and Microbiology), and sub-concepts were iteratively added until the level where component test results from the database were required.

To determine which component laboratory tests were available, a database query was executed to retrieve all unique entries in the Test_Abbr field of Pat_Test_HistV. The investigator then parsed these abbreviations, and subsets of the available tests were assigned to nodes within the semantic network. As discussed previously, some of the test abbreviations were not interpretable, and

these were not assigned to nodes. Concepts that fell outside the scope of the higher-level nodes were also excluded from the semantic network. In total, 101 nodes were assigned.

Because the component test results from Hospital B were available as a list, the semantic network representation for that database was built in a bottom-up fashion. A small auxiliary program was written to create the lowest hierarchical levels of the network using the table which links test orders to component tests. All the component tests were instantiated as leaf nodes of the network, and the test orders were instantiated as higher-level nodes. Building upon the test orders, concepts were iteratively assigned to group lower level concepts until the "root" concept of Laboratory Test was reached. 353 total nodes were assigned in the semantic network representing Hospital B.

For both Hospital A and Hospital B network representations, the relationship links and UMLS links for all nodes were assigned by the investigator.

To test the robustness of the semantic concept matching algorithms, variations of the semantic networks for both hospitals were created. The first variation eliminates all the UMLS links from non-leaf nodes of the network. In other words, all higher-level nodes were not instantiated with UMLS Metathesaurus concepts. This forces the concept matching process to function by only utilizing contexts for higher-level nodes. Forcing this mode of concept matching is a more "pure" test of the theory that useful semantic information is embodied in the relationship links of the networks.

Other variations in the semantic networks implement different relationship links to determine the effects those links may have on the matching process. These relationship variations represent alternative methods of encoding semantic information into a network. For example, "bacteriology" laboratory tests can be viewed as a subclass or subset of "microbiology", with different ramifications in terms of inheritance.

In total, four variations of the semantic network were produced for each hospital. These are:

1) Baseline network using subclass and subset relationships for higher-level group tests, with fully instantiated UMLS links,
2) Same network as above with instantiated UMLS links only for leaf nodes.
3) Network using subclass relationships for higher-level group tests, UMLS links only for leaf nodes.
4) Network using subset relationships for higher-level group tests, UMLS links only for leaf nodes.

Although variations in the relationship links can still result in a semantically "valid" network, constraints and dependencies between the relationships do exist. Perhaps the most important of these is the inheritance relationship, where subclasses are highly dependent upon superclasses for the establishment of the subclass attributes. Thus, changing bacteriology from a subclass of microbiology to a subset means that all the subclasses of bacteriology lose the properties they originally inherited from microbiology and the superclasses of microbiology. Breaking the subclass/superclass hierarchy has profound effects on the inheritance of attributes for all subclasses lower in the hierarchy. Modifications to the semantic network must take these inheritance effects into consideration.

4.3 **Database Queries**

Only the database from Hospital A was available in a scrubbed form suitable for testing. Using this database, sample queries were executed for multiple higher-level nodes (aggregates of lower level nodes) as well as leaf node nodes. Query results were also formatted and sorted by different combinations of data fields to test those functions.

Exhaustive querying of all the nodes was not performed. Instead, representative samples of nodes were queried using MEDIATE, and the results were compared with the results of direct SQL queries of the database.

The interface with the laboratory database was accomplished with MySQL, an open software database manager freely available for several operating systems. MySQL provides its own

software drivers for interfacing with Java applications (through the JDBC database classes), which eased integration of the native database into the MEDIATE test system.

4.4 **Concept matching**

Concept matching was performed between all configurations of semantic networks for both hospitals. Since there are 4 variations of each semantic network, a total of sixteen concept matching runs were performed, and the results were analyzed for the following measures:

1) Percentage of direct matches, generalized matches, and non-matches.
2) Quality scores for all matched nodes.
3) Comparison of variation in node matches based on semantic network configuration.

In addition to the concept matches which were performed during the concept matching runs, leaf matches were performed for all the aggregate (non-leaf) nodes in both networks, and these results are presented separately.

5 EXPERIMENTAL RESULTS

Several iterations of testing, evaluation, and modification of the system were performed before arriving at the results presented here. No major changes were made to the system architecture or the fundamental nature of the concept matching algorithms. There were many fine nuances, however, which required tweaking for optimal performance.

The hierarchical tree-like structure of the semantic network representations for both databases forced a modification in the matching algorithms to limit the number of candidate matches that were produced. Because of the large "fan-out" of linkages between some concepts and their subcomponents, the search patterns of the matching algorithms sometimes returned multiple leaf nodes that could not be distinguished based on contextual information. In this situation, literally dozens of nodes might be returned as specious candidate matches from one of the matching algorithms, overwhelming the signal of more reasonable matches from a different algorithm. Therefore, a threshold was enforced which limited the number of candidate matches from any given algorithm. If the threshold (currently set at three nodes) is exceeded, all the candidate matches from that algorithm are discarded as probable noise.

Another modification that produced minor improvements in the matching performance utilized the quality metric to dynamically assess candidate matches as the algorithms were executed. In a small number of cases, a node that was traversed earlier in the network search was a better candidate match than the "final" node discovered at the end of the search. Although using the quality metric slows down the matching process, the improved matching performance outweighs the inconvenience of a drop in efficiency.

Encouragingly, the experimental results did not undergo a radical change in nature after all the optimizations were in place (compared to the first "clean" run in which no gross programming errors were discovered). Quantitative results such as matching percentage and average quality scores did not change by more than a few percentage points. Manual inspection of the concept matches revealed a small number of improvements as judged by clinical relevance, but the majority of the matching results remained stable throughout the test iterations.

5.1 Database Queries

Sample queries using the scrubbed database from Hospital A revealed no differences between the data retrieved by MEDIATE and data retrieved through direct SQL queries. Although exhaustive testing was not performed, both leaf nodes and higher-level aggregate nodes were accurately retrieved when results were compared with direct database queries using manually coded SQL.

5.2 Overview of Concept matching Results

Of the 101 nodes in Hospital A, 68 nodes (67%) were "direct" matches that were found during the first two phases of the matching processes. The presence of a direct match implies that the target concept exists in both networks. 87 (25%) of the 353 nodes from Hospital B had direct matches identified.

Direct matches can be subcategorized into "UMLS" matches and "non-UMLS" matches. In a UMLS match, the matched nodes correspond to the match obtained through the UMLS link. In other words, the match found through the context-sensitive algorithms is the same as the match found through the UMLS terminology link. For example, node "CBC" from Hospital A matched node "cbc" from Hospital B, and both nodes are linked to the UMLS concept "blood count, complete".

There are cases, however, where the quality metric indicates that an optimal match differs from the UMLS match. For example, the node "ldlp" from Hospital B and the node "Lipid profile" from Hospital A match through the UMLS concept "test, lipids profile". This match, with a quality score of 40, is not as the good as the match between "ldlp" and the node "Lipids" from Hospital, A which has a quality score of 80. This discrepancy arises because "ldlp" actually consists of 5 subcomponents, whereas "Lipid profile" consists of 2 subcomponents and "Lipids" consists of 4 subcomponents. Manual inspection of the subcomponents reveals that the non-UMLS match of "ldlp" with "Lipids" is the better match. (See section 5.8 for another discussion of this example.)

Generally, most non-UMLS matches are performed because no matching UMLS concepts could be identified to instantiate the UMLS link for the pertinent nodes. In the experimental setup, this situation was mimicked by creating network configurations where only leaf nodes were instantiated with UMLS links. This tested the performance of the context-sensitive algorithms more rigorously and allowed a direct comparison with matches utilizing UMLS terminology.

Evaluation of the matches in the networks where UMLS links were not fully instantiated shows very few cases where they differed from the matches in the UMLS fully instantiated networks. These cases from Hospital A are shown in Table 2.

Concept	**Node match (networks with UMLS links fully instantiated)**	**Node match (networks with only leaf nodes UMLS links)**
Bacteriology	Bacteriology Culture	Bacteriology Culture
	Bacteriology Labs*	
CBC	cbc*	cbca
	cbca*	long1
	cbcd*	
	long1	
Lipid profile	ldlp*	ldlp
		Chemistry Labs
Proteins	bmauto	bmauto
	iepu	iepu
	tp*	
Virology	bmaut2	bmaut2
	Virology Labs*	
WBC differential	cbca	cbca
	difa*	difa
	diff*	
	WBC differential count*	

Table 2. Comparison of matches in networks fully instantiated with UMLS links vs. networks in which only leaf nodes had UMLS links. Concepts marked with a star (*) indicate a UMLS terminology match.

Direct matches are produced when the local target concept is also "found" in the remote database. Importantly, the lack of a direct match implies that the concept does not exist in the remote database. To evaluate this proposition, all the nodes that did not have direct matches were manually inspected for both hospitals. With the exception of attribute nodes (i.e. nodes for which the sole relationship is "attribute-of"), none of these remaining concepts could be identified in

the other remote network. [a] Therefore, the lack of a direct match has 100% negative predictive value for the existence of a concept in the remote network (with the exception of attribute nodes). This characteristic enables the automatic identification of concept disparities between databases, which may play an important role in data integration efforts such as the creation of a data repository.

The corollary proposition, that direct matches identify all concepts that exist in both databases, is true with two caveats.

The first caveat is that "terminological equivalence" is not the same as "semantic equivalence" within this system. In MEDIATE, semantic equivalence implies some degree of commonality in the semantic context of the two nodes. In particular, there must be some information content, as indicated by matched subcomponents, that both nodes have in common. For example, the fact that "WBC differential" directly matches with "difa" implies that both nodes have subcomponents that are equivalent (e.g. "PMN" = "neutrophils", "Bands" = "band", "Monocytes" = "mono", etc.)[b].

Referring again to Table 2, there are some a few concepts in which the UMLS matches (which correspond to terminological equivalence) are not found when the networks are not fully instantiated with UMLS links. For example, the match between "Virology" and "Virology Labs" is not found. Closer analysis reveals that although the two concepts are terminologically equivalent, they have absolutely zero commonality in the data that they contain (as represented by their respective subcomponents). This reflects in a quality score of 0 for this match. In cases such as the "CBC" concept, some UMLS matches are not discovered because the matching algorithms produced more optimal matches. Thus, "cbca" was found because it is a better match than "cbcd" or "cbc" for the "CBC" concept.

[a] Examples of attributes that were found in both databases: "Test name" and "Test name", "Test result" and "Result value", and "Result units" and "Units".

[b] To prevent this from becoming a circular argument, semantic equivalence is grounded at the level of the leaf nodes where the lack of a sub-hierarchy forces the equivalence inference to be based the terminological equivalence of the UMLS links.

The second caveat involves the issue of how semantic equivalence relates to real world practicality. The fact that two concepts are "semantically equivalent" by computation does not necessarily correspond directly with "clinical equivalence". For example, the direct match between Hospital B concept "newa" and Hospital A concept "Chemistry" indicates a high degree of overlap, but the clinical equivalence of this match remains an open question.[a] Manual inspection of the direct matches, however, reveals that the majority of matches have clinical equivalence in addition to semantic equivalence.

Overall, the experimental results support the assertion that MEDIATE enables automated identification of semantically equivalent concepts, bearing in mind the previously discussed caveats. Detailed quantitative results and further analysis are presented in the following sections.

5.3 Matching Percentages

The results from the sixteen concept matching experiments for Hospital A are displayed in Table 3. There were a total of 101 nodes in the semantic network representing the database from Hospital A. The sixteen matching runs are the result of applying the matching process to every cross-combination of network configuration from Hospital A and Hospital B (as explained previously in sections 4.2 and 4.4).

There is surprisingly little variation in the percentage of nodes matched throughout the sixteen experimental runs. As expected, there were more UMLS matches in the experiment involving full instantiation of UMLS links in networks from both Hospital A and Hospital B (run #1). Other than that, the percentage of matches remained unchanged throughout the matching experiments.

Similar results can be seen for the 353 nodes in the semantic network from Hospital B, shown in Table 4. In run #1, there is the expected rise in the percentage of UMLS matches due to the fact that both networks are instantiated with UMLS links to the fullest extent possible. Otherwise, there is essentially no variation between the sixteen different network configurations.

[a] Without access to the database designer, the name "newa" remains undecipherable at this time.

Match Run	Direct Matched	UMLS Matched	Non-UMLS Matched	Generalized Matched	Unmatched
#1: 1 x A	68 (67.0%)	54 (53.0%)	14 (14.0%)	19 (19.0%)	14 (14.0%)
#2: 1 x B	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#3: 1 x C	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#4: 1 x D	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#5: 2 x A	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#6: 2 x B	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#7: 2 x C	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#8: 2 x D	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#9: 3 x A	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#10: 3 x B	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#11: 3 x C	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#12: 3 x D	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#13: 4 x A	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#14: 4 x B	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#15: 4 x C	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)
#16: 4 x D	68 (67.0%)	46 (46.0%)	22 (22.0%)	19 (19.0%)	14 (14.0%)

Table 3. Results for Hospital A concept matching experiments. Directly matched nodes are comprised of UMLS matched and non-UMLS matched nodes (Direct = UMLS + non-UMLS). Total matches are comprised of Direct and Generalized matches (Total nodes = Direct + Generalized + Unmatched). **Legend** for Matching Run network configurations: Hospital A => 1: base network, 2: UMLS links only on leaf nodes, 3: subset/superset relationships excluded, and 4: subclass/superclass relationships excluded. Hospital B => A: base network, B: UMLS links only on leaf nodes, C: subset/superset relationships excluded, and D: subclass/superclass relationships excluded.

Match Run	Direct Matched	UMLS Matched	Non-UMLS Matched	Generalized Matched	Unmatched
#1: 1 x A	88 (25.0%)	64 (18.0%)	24 (7.0%)	148 (42.0%)	117 (33.0%)
#2: 1 x B	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#3: 1 x C	87 (25.0%)	49 (14.0%)	38 (11.0%)	150 (42.0%)	116 (33.0%)
#4: 1 x D	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#5: 2 x A	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#6: 2 x B	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#7: 2 x C	87 (25.0%)	49 (14.0%)	38 (11.0%)	150 (42.0%)	116 (33.0%)
#8: 2 x D	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#9: 3 x A	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#10: 3 x B	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#11: 3 x C	87 (25.0%)	49 (14.0%)	38 (11.0%)	150 (42.0%)	116 (33.0%)
#12: 3 x D	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#13: 4 x A	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#14: 4 x B	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)
#15: 4 x C	87 (25.0%)	49 (14.0%)	38 (11.0%)	150 (42.0%)	116 (33.0%)
#16: 4 x D	87 (25.0%)	49 (14.0%)	38 (11.0%)	149 (42.0%)	117 (33.0%)

Table 4. Results for Hospital B concept matching experiments. Format is identical to Table 3.

5.4 Match Quality

The percentage tables offer a rough evaluation of number of nodes that were matched, but do not contain information about the quality of the node matches that were made. The "score" portion of the match quality metric is a distillation of the match appropriateness between two nodes. Tables 5 and 6 show the average quality scores for the matches that were obtained for each of the network configurations.

Match Run	Direct Matched	UMLS Matched	Non-UMLS Matched	Generalized Matched	Total Matches
#1: 1 x A	80.06	92.39	32.50	0.00	53.90
#2 - #16	80.06	100.00	38.36	0.00	53.90

Table 5. Average quality scores for Hospital A concept matching experiments. Range of scores is 0 to 100. Legend for Matching Run network configurations: Hospital A => 1: base network. Hospital B => A: base network. The quality score for runs 2 through 16 did not vary, and are therefore presented as one row.

The lower average quality score for UMLS matches in run #1 arises from the fact that many non-leaf nodes in that particular network configuration have UMLS matches. Since quality scores are dependent upon variations in the sub-hierarchy, UMLS leaf matches are assigned the maximum score by default, whereas non-leaf matches will almost always have a slightly lower score. For example, the leaf node "Reticulocytes" from Hospital A is a UMLS match with the leaf node "ret" from Hospital B, with a quality score of 100. The non-leaf node "WBC differential", however, is a UMLS match with non-leaf node "difa", but only has a quality score of 46.

Despite the lower average quality scores for UMLS and non-UMLS matches in run #1 compared to the other matching runs, the overall quality score of 53.90 remains the same as for all the other experimental matching runs. This reflects the fact that there are a significantly higher proportion of UMLS matches in run #1 (see table 3), which balances the lower individual scores.

For Hospital B, run #1 again illustrates the lower average quality score for UMLS matches. In addition, non-UMLS matches also demonstrate a markedly lower average quality score. Again, the total quality score for runs #1 is the same as for runs #2 - #4 because of the greater percentage of UMLS matches in run #1 (see Table 4).

Match Run	Direct Matched	UMLS Matched	Non-UMLS Matched	Generalized Matched	Total Matches
#1: 1 x A	69.90	85.44	28.46	0.00	17.42
#2: 1 x B	70.70	100.00	32.92	0.00	17.42
#3: 1 x C	70.69	100.00	32.89	0.00	17.42
#4: 1 x D	70.69	100.00	32.89	0.00	17.42
#5: 2 x A	70.97	100.00	33.53	0.00	17.49
#6: 2 x B	70.97	100.00	33.53	0.00	17.49
#7: 2 x C	70.95	100.00	33.50	0.00	17.49
#8: 2 x D	70.95	100.00	33.50	0.00	17.49
#9: 3 x A	70.97	100.00	33.53	0.00	17.49
#10: 3 x B	70.97	100.00	33.53	0.00	17.49
#11: 3 x C	70.95	100.00	33.50	0.00	17.49
#12: 3 x D	70.95	100.00	33.50	0.00	17.49
#13: 4 x A	70.97	100.00	33.53	0.00	17.49
#14: 4 x B	70.97	100.00	33.53	0.00	17.49
#15: 4 x C	70.95	100.00	33.50	0.00	17.49
#16: 4 x D	70.95	100.00	33.50	0.00	17.49

Table 6. Average quality scores for Hospital B concept matching experiments. There are minor fluctuations in the scores through the various network configurations.

Legend for Matching Run network configurations: Hospital A => 1: base network, 2: UMLS links only on leaf nodes, 3: subset/superset relationships excluded, and 4: subclass/superclass relationships excluded. Hospital B => A: base network, B: UMLS links only on leaf nodes, C: subset/superset relationships excluded, and D: subclass/superclass relationships excluded.

Lower average quality scores are seen in runs #1 - #4. These scores result from the difference in a single concept match. For the Hospital B concept “iepu”, runs #1 - #4 had Hospital A concept “Chemistry” as the best overall match with a quality score of 6. For the remaining runs #5 - #16, the best overall match was with Hospital A concept “Proteins”, which had a quality score of 29. This result most likely represents an experimental artifact of the network configuration used for Hospital A in runs #1 - #4, in which a relationship link was inadvertently altered compared to the networks used for the remaining runs.[a]

[a] There is no computational reason to suspect that fully instantiating the UMLS links for Hospital A would otherwise produce this pattern of matching, since the corresponding networks for Hospital B have varying degrees of UMLS link instantiation in runs #1 - #4, and thus would be expected to produce varying match results if the UMLS link was the critical influence.

For both hospitals, generalized matches resulted in a quality score of zero. In retrospect, this is not surprising because generalized matches only occur when the system is unable to match the target node through any other means. This implies a lack of network "overlap" for the target node, and the lack of commonality in the local neighborhood of the node reflects in the quality score.

5.5 Unmatched Nodes

Unmatched nodes in both networks occurred in two categories. The first category consists of nodes that are attributes of other concepts. These nodes are connected to other nodes solely by the "attributeOf" relationship. Examples of these nodes include: Accession number, Lower Reference Range, Patient ID, Result status, and Result value.

As explained earlier in section 3.1.2.1.6, the attribute relationship is orthogonal to other relationships within this system. Therefore, the "attribute-of" relationship is not utilized in any of the concept matching algorithms, and the consequent result is that attributes are not matched.

The other category consists of disconnected nodes. These nodes were created at some point during construction of the semantic networks, but were not connected to the main network by any relationship links. This happened either through oversight, or because it was not possible to interpret the clinical meaning of the node from the hospitals' abbreviated name. Examples of these nodes include: hemogram, afp, ahbs, aldo, ana, apad, apai, b12, b2m, bhgbe, and biopsy.

5.6 Clinical Relevance

The quality scores give a "structural" measure of how well two nodes match, based upon the similarity between their network sub-hierarchies. To determine the clinical relevance of the node matches, however, requires detailed human examination of the actual matches. The following section provides a summary of pertinent results, while the complete list of matches is provided in Appendix A.

5.6.1 **Direct matches**

5.6.1.1 ***UMLS leaf matches***

UMLS matches of leaf nodes were straightforward and unexciting. For semantic networks from both Hospital A and Hospital B, all the UMLS matches of leaf nodes were clinically accurate and reflected the matching of synonymous concepts between the hospitals. There were a few cases in which there were multiple leaf nodes in the Hospital B network, which were semantically identical (e.g. neutrophils and poly). Under these circumstances, the matching algorithms appropriately matched the node from Hospital A (e.g. PMN[a]) with all of the synonymous nodes from Hospital B.

5.6.1.2 ***UMLS non-leaf matches***

Non-leaf node matches were more interesting because there was more inherent semantic ambiguity about the concepts being matched.

In cases where there was a 1-to-1 node match, all the node matches were clinically accurate and relevant. For example, the node "DIC screen" from Hospital A matched appropriately with the node "dic" from Hospital B. The quality score for this match was 50, indicating differences between the components which were contained in the test panel.

In some cases, one node matched with multiple nodes from the other hospital's semantic network through the UMLS link. For example, the node "WBC differential" from Hospital A matched nodes "difa", "diff", and "WBC differential count" from Hospital B. All three nodes from Hospital B represent variations of differential counts for white blood cells, and none are more "correct" than the others: they merely contain different component tests. In addition, the Hospital B nodes share similar UMLS links because no UMLS concepts differentiate between them. Choosing the best match for the Hospital A node involves choosing the node with the highest quality score. In this case, matching with "difa" gave the highest quality score of 46, vs. 35 for "diff" and 32 for "WBC differential count".

[a] PMN = polymorphonuclear white cell, synonymous with neutrophil and poly (abbreviation of PMN).

It is not always true, however, that the UMLS match for a non-leaf node presents the highest quality match. The node "Bacteriology" from Hospital A matches Hospital B node "Bacteriology Labs" with a quality score of 25. Although this is a UMLS match, the quality score is lower than the non-UMLS match with Hospital B node "Bacteriology Culture", which has a score of 27, indicating a higher degree of overlap between the concepts.

5.6.1.3 *Non-UMLS matches*

If nodes are matched through the primary matching process (phases 1 and 2 in section 3.3.1.1) but do not match through their UMLS links, then they are "direct Non-UMLS" matches. These node matches rely upon the interaction between the matching algorithms and the semantic context rather than the common terminology provided by the UMLS links. The utility of this type of matching was particularly tested by the network configurations in matching runs #6 - #16, where only leaf nodes in both hospital networks had UMLS links instantiated.

Many direct non-UMLS matches accurately matched synonymous medical concepts. Examples of these matches include: "Chem 7" = "basic7", "Blood gas" = "bg", "Liver Function Tests" = "hfp", and "Lipid profile" = "ldlp"[a].

Importantly, these matches were found despite differences in the composition of the test components. For example, "Liver Function Tests" is composed of the concepts "SGOT", "SGPT", and "bilirubin", while "hfp" is composed of the concepts "sgot", "sgpt", "bili, total", "bili, direct", "tp", "alb", and "ap".[b] Overall, 32% of the direct non-UMLS matches for Hospital A produced synonymous concepts, and 16% of the direct non-UMLS matches for Hospital B were synonymous.

Along with clinical accuracy and relevance, these matches also provided some interesting insights into the clinical usage of the tests. The "Chem 7" profile of serum chemistry measurements is a fairly standard panel of seven tests used in medical centers throughout the world. Yet the "basic7" panel from Hospital B actually contains 8 tests! The extra laboratory test

[a] ldlp = low density lipoprotein profile (interpretation by investigator).
[b] sgot = serum glutamic oxaloacetic transaminase, sgpt = serum glutamic pyruvic transaminase, tp = total protein, alb = albumin, ap = alkaline phosphatase (tp, alb, and ap interpretation by investigator).

is “ca” (calcium), which is a metabolite that is often irregular in patients with oncological disease.

There were no matches in which the nodes were completely unrelated on a clinical basis. Instead, the remaining matches had variable degrees of clinical relevance. In addition, there were no cases in which a more clinically “appropriate” match was apparent.

The degree of clinical relevance in the non-synonymous matches is more difficult to evaluate, although the quality score gives some idea of the “overlap” between concepts as measured by the similarity between sub-hierarchies. On the lower end of clinical relevance, some matched concepts fall into the same general “category” of laboratory test. For example, Hospital B node “g6p”[a] matches with Hospital A node “CBC”. The nodes are matched because both concepts contain the subcomponent “hemoglobin”. Clinically, “g6p” tests would be used to screen for anemia secondary to a deficiency in the enzyme G6PD, and a CBC gives information about various blood components, among which are red blood cell parameters which are important in the evaluation of anemia.

As the network configurations changed through the sixteen matching runs, only a few of the non-leaf node matches exhibited minor variations. For example, Hospital B node “comp12” was matched with two different Hospital A nodes (“Chem 7” and “Chemistry”) depending on the network configurations. As reflected by the quantitative results shown previously, the vast majority of matches were stable through all the different network configurations. For the few matches that exhibited variation, all of the variations were “reasonable” choices as judged by the component contents of the tests.

One of the most difficult circumstances to judge clinical relevance occurs when a concept has no corresponding clinical concept in the other hospital database or in the UMLS Metathesaurus. In other words, the semantic concept is nonexistent in the universe of the other network. In this investigation, this arose more commonly with concepts from Hospital B because of the large variety of oncology specific test panels present in the database. There are a large number of

[a] g6p = glucose 6-phosphate dehydrogenase profile (interpretation by investigator).

variations on hematological tests, and also several panels of tests used for bone marrow evaluations and bone marrow transplant profiling.

Generally, this type of node match showed useful areas of clinical overlap, but failed to capture the "essence" of the originating concept. For example, Hospital B node "bma"[a] matched with Hospital A node "WBC differential". In some ways, this is a very good match because many of the component tests are identical. The purpose of a bone marrow aspirate is to visualize the types of cells that are present in the marrow, and the WBC differential performs essentially the same task on circulating white blood cells. The crucial difference, however, is that the bone marrow aspirate panel contains components for evaluation of red blood cell elements as well as a comment field that can be used to evaluate other cell types, including tumor cells that have metastasized to the bone marrow. Although the difference between the component tests is small, the clinical "meaning" of a bone marrow aspirate is not totally captured by the WBC differential count.

Other matches for nonexistent semantic concepts proved even more problematic. The bone marrow transplant test panels from Hospital B often combine elements of several major test categories (e.g. "higher" level concepts such as hematology, microbiology, and chemistry). None of the concepts from Hospital A, however, inherit or combine elements from multiple higher-level concepts in the same manner. As a result, these matches generally pair the Hospital B bone marrow test with only one of the higher-level Hospital A concepts and exclude the other higher-level concepts. For example, the bone marrow test panel "bmallo" consists of 43 component tests contained within the sub-hierarchies of the concepts hematology, chemistry, and virology. This node matched with the node "Hematology" from Hospital A (because of the greater overlap with this concept) and excluded any concepts from the "Chemistry" and "Virology" sub-hierarchies.

In general, the clinical relevance of matches for concepts that are nonexistent in the other database is open to question. If a match is made at all, it implies an overlap between the concepts which MEDIATE exploits in order to define the match. Human clinical judgment, however, must still prevail when evaluating the usefulness of the overlap found in such matches.

[a] bma = bone marrow aspirate (interpretation by investigator).

5.6.2 **Generalized matches**

Generalized matches occur when no direct match is found during the initial phases of the matching process. MEDIATE subsequently attempts to match a higher-level concept that might encompass the concept that is being matched.

One gauge of utility is the number of generalized matches that correspond to the root concept of "laboratory test". This is a default match that does not contain any useful information, but very few matches fell into this category. For Hospital A, none of the nodes were generalized to the Hospital B node "Lab Test". For Hospital B, however, nine nodes were generalized to Hospital A node "Laboratory Test".[a]

Overall, clinical relevance is difficult to evaluate for generalized matches, since by definition the matching concept is nonexistent in the other database. This is reflected in the quality scores of zero for all generalized matches. Nevertheless, some matches clearly have the potential to be clinically useful. For example, Hospital B node "plasma cell" generalized to Hospital A nodes "CBC" and "Hematology". It is easy to envision that a clinician looking for information on plasma cells might find information about complete blood counts or hematology tests useful.

At the other end of the spectrum, it is not surprising to find that some generalized matches have little clinical relevance. Some of these matches involve concepts that cross sub-hierarchy boundaries, such as the bone marrow tests from Hospital B. But other matches that do not cross sub-hierarchy boundaries are still clinically irrelevant. For example, the Hospital B node "herpes ii antibody" matched with the Hospital A node "Chem 7". This match occurred because there are some matches between "Chem 7" and higher-level concepts in Hospital B that contain "herpes ii antibody" as a component. Clinically, however, there is no foreseeable circumstance under which a clinician searching for herpes antibody values would be satisfied by serum chemistries from a "Chem 7" panel.

[a] Nodes generalized to Hospital A node "Laboratory Test": balld3, bmall3, bmaut3, hbc, s/n ratio, samples to cell bank, serum storage, ua, and Virology Labs.

As is the case with direct matches, there is no way within this system to quantify "clinical relevance", and the final evaluation of clinical utility must still be rendered by human judgment.

5.7 Leaf Matches

Leaf matches were performed for only one representative network configuration, and the results are shown in Tables 7 and 8. The network configuration used for these leaf matches corresponds to run #6 from the previous matching experiments, where all relationships are instantiated but only the leaf nodes have UMLS links. Detailed leaf node matches are shown in Appendix B.

Node	*Score*	*Leaves matched*	*Leaves Unmatched*
Bacteriology	75	3	1
Blood gas	100	5	0
CBC	100	12	0
Chem 7	86	6	1
Chemistry	78	25	7
Cultures	100	3	0
DIC Screen	75	3	1
Electrolytes	58	7	5
Enzymes	60	3	2
Gram	100	1	0
Hematology	85	17	3
Laboratory test	74	46	16

Node	*Score*	*Leaves matched*	*Leaves Unmatched*
Lipid profile	100	2	0
Lipids	100	4	0
Liver Function Tests	100	3	0
Microbiology	40	4	6
Other Chemistry	100	1	0
Proteins	100	3	0
Serum lytes	100	6	0
Stains	50	1	1
Virology	17	1	5
WBC differential	100	8	0

Table 7. Hospital A Leaf Matches. Node: target node. Score: leaf match quality score = percentage of leaves matched. Leaves matched: number of leaves matched. Leaves unmatched: number of leaves unmatched.

The quality scores are the most pertinent parameter presented in the tables. Leaf matches are scored by the percentage of leaves that are matched for a given node. Thus, the score is a direct reflection of the amount of information retrievable from the remote database for an aggregate node. For leaf matches, the match quality score does not reflect "semantic equivalence" or the degree to which the target concept overlaps with a matching concept in a remote network. Instead, the quality score is a direct measure of the retrievable "information content" for the node. If the leaf match quality score is 100 (indicating all the leaves are matched), then the full information content of the node is available for retrieval.

Node	*Score*	*Leaves matched*	*Leaves Unmatched*
Bacteriology Culture	30	3	7
Bacteriology Labs	27	3	8
balld4	43	12	16
balld5	80	16	4
basic7	88	7	1
bg	33	5	10
bili	50	1	1
Blood Counts	33	14	28
BM Transplant Tests	38	33	55
bma	11	1	8
bmall2	13	1	7
bmall4	37	14	24
bmall5	79	15	4
bmallo	51	22	21
bmaut2	20	1	4
bmauto	51	25	24
cbc	50	4	4
cbca	35	13	24
cbcd	32	13	28
Chemistry	47	27	30
Labs			
comp12	87	13	2
dic	80	4	1
difa	55	6	5
diff	40	6	9
Electrolytes	86	6	1
fmmbmt	67	2	1
frap	50	1	1
g6p	50	1	1
Hematology Labs	30	18	42
hfp	86	6	1
iepu	33	2	4
iglb	33	1	2
Lab Test	27	49	130
ldlp	80	4	1
long1	38	12	20
lyte	75	3	1
newa	61	11	7
WBC differential count	41	7	10

Table 8. Hospital B Leaf Matches. Node: target node. Score: leaf match quality score = percentage of leaves matched. Leaves matched: number of leaves matched. Leaves unmatched: number of leaves unmatched.

Because of this difference between semantic equivalence and information content, leaf matches are complementary to concept matches in terms of their clinical relevance. Some nodes may have concept matches that have low clinical relevance, yet have leaf matches that have high information content and therefore a higher clinical relevance. This is particularly evident for concepts with leaves that are representative of several different categories in the remote network, such as the "BM Transplant Tests" from Hospital B. For example, the "bmallo" node from Hospital B has a concept match with the "Hematology" node from Hospital A, yet it has leaves which fit into the "Chemistry" sub-hierarchy. The leaf match for "bmallo" shows approximately 51% of the leaves matched, and a detailed examination of the matches shows nodes for both hematological and blood chemistry tests appropriately matched.[a]

[a] Matched leaf nodes for "bmallo". Node(matching node): lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); bili, total(Bilirubin); na(Serum sodium); ap(Alkaline phosphatase); eo(Eosinophils); ret(Reticulocytes); neutrophils(PMN); alb(Albumin); bun(BUN); plt(Platelet count); baso(Basophils); sgot(SGOT); cret(Creatinine); igg(IgG); sgpt(SGPT); k(Serum potassium); hematocrit(Hematocrit); blast(Blast); cl(Serum chloride).

Conversely, there are times when a concept match has a higher clinical relevance than the leaf match even if the leaf match quality score is 100. This is easily seen when there is true semantic equivalence between two nodes, but one of the nodes has a more extensive sub-hierarchy. For example, the "Liver Function Tests" node from Hospital A matches all of its leaves in a leaf match, but only has 3 leaves in its sub-hierarchy. The concept match "hfp" from Hospital B contains 7 leaves in its sub-hierarchy (which subsume the 3 leaves from "Liver Function Tests"), and represents a better match than the leaf match.

Like concept matches, the "clinical relevance" is only partially captured by the quality score. Thus, the utility of the leaf match vs. the concept match is still a judgment best left to the human user.

5.8 **Matching asymmetry**

Matching is not necessarily a symmetrical operation. For example, Hospital node "Lipid profile" is matched with Hospital B node "ldlp", but "ldlp" is matched with Hospital A node "Lipids", which is a more general concept than "Lipid profile". In this particular case, the match "ldlp" => "Lipids" occurs because "ldlp" shares more subcomponents with "Lipids" than it does with "Lipid profile", as illustrated in Table 9. In other words, "ldlp" is more semantically equivalent to "Lipids" than it is to "Lipid profile".

Matching symmetry can only be assured if there is a 1-to-1 relationship between semantically equivalent nodes in different networks. If there is a 1-to-many or many-to-many relationship, then MEDIATE attempts to find the best match in both directions and asymmetry may result.

Hospital	***Node***	***Subcomponent nodes***
A	Lipid profile	Cholesterol, Triglycerides
A	Lipids	Cholesterol, Triglycerides, HDL, LDL
B	ldlp	cholesterol, triglyeride, high dens. lipoprotein, ldl-cholesterol, very low density lipoprotein

Table 9. Subcomponents of lipid related nodes in semantic network representations.

6 DISCUSSION

6.1 Knowledge Representation

The basic problem of data exchange is one of knowledge representation. Within the domain of medical information, many different knowledge representation schemes have been investigated to facilitate the exchange of electronically stored data, although the most common representations are still those of organized free text and rigidly structured databases. MEDIATE attempts to leverage the ubiquitous presence and controlled structure of medical databases by tying those databases into the elements of a semantic network. The constraints of a formal network structure and the addition of procedural knowledge in the form of matching algorithms provide the basis for achieving the functionality provided by MEDIATE.

MEDIATE's knowledge representation system targets the "content" level of medical databases. At this point, no attempt is made to represent medical care processes or general medical knowledge that is not stored in the database. This level of representation reflects the goal of automatically identifying equivalent information content for exchange. Consequently, the complexity of more general knowledge representation systems can be avoided, the representation is easily understandable, and computation may take place more efficiently.

6.1.1 Semantic Networks

The choice of a semantic network as the basis for knowledge representation within MEDIATE was dictated by the design goals. The semantic network provides the power and the flexibility required to represent the myriad possible concepts of native databases, and the network structure provides an intuitive interface for a user to see and understand the native database.

The main reason for utilizing a semantic network data model is to capture more of the semantic content of an electronic data source. This semantic content may be explicit in the declaration of data elements, or it may be implicit in hidden relationships between the elements. Semantic networks help capture semantic content through the following mechanisms. [17, 81]

A semantic network makes relationships between concepts explicit. Since there are no restrictions on these relationships, they can be customized to provide the exact semantics required by the user. This contrasts with the "semantic overloading" that occurs frequently with relational databases, where multiple relationships between concepts are implicitly and imprecisely embodied in the table structures. These implicit relationships are not only hard to interpret, they are also difficult to communicate between information systems.

In the typical relational database schema, conceptual relationships must be inferred from the table structure and names. This may be problematic if the relationships are not apparent from casual inspection. In the database for Hospital B, for example, there are many examples of duplicated test result fields.[a] Since there is no way to indicate the synonymy relationship in the standard relational schema, these duplications will need to be documented either separately from the system or by creating an extraneous and complex "synonymy" table. This complexity differs markedly from the ease with which a "same-as" relationship can be added to a semantic network.

Semantic networks can also increase the separation between logical and physical components of information. This allows the system designer to explicitly delineate logical concepts and processes separately from the physical components with which they interact. Within the laboratory test domain used for this investigation, this capability is not yet fully exploited. It is easy to envision, however, the manner in which the current experimental system can be extended by representing the process by which certain laboratory tests are performed. For example, the sequence of steps used to type and cross-match a unit of blood may affect the risk of a transfusion reaction or the length of time before the unit is expired. Explicitly representing these steps within the semantic network can provide information about the blood unit that is separate from the information provided by the type and cross-match result.

The flexibility of a semantic network representation is crucial to capturing the variety and richness of native databases. The adaptability of the network structure enables the implementation of this representation over virtually any database structure. Compared to typical

[a] Duplicate data fields: trig = triglycerides, ap = alk phosphatase, abs neutrophil = abs polys, promyel = promyelo, and poly = neutrophil.

database schemas, a higher-level of fidelity to the granularity and conceptual structure of the data is possible. For example, hematology laboratories in large tertiary care hospitals may be subdivided into functional units that provide cell typing and microscopic examinations, clotting factor analysis, and functional tests on blood elements (e.g. platelet aggregation tests). Representing such a system is clearly a different task than representing the simple hematology laboratory of a community hospital, yet a semantic network easily accommodates both systems.

The ability to create "layers" of concepts within a network representation provides a natural framework for abstraction. On a practical level, concepts within the network can be grouped together for comparison purposes or to perform computation. Using one of the laboratory test networks in this investigation, it would be trivial to define an abstract concept named "expensive lab tests" and then assign data elements to be components of the new group. This abstract concept could then be used to calculate patient care costs and resource utilization.

As a user interface, a semantic network data model is easily understood and easy to navigate using a point-and-click interface such as the one implemented in MEDIATE. Relationships between concepts are clearly delineated, and it is easy to view the composition of aggregate concepts. The meta-information provided by the network representation is much richer than the typical relational database schemas. This allows users who are unfamiliar with the native database to quickly locate the data they seek. During informal presentations of the MEDIATE system to physicians, all the users easily comprehended the data model and successfully navigated to the data that they wished to view without any trouble.

The flexibility of a semantic network, however, needs to be constrained in order for meaningful comparisons to be made between different database representations. The relationships currently implemented within MEDIATE form an initial set of semantically useful relationships which allow adequate modeling of medical laboratory tests while limiting the possible network configurations enough to perform useful comparisons. The properties of the chosen relationships subsequently constrain the manner in which the semantic networks can be traversed, allowing repeated computations to be performed on networks representing different underlying native databases.

6.1.2 **Network Nodes and System Functionality**

Nodes within the MEDIATE semantic network function as more than just placeholders for concepts. The data structure of the node is designed to accomplish multiple purposes, including: 1) semantic identification, 2) facilitation of data interpretation, and 3) linkage of the concept with the underlying native database.

Semantic identification of the node concept is represented in several different ways. The basic semantic information about the node is contained within the "node name" and "node definition". The node name may sometimes be less useful, since it usually reflects the native database terminology and can be somewhat cryptic (as illustrated by the test names from the Hospital B database). The node definition, however, is a plain text message designed to enable an unambiguous description of the pertinent concept. This should be interpretable by any user.

"UMLS links" and "relationship links" embody the other ways in which a node contains semantic identification. By associating the concept with a standardized vocabulary through the UMLS links, terminology-associated semantic ambiguity is reduced, although it is not eliminated. It is the relationship links, however, that form the lynchpin of the representation system. The relationship associations with other concepts contain the crucial semantic information that allows the concept matching to take place.

It is not within the scope of this investigation to address the problem of interpreting the raw information contained in native databases. Nevertheless, some accommodation must be made in order to give MEDIATE practical functionality. The "format" data structure has been implemented to facilitate data interpretation by providing both semantic and syntactic information. As described previously, the two format parameters of "type" and "encoding" allow a basic explanation of how to interpret the data retrieved from the native database. Furthermore, a simple extension of the format data structure could be used to point to executable code that correctly displays or otherwise interprets the raw data. Although this feature is not currently implemented, the addition of this functionality is straightforward.

The "database link" data structure plays an important role in increasing the practicality of this representation system. By implementing a direct hook to the underlying database system, the network node creates the essential bridge between the semantic network representation and the raw data. Without this bridge, the network representation would merely be an interesting view of the data and would not facilitate data retrieval to nearly the same extent.

All of these functions (semantic identification, data interpretation, and database linkage) are tied directly into the node structure to create an encompassing "container" for the medical concept. Similar to the "self-describing objects" within the TSIMMIS system, the MEDIATE network node is self-contained and requires no other data structures to fully describe the concept that it encapsulates. [69, 70]

6.1.3 **Network Relationships and Inferences**

The current relationships implemented within MEDIATE support a flexible and descriptive set of network configurations. Although the relationships are far from all-inclusive, they are rich enough to support the representation of all the medical laboratory test concepts encountered in this investigation.

The relationships supported within this system differ from the stereotypical "isa" relationship by offering more semantic variety in the association between two concepts. As explained previously, the semantics for each implemented relationship is unique. Inheritance of attributes, for example, is only associated with the specialization (subclass/superclass) relationship, which is the direct analogue to the "isa" relationship. The limitations of the "isa" relationship are evident when trying to relate any two concepts that should logically be related by the composition (composed-of/component-of) relationship. To say that a serum sodium "isa" Chem 7 is clearly unreasonable, and serum sodium probably should not inherit all the attributes of Chem 7 as the "isa" relationship would mandate.

Support for computation and inference depends upon the nature of the semantic network links. In MEDIATE, the relationships support generalization and decomposition in a relatively straightforward manner, and these inferences are used in the concept matching algorithms.

Strictly speaking, generalization involves traversal of the "subclass-of" links up the hierarchy. From a functional viewpoint, however, climbing up the network using any kind of hierarchical relationship is a form of generalization (e.g. using "component-of" or "subset-of" relationships). The concept matching algorithms subscribe to this functional viewpoint and utilize all the hierarchical relationships when generalizing a concept for matching.

Similarly, strict decomposition should only utilize the "composed-of" relationship to descend the network hierarchy, but the matching algorithms actually use all the hierarchical relationships (e.g. "collection-of" and "superclass-of") to decompose concepts.

The rationale for using the broader forms of generalization and decomposition grows out of uncertainty about network configurations. Although the relationships themselves have clear semantics, the association between two concepts may include elements of several different relationships. Thus, "electrolytes" could correctly be related to "blood chemistries" through the "subset-of", "subclass-of", and "component-of" relationships.

There is no practical way of forcing users to choose a given relationship if they are all applicable, and instantiating all the possible relationships is somewhat redundant even if it is technically correct. These relationship overlaps produce an "intrinsic" form of semantic ambiguity in which multiple "correct" network configurations are possible for the exact same concepts. Because of this uncertainty, broader forms of inferences that utilize network traversal may be more practically useful than the strictly correct inferences. This was the motivation for utilizing all the hierarchical relationships during generalization and decomposition within the matching algorithms.

Inferences that are supported by the relationship links depend not only upon the semantics of the relationship, but also upon some of the basic properties of the relationship (as outlined previously in Table 1). The most important of these properties is transitive closure, which supports unidirectional traversal across the network using the pertinent relationship. Transitive closure and hierarchy are the properties that support the inferences of generalization and

decomposition. Other inferences are possible based upon other properties, although they are not currently utilized within MEDIATE. For example, the transitive closure and dependency properties could be used to generate a list of concepts that must be examined for a change in their semantics when a concept is deleted from the system.

The functional distinction between relationships blurs a bit when the considering the differences between the specialization (subclass/superclass) and set (subset/superset) relationships. On the surface, the semantic distinction is obvious. But from another perspective, set elements can be viewed as instantiated instances of classes. This corresponds to the "extensional" notion of a class, where the class is defined by the elements that are members of the class.[a]

Using this viewpoint, subsuming the set relationship within the specialization relationship may have little functional impact. Within the experimental setup of the current investigation, there was essentially no effect on the matching outcomes when set relationships were excluded from the network configurations. Of course, much more data is required before the utility of the set relationship can be addressed.

6.1.4 **Procedural Information and Inferences**

Within any knowledge representation system, inferences are performed not only by manipulating the data structure, but also by more general computational methods. Within MEDIATE, the concept matching algorithms and the quality metric calculations store procedural information that provides two forms of inter-network inferences, equivalence and subsumption.

The equivalence inference is a result of the direct matching process, where two concepts in different networks are inferred to be semantically equivalent if they are produced as the output of a match. This inference creates the foundation for automating the data exchange process between heterogeneous databases.

The subsumption inference is a product of the generalized matching process, where a target concept in one network is subsumed within the hierarchy of a higher-level concept in another

network. In this particular process, the subsumption inference can itself be decomposed into a generalization inference followed by an equivalence inference. Overall, the subsumption inference adds utility to the data exchange process by finding alternative concepts that may encompass the target concept.

Leaf matching provides a complementary pathway for data retrieval by utilizing the decomposition and equivalence inferences. By decomposing an aggregate node into its constituent concepts and finding the equivalents for those concepts, the leaf match retrieves information that is different from either direct or generalized concept matching. Viewing the matching computations through the perspective of the inference processes helps delineate the differences between the types of matches that are performed.

By modifying the basic inference processes, slightly different results may be obtained. For example, if the decomposition process were modified to stop after only one level of decomposition (rather than continuing until the leaves of the network are reached), the "leaf match" would become a "decomposition match" that may retrieve different information from the remote database.

In order to measure the variations produced by changes in the inference processes, a metric must be used. The match quality metric currently implemented within MEDIATE essentially measures the set "coverage" or overlap between two concepts. The quality metric functions as a proxy for the degree of semantic equivalence between two concepts, since there is no direct measurement available. Similarly, in the case of a leaf match, the quality score measures the set coverage for the target concept itself. In this setting, the quality score functions as a proxy for information content, or the "amount" of information that is available for a given concept.

Unfortunately, there is no way to capture "clinical relevance" directly within a metric, since clinical relevance is a subjective judgment that varies depending upon the circumstances and motivation of the user. For generalized concept matches in particular, the current quality metric

[a] The "intensional" definition of a class is given by defining parameters of the class which must then hold true for instantiated instances of that class.

is fairly useless, and the clinical relevance of the match depends entirely upon the nature of the data the user is seeking.

As the central computational mechanism for MEDIATE, the success of the equivalence inference process drives the utility of the system. There are many ways in which this inference may fail, from both computational as well as semantic standpoints.

Semantically, the most obvious way in which the equivalence inference fails is if a concept is absent from the universe under consideration. There were many examples of this phenomenon within this investigation, such as the concept "newa" which was present in Hospital B's database but not present in Hospital A's database.

Yet the issue of semantic absence is not black and white, but instead exists on a continuous scale. There is a gray zone where it is difficult to discern whether a concept is present or absent. For example, Hospital B's database contains the concepts "cbc", "cbca", and "cbcd", and Hospital A only has the concept "CBC".[a] Does this indicate that the concepts "cbca" and "cbcd" are absent from the universe of concepts in Hospital A? It is true that compared to the concept "CBC", the match quality scores indicate that the Hospital A concept "Hematology" is actually a better match for "cbca" and "cbcd". However, the semantics conveyed by the names "cbca" and "cbcd" seem to indicate at least some degree of equivalence with "CBC".

The current matching algorithms in MEDIATE support a liberal equivalence inference process. The algorithms err on the side of producing an equivalence match even if there is very little similarity between the concepts (as measured by the quality score). Additionally, the subsumption inference is also very liberal. The end result is that unmatched nodes represent concepts that are not only absent from the remote network, but also disconnected from other nodes within the local network, or connected only through the attribute relationship.

[a] The abbreviation "cbc" stands for "complete blood count", which implies that "cbca" and "cbcd" are variations of a complete blood count.

From the computational standpoint, the equivalence inference may be affected by many factors, including: network configuration, UMLS links, search algorithms, and the matching algorithms. Failure to find existing semantically equivalent concepts can arise from problems within any of these areas.

Network configurations have an obvious effect since they provide the semantic "context" which is used for the equivalence inference. Since there is no guarantee of how a user will configure a given network, concepts may be assigned relationships in a way that reduces the effectiveness of the matching process. In the worst case, a node may fail to be connected to the network (as in the case of the "Hemogram" node for Hospital A).

UMLS links are an important source of semantic information about the concept, and certain types of link assignments may hinder matching. The specificity of Metathesaurus terms may be problematic in some cases. For example, if a concept for "serum sodium" is linked only to a Metathesaurus term that indicates a specific technique for measuring sodium, it will not match other "serum sodium" concepts that do not include that technique. In a similar fashion, the number of Metathesaurus terms used in the UMLS link affects matching. For concepts with a large number of potential links to the Metathesaurus, instantiating fewer links will decrease the possibility of matching another concept with the same pool of potential links.

MEDIATE addresses these problems with UMLS links in an ad hoc fashion by enabling users to link all applicable Metathesaurus terms to a concept, from the specific to the general. This blurs the semantic distinctions between Metathesaurus terms and creates a "possibility set" of associations, which allows a more flexible matching process.

A more elegant solution, however, would be to create a semantic network of the Metathesaurus terms and allow MEDIATE to apply the computational machinery which already exists. Thus, if equivalence inferences did not find a match for a concept linked to a very specific Metathesaurus term, generalization inferences could be used to match concepts linked to more general Metathesaurus terms. Although this technique was not explored in the scope of this current

investigation, it would certainly be worthwhile to implement in future investigations of this system.

At the heart of the equivalence inferences lay the search and matching algorithms, which perform the requisite computations. Changes in these algorithms will certainly affect the outcome of the matching process.

Searching by BFS is uncomplicated, but the parameters used for terminating the search can affect the search outcome. Currently, a simple limit on search distance is enforced. Changing these limits affects the number of nodes searched and consequently affects the number of nodes that are considered as potential matches for the target node. All of the BFS searches in MEDIATE are currently limited to a single link traversal before terminating. This may not be appropriate if there are large differences in the size of the networks to be matched, especially if the size differences are reflective of differences in granularity of the concept representations.

The current matching algorithms implemented within MEDIATE are certainly not exhaustive, although they represent a studied attempt to exploit the semantic network linkages and patterns. In addition, there is some overlap between the algorithms, which may not be an optimal way to explore the entire solution space. Clearly, expanding the number of matching algorithms or modifying the current algorithms affects the results of the equivalence inference. Proving the correctness of the outputs, however, is still an open question.

In the end, the correctness of the equivalence inference depends upon some measure of what it means for two concepts to be semantically equivalent. As stated previously, the quality metric attempts to capture semantic equivalence by utilizing a form of set coverage, but ultimately only acts as a proxy for the subjective judgment of clinical relevance.

The quality metric can be used in conjunction with the matching algorithms to help choose the "best match", or the candidate node most likely to be semantically equivalent to the target node. In order to allow the user to perform the ultimate judgment of semantic equivalence, however, this technique is not automatically implemented in MEDIATE.

6.1.5 **Context Representation**

One of the advantages of a semantic network representation arises from the natural association between network neighborhoods and concept context. By definition, the nodes surrounding a target concept are related to that concept. Nodes that are more than one link distance away from the target concept are also related in either a direct way (if the relationships support transitive closure) or an indirect way. Of course, the strength of the relationship falls off as some function of the distance from the target node.

These neighboring nodes create a semantic context grounded in the relationship links and in the nodes themselves. This context contains information that facilitates the semantic interpretation of a given node. An example from the semantic network construction phase of this investigation illustrates the power of these contexts.

During the creation of the semantic network for Hospital B, many of the test abbreviations were so terse that they were not interpretable. This caused a problem during creation of the UMLS links, since the target concept needs to be clearly identified to create the link. For many nodes, however, the identification problem was resolved when relationship links that were specified in the native database were instantiated. These links were composition links that specified test panels composed of other tests. The node “bg”, for example, was an unknown entity until its component nodes were instantiated. Once the relationship to nodes “pH”, “pCO2”, and “pO2” was established, it became obvious that node “bg” represents the concept of a “blood gas” test.

The neighboring nodes also create a topological context that translates naturally into a graphical user interface. As stated previously, the graphical network interface was easily understood by many users, and enabled the location of desired data without any problems.

Capturing the semantic context in the network representation not only facilitates interpretation of the data, but also supports the inference process that forms the basis of concept matching. Conceptually, the matching algorithms boil down to methods of matching node contexts. Thus, the search algorithms explore the neighborhood nodes in various ways, and the matching

algorithms attempt to recognize equivalence by finding similarities between the patterns of the nodes explored.

Although MEDIATE provides the framework to construct representations that reflect the underlying database structure, strict faithfulness to the logical structure of the native database may not be the most informative representation scheme. Generally, the most complex networks will also be the most informative in terms of providing semantic context for searching and matching. Therefore, network representations that mirror very simple database structures may contain less semantic context and provide a poorer substrate for the matching algorithms.

The most extreme example of this phenomenon is an entire database consisting purely of attribute-value pairs, i.e. one table column with labels and another column with values. If the network representation only reflects this database structure, the network would assume a completely flat topology with no relationships between the concepts. This representation clearly provides no semantic context and blocks the concept matching algorithms. Figure 10 illustrates this problem.

In this situation, the user can improve the context representation by superimposing semantic structure based on common clinical usage of the concepts. This will likely occur to some extent for all database representations because of semantic overloading and the implicit nature of relationships within database schema. Many databases, however, will have more logical structure than simple attribute-value columns, and will provide a richer substrate for context representation.

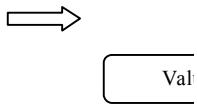
6.1.6 **Semantic Representation Summary**

Within the framework of the functional goals established for MEDIATE, the representation system supports both human interpretation and machine computation of concept semantics.

From a human viewpoint, the semantic information contained within the network allows accurate recognition of the concept embodied within a node, and the graphical representation of the network enables facile navigation to locate desired data.

Blood Culture Result	Positive
Blood Culture Organism	S. aureus
Blood Culture Antibiotic Sensitivity	Vancomycin
Sodium Level Fluid	Urine
Sodium Level Result	45
Sodium Level Units	meq
Sodium Level Volume	2 liters

Tables ***Se***



Analyte	**Fluid**	**Result**	**Units**	**Volume**
Sodium	Urine	45	meq	2 liters

Source	**Result**	**Organism**	**Abx Sensitivity**
Blood	Positive	S. aureus	Vancomycin

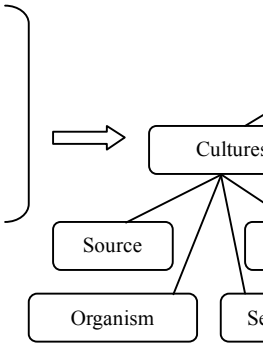


Figure 10. Table structure and network structure. If the semantic network
attribute-value pairs will induce a simple network with minimal concept diffe
complex structure (shown in the lower half of the figure) induce networks with

From the computation viewpoint, the representation enables the execution of several inference methods that support the goals of automatic data exchange. The most important of these inferences are the equivalence and subsumption inferences between networks. In addition, the calculation of a match quality metric allows comparison of various algorithms used to carry out these inferences, and also allows a rough evaluation of concept equivalence.

Other forms of semantic information can be computationally derived and represented without changing the topology of the network. For example, the extensional definitions from Zollo and Huff can be used as another form of semantic identification. This type of information could easily be added to the network node as another parameter for the equivalence inference.

6.2 Engineering Considerations

To incorporate MEDIATE into a real world production system for data exchange, many engineering issues need to be addressed. The following sections examine some of these issues, although this investigation did not attempt to implement all the processes required to support a production system.

6.2.1 Supporting Environment

A general infrastructure to support data exchange will require at least the following elements: communication protocols, data interpretation/decoding, and security measures.

Thankfully, basic communication protocols for medical information are mature enough that there is no need to create new processes for MEDIATE. In particular, the HL7 protocol is now widely implemented and supported on many platforms and medical information systems. This protocol is suitable for communicating both the semantic network representation and medical data between systems that are MEDIATE-enabled.

The XML extensions to HL7 are even more suitable as an underlying communication protocol, since all the data structures within MEDIATE have already been encoded in an XML-like syntax. It would take trivial modifications to make the system fully XML compliant.

At a more basic level, the hypertext transfer protocol (HTTP) has become a standard platform onto which other communications can be layered. There are many Internet based medical information systems that implement this basic communication layer, and the utility of using HTTP for data exchange has been proven repeatedly in the financial arena.

Interpreting or decoding data requires a combination of tools within both MEDIATE and the supporting information system. One of the benefits of the representation system within MEDIATE is that complex data can be decomposed into simple parts, which are then much simpler to interpret. Any laboratory test result, for example, is likely to have multiple pieces of associated information (e.g. result value, units, specimen number, etc.) that can be represented as

attributes. Since each attribute is an individual concept, the format data structure can be used to represent information about interpreting that concept.

Once concepts have been decomposed to the appropriate level within MEDIATE, standard tools can be used to decode the more elemental formats. Tools to convert and format text strings, convert between number systems, convert measurement units, and perform other simple transformations could easily be incorporated within MEDIATE. Tools to interpret binary objects such as images, sound files, video, etc. might be incorporated into the supporting information systems for transmission with the data. These tools can even be represented as concepts and included as attributes of other concepts within the semantic network.

Security and confidentiality are vital concerns for communication of any medical information, but the complexity of this topic is well beyond the scope of this investigation. It is worth noting, however, that system and human processes are much more important than any technical solutions in protecting the transmission of sensitive medical information. Therefore, the use of access control and encryption technology or protocols such as Secure Socket Layer transmissions will be necessary, but not sufficient to protect data transmission. The bulk of the design work and implementation of security measures will need to be performed not just for MEDIATE, but for the underlying information system as a whole.

6.2.2 **Performance Issues**

Speed of execution and space requirements is always a consideration in production systems.

The matching algorithms within MEDIATE currently operate in $O(n^3)$ time, where n is the greatest number of nodes in either of the two semantic networks to be matched. This worst case scenario may occur because the iteration matching algorithms may traverse every node of a network during a match for each node in the other network, and may perform this procedure up to n times before the algorithm terminates. Although this geometric growth is not a theoretical computational barrier, it may have real world consequences depending on the size of the networks and the actual execution speed of the system.

Space requirements are $O(n^2)$ in the size of the semantic networks. In the current implementation, all of the data structures used to represent the semantic network and to perform the inferences are kept in memory. Depending on the available memory and size of the networks, there will be the usual trade-offs in space vs. performance if virtual memory is required.

Optimization of the algorithms could conceivably decrease the execution time, but it is not apparent that any procedures exist which will operate in less than $\Omega(n^2)$ time. Optimization seems most likely to occur in the constant factors of the algorithms. This optimization should certainly be explored for a production system, but only empirical testing can prove the utility of such an effort.

On an architectural level, processing of semantic nets using a dataflow model on massively-parallel computers has been explored by Bic. [82] Although this approach does not change the fundamental nature of the problem, it may offer an advantage in execution time given the right computing environment.

6.2.3 **Representation Construction**

Construction of the semantic network will almost certainly form the bulk of the work required to implement MEDIATE within an information system. A systematic approach to facilitate this process will pay dividends not only in saving time, effort, and money, but may also result in a more complete and useful network representation.

Improvements to the user interface will certainly ease the process of network construction. For example, adding the ability to choose which relationship links to display will allow the user to focus on relevant links and nodes. Other graphical interface modifications, such as selective collapse/expansion of links and 3-D display techniques, may also help the user to organize the network more easily. Improving random access to the network nodes (rather than having to navigate through the network) will decrease the time users require to check or modify existing nodes. The implementation of a grep-like pattern-matching utility would be a great improvement over the alphabetical drop-down menu that is currently implemented.

Programs to help perform batch-processing of network node creation greatly speed up the creation of a representation. As stated previously, an auxiliary program to perform this task was used for the Hospital B network in the experimental setup. This program cut the amount of time required to create the network by an order of magnitude compared to the Hospital A network. Further investigation of the nature of these programs may reveal a more general framework that can be applied to a variety of information systems. This would ameliorate some of the work required to build customized batch-processing programs for every database system.

To retrieve information from a database, the semantic network representation must be correctly hooked into the database. This requires two components: procedural knowledge about how the database system functions, and a database driver that can be called by other applications. In the current investigation, MEDIATE contains the procedural knowledge to interface with most relational databases. The database link component of a node contains data structures and algorithms to specify the elements of relational tables and generate SQL queries for data retrieval. Fortunately, most common relational databases also provide various drivers for use by external programs.

For other database systems, new procedural knowledge and interface drivers must be provided to enable MEDIATE to cooperate with the database. This type of functionality may be provided in an approximate form for general types of databases (e.g. hierarchical, flat file, CORBA-mediated, etc.), but is likely to require some customization to attain complete functionality and integration with the host database system. This work contributes to the overall task of creating the network representation, but does not need to scale with the number of concepts represented, i.e. it is a constant factor.

Another tool that would be useful in constructing the semantic network is a network validation program. Checking for cycles and unconnected nodes is relatively straightforward, but searching for possible logical inconsistencies is more complex and not easily specified. It may be unlikely, for example, that a node is related by the “component-of” relationship to two other concepts that are in the same specialization hierarchy, but it is not impossible. Thus, the implementation of logical filters is likely to be based on heuristics rather than strict constraints on the network.

6.2.4 **Usability Issues**

Like any system, the design of the user interface greatly affects the ability to effectively utilize the system. The graphical interface for semantic network construction was discussed previously, but another important problem is the display of data from multiple sources. This issue was not addressed within this investigation because only two databases were represented. The problem is obviously compounded as more databases are added. Although multi-database information display is outside the scope of this investigation, it is likely that easy access to the semantic network representation of the data will be an important part of the interface.

Another issue that affects usability of the system is the amount of network traffic that is generated by data retrieval requests. For MEDIATE, the amount of data that must be communicated is greater than for a simple data request because the semantic network representations must be transmitted beforehand. In the worst case, the semantic network might need to be transmitted before every data query. In reality, however, this is highly unlikely since changes to the structure of a database do not occur several times a day.

Version control of the semantic networks can reduce the network traffic. One method is to have each semantic network maintain a modification field indicating the date and time that it was last modified. Other systems would then check this field to determine whether the semantic network needs to be transmitted again for matching before a data request is processed. The tradeoff for decreasing the network load and access time is an increase in the local storage required for each system to keep track of concept matching that has been performed with other systems.

Finally, one of the most important usability issues is the degree of automation employed in equivalence matching and retrieving data. This issue requires much more empirical data to resolve than this investigation provides. The degree of automation, however, clearly affects the function of the system in the following areas: estimation of the best concept match, processing of leaf matches, and retrieval of data.

As stated previously, the match quality metric could be used to automatically select the “best” match based on the quality score, with the highest scoring match theoretically representing the

most semantically equivalent concept. Currently, MEDIATE displays all candidate node matches with their respective quality metrics. This allows the user to choose the most appropriate match based on human judgment. If, however, the highest scoring match were chosen the vast majority of the time, then automating this choice would save the user time and effort.

Processing of leaf matches is currently performed automatically for all nodes, and the user is given an option to view the leaf matches within the user interface. This process could be customized so that less automation is performed in order to speed up the system response time. For example, the system could defer performing a leaf match until a user specifically requests it. Or a leaf match could be performed if the concept match has a quality score below a certain threshold. The tradeoffs for these decisions are difficult to quantify without further empirical data.

The actual retrieval of data from the native database can be automated if certain decisions are made about the contents of the query. Specifically, the query can be constructed in a very broad fashion to retrieve the data for all candidate matching nodes as well as leaf node matches, or the query can be narrowed to a given match type based on criteria such as the quality score. These choices are currently determined manually by the user, which requires time and effort that could be eliminated through automation.

In the end, a system interface that allows the user to set the degree of automation for each function will allow the greatest flexibility and most likely result in optimal usability.

6.3 System Evaluation

The overall goal of MEDIATE is to facilitate data exchange across multiple heterogeneous databases by automatically identifying semantically equivalent concepts between those databases. The experimental setup detailed in this investigation is a proof-of-concept, and the results demonstrate that within the boundaries of this study, MEDIATE achieves the goal of automatic concept matching.

6.3.1 **Match Types**

Direct concept matching drives the equivalence inference for all nodes. Leaf nodes of the semantic network representations are often analogous to atomic data elements in the native databases. These match without problems through the UMLS links if the equivalent concept exists in the other network.

UMLS links, however, are not required for direct concept matching to succeed. In the majority of the matching runs (nine out of sixteen), UMLS links were instantiated only for the leaf nodes. In this circumstance, the direct non-UMLS node matches illustrate the capability of MEDIATE. The ability to automatically find clinically relevant matches between concepts that do not share a common terminology distinguishes MEDIATE from almost all other existing systems. At this time, the only other system known to demonstrate this ability is the previously mentioned "extensional definition" system of Zollo and Huff, which does not support relationships between concepts.

As reported in the results section, detailed examination of the direct matches indicates that all possible matches were found when equivalent concepts were present in each of the two networks. This level of automated performance is highly encouraging, even for a limited test system.

When the concept was absent from one of the networks, however, the matches were much more variable and open to question. Since the matching algorithms utilize concept context to find equivalence, any overlap between neighboring nodes may be construed as possible equivalence. For example, the Hospital B concept "bmall5" matched with Hospital A concepts "Chemistry", "Serum lytes", and "Chem 7", with the highest quality score for the "Chemistry" match. Even though the test panel "bmall5" does not exist in Hospital A's database, there is a large overlap between the semantic contexts for "bmall5" and "Chemistry". Therefore, the inference of semantic equivalence may be reasonable, depending upon the motivation or judgment of the user.

By performing the subsumption inference, generalized matches provide useful information in the following form. First, the generalized match allows automated retrieval of information that may encompass the target concept. Second, the generalized match narrows the search space for equivalent information to a set of concepts with a high probability of semantic similarity to the target concept. This allows the user to locate pertinent information more efficiently if automated retrieval of information does not satisfy the user's requirements.

The current implementation of MEDIATE performs generalized matching after all attempts at direct concept matching fail. This algorithm is biased towards finding a direct concept match even if there is minimal semantic overlap between the concepts that are matched. Generalized matches are secondary matches that only performed if the target concept is absent from the remote database. This approach completely separates the equivalence and subsumption inferences.

As demonstrated by the experimental results, the equivalence inference sometimes produces a concept match that is only marginally relevant. In this case, it might be useful to perform the subsumption inference on the same target node to cast a broader data retrieval net. The decision to perform a generalized match could be based on the quality score (e.g. by using a threshold), or generalized matches could be performed on all nodes.

Since MEDIATE does not currently have a good metric for the relevance of a generalized match, automating the data retrieval in a useful way for these matches does not seem feasible. As in other parts of this system, the user has the ability to choose from candidate generalized matches and exercise judgment about the clinical relevance of the matches.

Leaf matches provide a complementary source of information for aggregate nodes by executing the decomposition and equivalence inferences in succession. This is a reductionist form of semantic equivalence, where the aggregate concept is ignored in favor of its constituent parts. If the leaves of a semantic network are fully matched, this form of data retrieval works well and provides all of the information content inherent in the aggregate concept.

But even if a leaf match has fully matched leaves, the concept match may still be a preferable method of retrieving data, particularly if the concept is present in both networks. Test panels are often ordered to assess the function of a particular organ system or physiological process. When retrieving data from a remote hospital, the clinical intent is often to assess a particular organ system or physiological process, not merely to see the results of a few tests. Thus, retrieving the concept match from a remote hospital may be more true to the clinical intent than just retrieving the matching leaves.

Concept matches such as the one between Hospital A node "Liver Function Tests" (LFTs) and Hospital B node "hfp" illustrate this line of reasoning. The leaf match results for "LFTs" reveals that all three leaf nodes – SGPT, SGOT, and Bilirubin – are matched. It is thus possible to retrieve the entire informational content of "LFTs" through the leaf match. But the matching concept "hfp" has seven component nodes that include SGPT, SGOT, and total bili, thus providing all the information in the leaf match and more. In this case, the concept match obviously surpasses the leaf match as a method to retrieve data.

Even in the other direction, however, retrieving the concept match for "hfp" (which happens to be "LFTs") might be more clinically useful than retrieving the seven leaf components. This may be true because the three components of "LFTs" will always be ordered together in a compatible clinical scenario. On the other hand, the other four components of "hfp" might be ordered for completely different reasons at different times, and thus might represent extraneous and possibly confusing information.

The leaf match reveals its true utility when a concept is absent from the remote network. In this situation, the direct and generalized concept matches are open to interpretation and are variably relevant in terms of their ability to capture the information content desired by the user. The leaf match, however, directly retrieves as much of the information content as possible for the target concept. In addition, the quality score for a leaf match is precisely related to the amount of information retrievable.

MEDIATE does not automate the choice of concept vs. leaf matches for data retrieval. Until a better model can be created to capture the idea of clinical relevance, this choice has been left for the user.

6.3.2 Network Configuration Effects

The experimental setup for this investigation employed four different network configurations for each hospital in order to explore the effects that different relationships might have on concept matching (as discussed in section 4.2). Enumerating all the unique combinations of these network configurations produced sixteen matching runs. Through all sixteen of these matching runs, the experimental results displayed remarkably small variations.

Analysis of the matching algorithms reveals the reason for this strong consistency in the results. In order to accommodate variations in network representations, the matching algorithms treat all the hierarchical relationships in almost the same manner for generalization and decomposition inferences (previously discussed in section 6.1.3). Whenever network traversal is required as part of the algorithm, the traversal of one of the hierarchical relationship links implies that traversal of the other relationships will also be utilized. Thus, the consistency in the matching results actually springs from one of the design goals for the matching algorithms, although the extent to which this consistency was achieved was a little surprising.

The experimental results demonstrate that the chief benefit of the current approach to generalization and decomposition is robust matching behavior with respect to different network configurations. This makes the performance of the system less sensitive to the vagaries of representation construction that are sure to arise among disparate database systems.

Conversely, the insensitivity to network configurations might signify a drawback to the current matching algorithms. Since one of the primary goals of MEDIATE is appropriate representation of semantic context, delineating distinctions between relationships is an important way of differentiating concepts. If these relationship distinctions are important in the semantic representation but unimportant in the matching process, some of the system functionality is lost.

In addition, finely delineated relationship links only add complexity to representation system if the distinctions between them are not functionally relevant.

Exploring the balance between robust performance and fine-grained semantic representations requires further investigation. Accumulation of empirical data will help reveal the circumstances that may tip the balance one direction or the other.

6.3.3 **Clinical Use**

Because of the semantic modeling capabilities built into the system, MEDIATE can be used for more than just information exchange between databases. It can be also be used as a tool to organize information and create new concepts, as a navigational tool to browse databases, and as a form of documentation for the native database system.

Organizing information into customized structures allows users to view and manipulate data in new ways. An immunologist, for example, could create a sub-network that groups together tests for white blood cells from the hematology laboratory, tests for antibody levels from the chemistry lab, and functional stimulation tests from the immunology lab. This sub-network represents a certain view of the data that helps the immunologist to assess a patient's immune status.

Once a sub-network has been created, it is even possible to create a new concept to label the sub-network using the composition relationship. The aggregate concept "immune function panel" becomes a new semantic entity that precisely captures the information that the immunologist seeks. This new concept could remain a permanent part of the network representation, or it could just be used in a temporary fashion and eventually be discarded. The ease with which such new views of the data can be created, labeled, and used demonstrates one of the most powerful features of this representation system. In essence, this tool enables users to create new semantic concepts that embody the precise amount of information they wish to analyze.

These novel semantic concepts are useful for viewing data in both local and remote databases. Retrieval of information from the local database should return the desired elements without

complication. For data retrieval from remote databases, identifying direct semantic equivalence may prove difficult. Performing leaf matches for the novel concepts, however, will yield the maximal amount of available information from the remote databases.

Utilization of novel data views enables MEDIATE to fulfill the data collection role for multi-institutional research projects. By defining a "data collection" aggregate concept at the central analysis site, all the pertinent data elements can be collected from participating remote sites with minimal effort (assuming all the systems are MEDIATE-enabled). Additionally, research investigators can easily modify the data elements by simply changing the composition of the "data collection" concept.

Another example of the utility of this feature is the ability to perform public health surveillance. A panel of pertinent test results could be aggregated as a "weekly surveillance" concept that is used to retrieve information from multiple institutions. If an event such as a disease outbreak occurs, the panel may be modified or a new panel created to retrieve additional pertinent data elements. The fact that representation modifications only need to occur at the data collection site showcases the simplicity and efficiency of this system.

As noted previously, the semantic network representation is useful not only as a means of data collection, but also as a means of data navigation. If all attempts at concept matching fail to meet the user's needs, or if the user merely wishes to explore some portion of the database, the semantic network provides an organized and facile way to search for pertinent data.

The navigational aspects of the semantic network reveal another feature of this system that may be utilized: the representation serves as another form of documentation for the underlying native database. The data structures within MEDIATE enable the documentation of concept location in the database (database link), associated concepts (attributes), concept interpretation (format), and concept relationships. The additional semantic information in the form of concept definitions and UMLS links may also prove useful at times for disambiguating a concept from other similar concepts.

This type of database documentation may also help prevent problems with duplicate data elements, as encountered in the Hospital B database. Many data elements are differentiated in relational tables being a unique key value for a column domain. These column domain elements may not be as visible or easy to find, and thus are easier to duplicate by mistake. In contrast, the semantic network concepts are structural elements of the representation that clearly define their place in the concept space. This is true even if the database link connects to a data element that is a column domain value.

Although data retrieval and exchange provide the primary functionality in MEDIATE, the other features that have been discussed enhance the clinical utility of this system.

6.3.4 **Summary**

The experimental results show that when a concept is present in both networks, MEDIATE always finds the match between the networks. Furthermore, MEDIATE goes beyond simple terminology matching by discovering matches between concepts based on their semantic context. These matches may differ from matches based solely on UMLS links, and often offer more information.

When a concept is absent from one network, alternative matches are found which may prove useful. Generalized matches provide some of the same functionality found in the Chu's Cobase system by utilizing a subsumption inference to encompass the target concept. Leaf matches provide a complementary method of retrieving data based on information content rather than semantic equivalency. At this time, the final choice of match retrieval remains with the user due to the absence of a suitable clinical relevance metric.

The databases used in this experiment were truly disparate, both in size and in concepts. Given this heterogeneity, the experimental results provide good evidence that MEDIATE achieves its primary goal of automated concept matching.

6.4 Experimental and System Limitations

Conclusions about the real world utility of MEDIATE are limited by drawbacks in the experimental setup. In addition, there are system limitations which were outside the scope of this investigation, but still worthwhile to address in anticipation of future work.

6.4.1 Single User Construction of Experimental Model

Construction of semantic network representations is open to influence by the user's perspectives and goals. Thus, having a single user (the investigator) construct the network representations for both of the experimental databases creates bias towards increased similarity between the databases. The aggregate nodes and levels of hierarchy (not already specified by the native database structure) are likely to exhibit increased resemblance compared to networks constructed by different users. This increases the probability that nodes will have similar semantic contexts, and consequently the probability that the matching process will successfully identify the match.

On the other hand, many elements of the semantic network representation for a given database would be similar even if different users were asked to construct the network. This similarity occurs because all the leaf nodes and predefined aggregate concepts (e.g. test panels) will be the same no matter who performs the construction. The main degrees of freedom are in the more abstract concepts or higher-levels of the network hierarchy. Although the abstract concepts are likely to exhibit more variety, the restricted knowledge domain helps to increase the likelihood that some similarity exists.

In addition to the network structure, the UMLS links may also be biased towards similarity because they were instantiated by a single user. This directly affects the matching process for leaf nodes, which depend almost entirely on UMLS matching for the equivalency inference. The net effect again biases the system towards increased matching and better experimental results.

A more insidious problem than representation bias is the possibility that the matching algorithms have been over-fit to the experimental model. Like most new systems, MEDIATE progressed through repeated iterations of the develop-test-evaluate loop. All the iterations of this

development loop, however, were based on the semantic network representations created for the two test databases from Hospital A and Hospital B.

Modifications to the matching algorithms were made to improve performance after analysis of the results from each test phase. The theoretical intent of these modifications, of course, was to correct "logical errors" in the matching algorithms. But no clear way of separating the errors from the experimental framework exists. Given the dependency between the experimental model and testing, the possibility arises that over-fitting of the matching algorithms may impede generalization of the techniques to other databases and network representations. Or, even if the techniques can be used, the performance may degrade in other environments.

Assessing the effects of single user representation construction and possible over-fitting of the experimental model requires further experiments with different users constructing representations for other databases. This expansion of the experimental model is a normal phase in the evolution of a new system, and the data derived from these experiments are crucial for further development of the system.

6.4.2 **Insufficient Sample Size**

As indicated previously in section 4.1.3, other databases were also considered for this investigation, but were eventually excluded. Therefore, any conclusions about the MEDIATE system must be tempered by the fact that only two databases were used in this investigation.

As in any empirical experiment, the samples used for data collection should reflect the characteristics of the general population towards which the experiment is targeted. Using only two laboratory databases increases the possibility that important characteristics of the general population of laboratory databases remain unexplored, which consequently skews the results. Although the two databases used were quite different in both size and content, it is unlikely that they capture the full range of concept and relationship possibilities. This raises the risk that the conclusions cannot be generalized to other laboratory databases.

6.4.3 **Restricted Medical Domain**

The ultimate goal of MEDIATE is to facilitate the exchange of all medical data, not just laboratory test results. The reasons to restrict the scope of this investigation were previously delineated in section 4. Given the restricted domain, however, generalizing the findings of this investigation to the broader medical database arena is not a feasible task.

On a theoretical basis, the structures required for representation of general medical information exist within MEDIATE. The UMLS Metathesaurus provides grounding for the atomic medical concepts, and the relationship links were designed to accommodate a broad range of concepts. But no method exists to predict in advance the performance of the system on general medical databases.

Assessment of MEDIATE's suitability for general medical information exchange awaits further experiments that utilize broader medical databases.

6.4.4 **Information Required for Representation Construction**

One of the real world limitations for constructing the semantic network limitations became evident during the course of this investigation. The type of database information and documentation needed may be difficult to acquire.

On casual inspection, the required parameters seems fairly straightforward: 1) a list of data elements which correspond to atomic elements in the native database, 2) the database call or routine to retrieve each of these data elements, 3) the database schema or general structure, and 4) a list of component data elements for each aggregate concept (if aggregate concepts exist).

In reality, each of the parameters listed above can be difficult to obtain if the database system does not have adequate documentation that is kept up to date. The list of data elements, for example, may be difficult to obtain from a relational database where the data element is a value in a column domain. Unless the values of the column domain are documented separately, the only way to ascertain the values present in the system is to perform a query on all the unique values of that column (as was done for Hospital A). The documentation of the database call for

each data element sometimes gets muddled over time, particularly if the underlying database structure has changed. In the case of one of the hospitals (which was not included in this study), the problem was compounded because the syntax for the database call was constructed on the basis of a hard-coded print form. All links between the syntax and the semantics of the data concepts were lost, and the documentation was scattered and not well maintained.

During the experimental setup for this investigation, problems were encountered in each of the four previously listed parameters. In the end, solutions were found for only two of the hospital databases in a time frame that allowed inclusion in this investigation. With sufficient time and motivation, however, all the requisite information could be obtained. In a production system, the problems with acquiring the needed information might delay the implementation of MEDIATE, but is unlikely to completely prevent it.

Once the requisite database information is available, the construction of the semantic network requires skills that include: 1) familiarity with the database and all four parameters of the required information, 2) sufficient computer proficiency to utilize the MEDIATE interface to construct the network, and 3) enough medical domain knowledge to instantiate the UMLS links and create the overall structure of the network.

Personnel in the information technology (IT) department are likely to possess the first two skills, and clinicians are the most likely to possess the third skill. If a single person who possesses all the skills cannot be found, then the task of creating the semantic network can be divided as follows. The IT department can create the network nodes corresponding to atomic data elements (leaves of the network) and instantiate the database link for each node. A computer savvy clinician can then instantiate the UMLS links and create the overall network structure with aggregate nodes and hierarchy that are appropriate for the institution.

6.4.5 **Attribute Relationship Representation**

During the course of this investigation, it became apparent that the attribute relationship lacked the crispness of definition that was present in the other relationships. Difficulties in precisely defining attributes are well known in the knowledge representation field. This imprecision

influences the decisions made about semantic network structure and increases the possibility of semantic ambiguity.

The working definition of an attribute states that concept A is an attribute of concept B if all the subclasses of concept B should inherit the attribute concept A. Semantically, concept A is some property of concept B that is so closely associated that all examples and variations of concept B should also have concept A associated with them. Functionally, relational database columns from a single table can be often be directly mapped into attributes for the concept that represents the entire table. For example, the columns of a laboratory results table form the attributes for the concept laboratory test.

The main difficulty with the attribute relationship arises in attempting to differentiate it from the composition relationship. The composition relationship also states that two concepts are tightly associated, so that the aggregate concept depends upon the component concept for its semantic value. When constructing a network representation, the differences between the attribute and composition relationships can blur.

For example, the concept “address” usually has sub-concepts “street address”, “city”, “state”, and “zip code”. Are these sub-concepts components or attributes? If the sub-concepts are the columns in a relational table titled “address”, the argument can be made that they are attributes that would be inherited by subclasses such as “home address” and “business address”. On the other hand, a home address could be “composed-of” all these sub-concepts, while a business address would be “composed-of” these sub-concepts plus “business name” and “department” sub-concepts. Neither representation seems more inherently correct than the other.

From the semantic context perspective, some attributes may not allow distinguishing between concepts. In the entire “laboratory test result” network, for example, all the concepts share common attributes of “test name”, “result value”, and “units”. Thus, attempting to distinguish between laboratory tests based on these attributes is an exercise in futility. This is the main reason that the attribute relationship is not included in the search algorithms used in the matching process for MEDIATE. Conversely, components usually create distinguishing semantic contexts.

Yet it is also clear that choosing between attribution vs. composition can sometimes be arbitrary and depend completely upon the user's judgment. This semantic ambiguity can create problems in the matching process, since network searching does not traverse the attribute links at this time.

After more experiments, the usefulness of including the attribute relationship in the matching process may become clear. Until that time, or until a more lucid definition of the attribute relationship is created, all semantic networks are at some risk for disparity in the choice between composition and attribute links.

6.4.6 **Concept Ordering and Cardinality**

Ordering of elements is a fundamental property of many types of data, but MEDIATE currently lacks a principled way of applying ordering to concepts within the semantic network. Different methods of implementing ordering include a new type of ordering relationship, or subtypes of the current relationships. The effects of the ordering scheme on the semantic context and matching of concepts must be considered before any implementation is included within the system.

6.4.7 **Relationship Composition**

Composing relationships implies the traversal of relationship links of different types across the network. Certain relationships have semantics that support composition, such as: Concept A is an element-of Concept B that is a subset-of Concept C => Concept A is an element-of Concept C. The effects of relationship composition are not fully explored in the current investigation, nor is the validity of the compositional relationships fully delineated. Many of the relationships are not commutative when composed, and only some compositions have logical consistency. Cohen explored the induction of plausible inferences from composing relationships, and found a correlation with relationship properties such as transitivity and inheritance. [83]

6.4.8 **Lack of Storage Model**

MEDIATE exists to facilitate the automated retrieval of information from remote medical databases. Once the information is received, however, the system does not address the issue of what to do with the information beyond displaying it.

Storage of information from multiple sources presents entirely new issues that are beyond the scope of this investigation. Whether the goal is local storage of information or creation of a data repository, many issues require thought beyond the consideration of semantic equivalence. The choice between a unifying data model vs. separate storage of information remains one of the primary issues.

Information storage could utilize MEDIATE interfaces to manage multiple information caches, but there are clearly different tradeoffs in terms of space, efficiency, and performance compared to the data retrieval problem.

Overall, storage of information from multiple sources is a large-scale problem that requires extensive investigation in its own right.

6.4.9 **UMLS Link Dependency**

The use of a standard terminology to "ground" the system is the closest that MEDIATE gets to utilizing a central data model. To some extent, this exposes the system to some of the weaknesses inherent in central data models. Namely, modifications to the Metathesaurus, absence of terms, and addition of new terms may all affect UMLS links within the semantic network.

On a practical level, however, MEDIATE depends on the UMLS links only for leaf concepts. These atomic concepts are much easier to associate with a standardized terminology, and no relationship or structural considerations involving other concepts interfere with this association. The main problem is semantic ambiguity within the Metathesaurus itself, which MEDIATE addresses by using a "possibility set" model for the UMLS link.

As discussed previously, the UMLS link provides one of the many forms of semantic information that MEDIATE incorporates. For leaf nodes, the UMLS link provides good functionality because the Metathesaurus terms associate with semantic network concepts with high degrees of semantic equivalence, and the link supports a simple computation for matching

concepts. For non-leaf nodes, the semantic context of the concept assumes greater importance, and the dependency upon the UMLS link does not have as great an influence.

6.4.10 **Lack of Clinical Relevance Metric**

Semantic equivalence in MEDIATE is based on structural similarities between the network representations, i.e. the semantic context. This is not, however, a direct proxy for "clinical relevance", which is much harder to quantify.

Very complex models would be required to capture user motivation, goals, and preferences. All of these parameters affect the manner in which a user judges clinical relevance. Unfortunately, the problem remains intractable at this time, and it seems unlikely that a rigorous metric can be developed in the near future.

6.4.11 **Lack of Process Modeling**

MEDIATE is intended to facilitate the exchange of database "content" and currently lacks the capability to represent medical processes or general medical reasoning. This narrow scope provides advantages in terms of system complexity and understandability but has obvious drawbacks in terms of general knowledge representation. However, an automated content exchange system such as MEDIATE could provide the foundation for other representation and inference systems with more ambitious goals.

6.4.12 **Functional Decentralization**

The decentralized architecture and computation of this system provide a benefit in terms of system scalability robustness. For certain functional processes, however, centralization of the computation offers advantages in efficacy and efficiency. For example, Zeng described a method of displaying different "views" of medical information (e.g. time-oriented, source-oriented, and concept-oriented) by using semantic networks to construct a central ontology. This ontology supports the inferences necessary to generate the different views from local databases. [84]

Although the functional goals enumerated for MEDIATE led to a decentralized system implementation, adding any new functionality will entail additional examination of the benefits and drawbacks of this distributed system. The current architecture does not preclude adding

centralized components, but the balance between centralized and distributed computation obviously requires careful consideration.

6.4.13 **Limitations summary**

This current investigation into the application of MEDIATE is a proof-of-concept, rather than a large-scale data collection experiment. As such, the limitations of the experimental setup clearly affect generalization of the results and conclusions. In addition, MEDIATE has inherent limitations that require further exploration before firm statements can be made about the performance and utility of the system.

The experimental limitations can be expected at this stage of development, and further investigation can expand the conclusions in a fairly straightforward manner. The inherent system limitations, however, require much more thought and consideration, and addressing some of the limitations may remain beyond the scope of implementing an effective production system.

6.5 **Future Direction**

Further investigation of MEDIATE will explore several different areas involving generalization of the system, consideration of current limitations, and extension of the system to add more functionality.

6.5.1 **Generalization to Full Medical Record**

The structure of an experiment to represent a full medical record can be very similar to the structure used in this investigation. Online medical records from different institutions will have semantic network representations created, and the performance of the matching process can be tested on these representations. The sample size will depend upon the availability of databases and the pertinent parameters for those databases (section 6.4.4), with the attendant correlation between sample size and the ability to generalize the results.

Modifications required within MEDIATE mainly involve revision of the UMLS link to include more of the Metathesaurus. Because the entire Metathesaurus might potentially be required, a local database version (provided on CD-ROM by the National Library of Medicine branch of the

National Institutes of Health) might need to be implemented. Utilizing the rudimentary synonym links within the Metathesaurus could also augment the functionality of the UMLS link.

6.5.2 **Generalization of Concept matching**

The core process of MEDIATE is execution of the equivalence inference based on the semantic context (neighboring nodes) of concepts represented within a semantic network. It may be possible to generalize the process to facilitate automated concept matching within other semantic networks.

In order to test the concept matching process in other systems, several issues require extensive consideration. These include:

1) *The nature of relationships between concepts within the system.* The relationship links currently implemented within MEDIATE may not apply in other systems, or other relationships may already be in widespread use. The characteristics of the relationships that might be utilized in the semantic network representations need analysis to see if they support the inferences required for concept matching.
2) *The availability of a starting point to "ground" the match process.* This necessitates some form of commonality or structure that can be exploited for matching of atomic concepts. In many cases, this may be a standard vocabulary or data model. Eventually, natural language processing may be efficacious enough to fill this role. The use of "possibility sets", as implemented in MEDIATE, can help ameliorate problems with semantic ambiguity in the grounding system.
3) *Customization of matching algorithms to achieve system goals.* Although the semantic context of a concept is easily understood to be the neighboring nodes in the network, the true goal of a match might include optimization for some particular relationship or local network structure. For example, in network representations of financial concepts, it might be desirable to maximize the matching of relationships relating to monetary flow.
4) *An appropriate metric of utility for the concept match.* A metric that truly captures the most important parameter of utility may be difficult to derive, as evident in the lack of a true clinical relevance metric for this investigation. Nevertheless, proxy parameters such as the MEDIATE quality score enable the objective evaluation of concept matches.

5) *The final objective of concept matching.* Within MEDIATE, the process of concept matching is merely a means to an end goal of automated data exchange. For other systems, concept matching may be the end goal, or it might be utilized for other purposes such as measuring the content similarity between two knowledge bases.

Some examples of areas where automated concept matching by semantic equivalence may prove useful are:

1) *Data exchange in other knowledge domains in which no standard data model exists, or for which the existing standards are inadequate.* The need for this functionality is evident in the proliferation of standards for data exchange, particularly in the financial and business arena.
2) *Integration or interchange of ontologies.* Since many ontologies are already expressed as semantic networks, this appears to be a natural area to apply the methods used in this investigation. However, conversion of an existing ontology to a form amenable for concept matching may prove to be a complex exercise, since many ontologies are heavily invested in the concepts and relationships used within their representations.
3) *Navigation of semantic nets, including the World Wide Web (Web).* As the Semantic Net efforts of the World Wide Web Consortium gain traction, more of the Web will have organized semantic content. The methods utilized within the investigation may be incorporated into automated systems that facilitate concept matching across different Web sites. These matching efforts could serve as the foundation for navigational directories that enable users to locate pertinent information.
4) *Search utilities.* Similar to the previous item, the location of pertinent information could utilize directories constructed by previous concept matching. Searching for system-wide information through real-time concept matching may be appropriate for limited systems, but will not scale to large systems such as the Web.

Further exploration of automated concept matching will undoubtedly reveal more issues and problems that require intensive investigation. But the potential for broader application of the methodology seems high, and the benefits may provide sufficient motivation to drive further research in these areas.

6.5.3 **Addressing Current Limitations**

The issues with the current attribute relationship deserve further investigation immediately. A better definition of the relationship will allow either elimination of the relationship, or a clearer picture of appropriate applications of the relationship.

Adding new relationships to enable ordering of concepts is an important objective that also assumes a fairly high priority. Examples of ordered lists that may require representation include patient problem lists, clinical practice guidelines, and historical lists of significant events and procedures. The nature of these ordered relationships, however, is complex, and their effect on the semantic context deserve close scrutiny.

Better metrics to measure either semantic equivalence or clinical relevance may be developed with further study. In particular, a metric for generalized matches would help objectively quantify the relative usefulness of different matches. In addition, a metric to measure the "match quality" of the entire network could be utilized to refine the concept matches on a network-wide basis. An iterative process to optimize this network metric may improve the overall performance of the system.

Other system limitations, such as relationship composition, the lack of a storage model, and system interactions with the Metathesaurus, require more long-term investigations to address. The cost-benefit ratio of tackling these limitations depends upon the ultimate use of MEDIATE. For data retrieval within a small network of systems, the current capabilities may suffice. More extensive implementations or attempts at data aggregation and storage may mandate further development of a storage architecture and better methods of utilizing the Metathesaurus or other concept grounding models.

6.5.4 **Augmenting System Capabilities**

One natural way to extend the capabilities of MEDIATE is to add additional relationships to create more complex and precise semantic representations. Temporal relationships in particular have a high priority because many concepts in medicine as well as other domains have a temporal component. Other types of relationships have already been implemented in systems

such as the UMLS Semantic Network, and these relationships may be further explored within MEDIATE. Anatomic, spatial, and process-related relationships all have the potential to provide increased functionality to the system.

Additional matching algorithms may be used to take advantage of richer network representations. Heuristic search algorithms might add efficiency and increased performance. And matching algorithms tuned to nuances of semantic context representation could provide better or more specific matches.

As the tools to create a rich semantic context become more elaborate, the risk for disparity in the network representations also increases. Therefore, methods to accommodate varying degrees of network diversity also need continued development. One approach might involve a "simplification" process, in which complex networks are reduced to simpler structures. The simplification process would facilitate comparisons between complex and simple networks. A corollary need arises for a metric to measure network complexity. Such a metric would support the ability to modify network structures in order to increase the comparability of the networks, which in turn supports better concept matching.

More elaborate computations with new functionality can be created to leverage new relationships. For example, processing of time sequences may be possible using temporal relationships, and following the natural history of a disease may be a feasible task. Even the task of automated or assisted disease diagnosis can be approached through concept matching of real world data with stereotypical disease concepts.

Two general themes emerge for expanding the capabilities of MEDIATE beyond simple data retrieval. One theme centers on the creation of ever richer representations and semantic contexts, allowing more accurate and complex descriptions of information. The second theme involves the creation of more powerful inference engines that are supported by the more informative representations. Both themes are well known in the knowledge representation field. But this system distinguishes itself by the central role of automated concept matching, which forms the basis for virtually any process that requires information comparison or exchange.

7 CONCLUSION

The goal of facilitating information exchange between heterogeneous databases can be approached in many ways. MEDIATE was designed to address the critical issue of identifying semantically equivalent concepts, a task that must always be performed at some level in order to correctly interpret information transmitted between disparate systems. The representation system and computational processes chosen for MEDIATE enable the equivalence inference to be performed in an automated fashion, and support the functional goals delineated at the start of this investigation. To reiterate, these goals include reducing the semantic ambiguity of transmitted data, representing the internal structure and granularity of native databases, and facilitating the retrieval of "useful" information even in the absence of direct correspondence between data concepts.

Although the limitations inherent in the experimental system must be kept in mind, the results obtained in this investigation support the assertion that MEDIATE achieves these goals. Automated matching of equivalent concepts from two different databases was accomplished, the representation system supported all levels of information granularity, and the implementations of generalized and leaf matches provided clinically relevant information for many concepts that would otherwise have produced null fields in a database query.

The system limitations of MEDIATE appear resolvable with further investigation and sufficient motivation. As in all real world systems, compromises and optimizing assumptions will inevitably be required. But for the declared goal of data exchange, this investigation did not uncover any insurmountable obstacles. Indeed, the results show promising performance characteristics given the disparity between the test databases.

Compared to other systems, MEDIATE offers potential benefits in the areas of scaling, robustness, efficient use of legacy databases, information navigation, documentation, and preservation of local semantics for each participating institution. Further testing will prove whether these benefits are realizable on a more ambitious level.

As an information platform, the fundamental mechanisms of MEDIATE provide a fertile environment for exploring new functionality in the areas of data sharing and information location and retrieval. With a sufficiently rich representation of semantic context, high level knowledge-based computation can also be supported. Future investigations of this system harbor great promise for contributions in the venture of information management.

8 REFERENCES

1. Beeler, G.W., Jr., *On the Rim: the making of HL7's Reference Information Model.* MD Comput, 1999. **16**(6): p. 27-9.
2. Russler, D.C., et al., *Influences of the Unified Service Action Model on the HL7 Reference Information Model.* Proc AMIA Symp, 1999: p. 930-4.
3. van Wingerde, F.J., et al., *Linking multiple heterogeneous data sources to practice guidelines.* Proc AMIA Symp, 1998: p. 391-5.
4. Levy, D., *Left ventricular hypertrophy. Epidemiological insights from the Framingham Heart Study.* Drugs, 1988. **35 Suppl 5**: p. 1-5.
5. Lloyd-Jones, D.M., *The risk of congestive heart failure: sobering lessons from the Framingham Heart Study.* Curr Cardiol Rep, 2001. **3**(3): p. 184-90.
6. Lloyd-Jones, D.M., et al., *Applicability of cholesterol-lowering primary prevention trials to a general population: the framingham heart study.* Arch Intern Med, 2001. **161**(7): p. 949-54.
7. Millen, B.E. and P.A. Quatromoni, *Nutritional research within the Framingham Heart Study.* J Nutr Health Aging, 2001. **5**(3): p. 139-43.
8. Murabito, J.M., *Women and cardiovascular disease: contributions from the Framingham Heart Study.* J Am Med Womens Assoc, 1995. **50**(2): p. 35-9, 55.
9. McDonald, C., *The Barriers to Electronic Medical Record Systems and How to Overcome Them.* J Am Med Inform Assoc, 1997. **4**(3): p. 213 - 221.
10. Dashti, A.E., et al., *Database challenges and solutions in neuroscientific applications.* Neuroimage, 1997. **5**(2): p. 97-115.
11. Bright, M.W., A.R. Hurson, and S. Pakzad, *Automated Resolution of Semantic Heterogeneity in Multidatabases.* ACM Transactions on Database Systems, 1994. **19**(2): p. 212-253.
12. Worboys, M.F. and S.M. Deen, *Semantic Heterogeneity in Distributed Geographic Databases.* Sigmod Record, 1991. **20**(4): p. 30 - 34.
13. Liu, Z., X. Du, and N. Ishii. *Integrating databases in Internet*. in *Second International Conference on Knowledge-Based Intelligent Electronic Systems*. 1998.
14. Kim, W., et al., *On Resolving Schematic Heterogeneity in Multidatabase Systems.* Journal of Parallel & Distributed Databases, 1993. **1**(3): p. 251-279.
15. Bressan, S. and C. Goh, *Semantic Integration of Disparate Information Sources over the Internet Using Constraints.* 1997.
16. Mori, A.R. and F. Consorti, *Integration of clinical information across patient records: a comparison of mechanisms used to enforce semantic coherence.* IEEE Trans Inf Technol Biomed, 1998. **2**(4): p. 243-53.
17. Peckham, J. and F. Maryanski, *Semantic Data Models.* ACM Computing Surveys, 1988. **20**(3): p. 153 - 189.
18. Sciore, E., M. Siegel, and A. Rosenthal, *Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems.* ACM Transactions on Database Systems, 1994. **19**(2): p. 254-290.
19. Bergamaschi, S. and C. Sartori, *On Taxonomic Reasoning in Conceptual Design.* ACM Transaction on Database Systems, 1992. **17**(3): p. 385 - 422.

20. Feng, L., H. Lu, and A. Wong. *Integrating legacy systems towards intelligent enterprise computing*. in *IEEE International Conference on Systems, Man, and Cybernetics*. 1999: IEEE.
21. Goh, C., et al., *Context Mediation: New Features and Formalisms for the Intelligent Integration of Information.* Sloan Working Paper 3941, 1997.
22. Leong, T.-Y. *Representing Context-Sensitive Knowledge in a Network Formalism: A Preliminary Report*. in *Uncertainty in Artificial Intelligence*. 1992: Morgan Kaufmann.
23. Palopoli, L., et al. *A unified graph-based framework for deriving nominal interscheme properties, type conflicts and object cluster similarities*. in *International Conference on Cooperative Information Systems*. 1999: IEEE.
24. Palopoli, L., D. Sacca, and D. Ursino. *Semi-automatic, semantic discovery of properties from database schemes*. in *International Database Engineering and Applications Symposium*. 1998.
25. Rector, A.L., et al., *The GRAIL concept modelling language for medical terminology.* Artif Intell Med, 1997. **9**(2): p. 139-71.
26. Rector, A.L., et al., *Medical-concept models and medical records: an approach based on GALEN and PEN&PAD.* J Am Med Inform Assoc, 1995. **2**(1): p. 19-35.
27. Tange, H.J., et al., *The granularity of medical narratives and its effect on the speed and completeness of information retrieval.* J Am Med Inform Assoc, 1998. **5**(6): p. 571-82.
28. Chu, W., M. Merzbacher, and L. Berkovich. *The Design and Implementation of CoBase.* in *Proceedings of ACM SIGMOD*. 1993. Washington, D.C.
29. Chu, W., Q. Chen, and M. Merzbacher, *CoBase: A Cooperative Database System.*, in *Non-Standard Queries and Answers*, R. Demolombe and T. Imielinski, Editors. 1994.
30. Chu, W., et al., *CoBase: A Scalable and Extensible Cooperative Information System.* Journal of Intelligent Information Systems, 1996. **6**.
31. Rector, A.L., *Thesauri and formal classifications: terminologies for people and machines.* Methods Inf Med, 1998. **37**(4-5): p. 501-9.
32. Barrows Jr, R. and S. Johnson, *A data model that captures clinical reasoning about patient problems.* Proc Annu Symp Comput Appl Med Care, 1995: p. 402 - 405.
33. Dore, L., et al., *An object oriented computer-based patient record reference model.* Proc Annu Symp Comput Appl Med Care, 1995: p. 377-81.
34. Friedman, C., et al., *The Canon Group's Effort: Working Toward a Merged Model.* Journal of the American Medical Informatics Association, 1995. **2**(1): p. 4 - 18.
35. Gouveia-Oliveira, A. and L. Lopes, *Formal representation of a conceptual data model for the patient-based medical record.* Proc Annu Symp Comput Appl Med Care, 1993: p. 466-70.
36. Pollard, D. and J. Hales, *Evaluation of an object-based data model implemented over a proprietary, legacy data model.* Proc Annu Symp Comput Appl Med Care, 1995: p. 367 - 371.
37. Burdis, C., B. Eaglestone, and P. Procter, *A unified model to support an information intensive health care environment.* Stud Health Technol Inform, 1999. **68**: p. 171-4.
38. Canfield, K., M. Silva, and K. Petrucci, *The standard data model approach to patient record transfer.* Proc Annu Symp Comput Appl Med Care, 1994: p. 478-82.
39. Campbell, J.R., et al., *Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures.* J Am Med Inform Assoc, 1997. **4**(3): p. 238-51.

40. Kohane, I., et al., *Sharing electronic medical records across multiple heterogeneous and competing institutions.* Proc AMIA Fall Symp, 1996: p. 608 - 612.
41. Dolin, R.H., et al., *HL7 document patient record architecture: an XML document architecture based on a shared information model.* Proc AMIA Symp, 1999: p. 52-6.
42. Shakir, A.M., *HL7 Reference Information Model. More robust and stable standards.* Healthc Inform, 1997. **14**(7): p. 68.
43. HL7.http://www.hl7.org.
44. Kirby, J. and A.L. Rector, *The PEN&PAD data entry system: from prototype to practical system.* Proc AMIA Annu Fall Symp, 1996: p. 709-13.
45. Pole, P.M. and A.L. Rector, *Mapping the GALEN CORE model to SNOMED-III: initial experiments.* Proc AMIA Annu Fall Symp, 1996: p. 100-4.
46. Rector, A.L. and W.A. Nowlan, *The GALEN project.* Comput Methods Programs Biomed, 1994. **45**(1-2): p. 75-8.
47. Rector, A., et al., *Medical-concept Models and Medical Records: an Approach Based on GALEN and PEN&PAD.* Journal of the American Medical Informatics Association, 1995. **2**(1): p. 19 - 35.
48. Rector, A.L., et al., *A Terminology Server for medical language and medical information systems.* Methods Inf Med, 1995. **34**(1-2): p. 147-57.
49. Rector, A., et al., *Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation.* Int J Med Inf, 1998. **48**(1-3): p. 71-84.
50. Rector, A.L., et al., *Reconciling users' needs and formal requirements: issues in developing a reusable ontology for medicine.* IEEE Trans Inf Technol Biomed, 1998. **2**(4): p. 229-42.
51. Rogers, J.E., et al., *Validating clinical terminology structures: integration and cross-validation of Read Thesaurus and GALEN.* Proc AMIA Symp, 1998: p. 845-9.
52. Rogers, J. and A. Rector, *GALEN's model of parts and wholes: experience and comparisons.* Proc AMIA Symp, 2000: p. 714-8.
53. Solomon, W.D., et al., *Having our cake and eating it too: how the GALEN Intermediate Representation reconciles internal complexity with users' requirements for appropriateness and simplicity.* Proc AMIA Symp, 2000: p. 819-23.
54. Hardiker, N.R. and A.L. Rector, *Modeling nursing terminology using the GRAIL representation language.* J Am Med Inform Assoc, 1998. **5**(1): p. 120-8.
55. Rogers, J.E., et al., *Rubrics to dissections to GRAIL to classifications.* Stud Health Technol Inform, 1997. **43 Pt A**: p. 241-5.
56. Solomon, W.D., et al., *A reference terminology for drugs.* Proc AMIA Symp, 1999: p. 152-6.
57. Boye, N. and N.E. Veirum, *Ontology-based, medical domain-specific, use-case driven EMRs for use in clinical quality assurance and passive decision support.* Stud Health Technol Inform, 2000. **70**: p. 36-8.
58. Grimson, W. and P.A. Sottile, *Synapses in the context of healthcare information systems.* Stud Health Technol Inform, 1997. **45**: p. 30-9.
59. Grimson, W., et al., *Federated healthcare record server--the Synapses paradigm.* Int J Med Inf, 1998. **52**(1-3): p. 3-27.
60. Heimbigner, D. and D. McLeod, *A Federated Architecture for Information Management Systems.* ACM Trans Office Info Systems, 1985. **3**(3).

61. Hurlen, P. and K. Skifjeld, *Design and functional specification of the Synapses federated healthcare record server. Synapses Consortium.* Stud Health Technol Inform, 1997. **43 Pt A**: p. 334-8.
62. Mostardi, T. and C. Siciliano. *An overview of WIND (Wide Interoperable Networked Databases)*. in *Twenty-Seventh Hawaii International Conference on System Sciences*. 1994.
63. Sheth, A. and J. Larson, *Federated Database Systems for Managing Distributed Heterogeneous and Autonomous DataBases.* ACM Computing Surveys, 1990. **22**(3): p. 183 - 236.
64. Gligor, V. and G. Luckengaugh, *Interconnecting Heterogeneous Database Management Systems.* IEEE Computer, 1984: p. 33 - 43.
65. Wiederhold, G., et al. *KSYS: An Architecture for Integrating Databases and Knowledge Bases.* in *Integration of Information Systems: Bridging Heterogeneous Databases*. 1989: IEEE Press.
66. Thomas, G., et al., *Heterogeneous Distributed Database Systems for Production Use.* ACM Computing Surveys, 1990. **22**(3): p. 237 - 266.
67. Arens, Y., et al., *Retrieving and Integrating Data From Multiple Information Sources.* International Journal on Intelligent and Cooperative Information Systems, 1993. **2**: p. 127-158.
68. Bressan, S., et al., *The COntext INterchange Mediator Prototype.* ACM SIGMOD International Conference on Management of Data, 1997.
69. Garcia-Molina, H., et al., The TSIMMIS Approach to Mediation: Data Models and Languages.1997.http://www-db.stanford.edu/tsimmis/publications.html.
70. Papakonstantinou, Y., et al., A Query Translation Scheme for Rapid Implementation of Wrappers.1995.http://www-db.stanford.edu/tsimmis/publications.html.
71. Roth, M. and P. Schwarz, *Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources.* Proceedings of the 23rd VLDB Conference, 1997.
72. Masarie, F.E., et al., *An Interlingua for Electronic Interchange of Medical Information: Using Frames to Map between Clinical Vocabularies.* Computers and Biomedical Research, 1991. **24**: p. 379-400.
73. Gruber, T., *A translation approach to portable ontology specifications.* Knowledge Acquisition, 1993. **5**(2): p. 199-220.
74. Baorto, D.M., et al., *Using Logical Observation Identifier Names and Codes (LOINC) to exchange laboratory data among three academic hospitals.* Proc AMIA Annu Fall Symp, 1997: p. 96-100.
75. Genesereth, M. and R. Fikes, *Knowledge Interchange Format, Version 3.0 Reference Manual*, in *Computer Science*. 1992, Stanford University: Stanford.
76. Jang, Y., KOLA: Knowledge Organization LAnguage.MIT/LCS/TR-396.Laboratory for Computer Science, MIT.1988.
77. Brachman, R.J. and J.G. Schmolze, *An Overview of the KL-ONE Knowledge Representation System.* Cognitive Science, 1985. **9**(2): p. 171-216.
78. Moser, M.G., An Overview of NIKL, the New Implementation of KL-ONE.Technical Report 5421.Bolt, Beranek and Newman, Inc.1983.
79. Rossi Mori, A. and F. Consorti, *Integration of clinical information across patient records: a comparison of mechanisms used to enforce semantic coherence.* IEEE Transactions on Information Technology in Biomedicine, 1998. **2**(4): p. 243-253.

80. Zollo, K.A. and S.M. Huff, *Automated mapping of observation codes using extensional definitions.* J Am Med Inform Assoc, 2000. **7**(6): p. 586-92.
81. Hull, R. and R. King, *Semantic Database Modeling: Survey, Applications, and Research Issues.* ACM Computing Surveys, 1987. **19**(3): p. 201 - 260.
82. Bic, L., *Processing of Semantic Nets on Dataflow Architectures.* Artificial Intelligence, 1985. **27**: p. 219-227.
83. Cohen, P.R. and C.L. Loiselle. *Beyond {ISA}: Structures for Plausible Inference In Semantic Networks*. in *7th National Conference on Artificial Intelligence*. 1988. St. Paul, MN: Morgan Kaufmann.
84. Zeng, Q. and J.J. Cimino, *A knowledge-based, concept-oriented view generation system for clinical data.* J Biomed Inform, 2001. **34**(2): p. 112-28.

APPENDIX A. LISTING OF CONCEPT MATCHES

The following tables display the detailed results of matching run #1 from the experimental results. In this matching run, the network configurations for both hospitals had all relationships instantiated, and all possible UMLS links were instantiated. This configuration shows node matches that are both terminologically based and context based.

1. Hospital A node matches

Table 1. Direct matches for Hospital A semantic network nodes.
Node: Hospital A node name. Matching nodes: Hospital B node names. UMLS: match corresponds to UMLS link. Coverage: matching set coverage for the node from Hospital A and all its leaf nodes. Score: quality score.

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
Albumin	alb	Yes	Full	100
Alkaline phosphatase	ap	Yes	Full	100
	alk phosphatase	Yes	Full	100
Atypical Lymphs	atyps	Yes	Full	100
Bacteriology	Bacteriology Culture	No	Full	27
	Bacteriology Labs	Yes	Full	25
Bands	band	Yes	Full	100
Base deficit	base excess	Yes	Full	100
Basophils	baso	Yes	Full	100
Bilirubin	bili, total	Yes	Full	100
Blast	blast	Yes	Full	100
Blood culture	blc	Yes	Full	100
Blood gas	bg	Yes	Full	33
BUN	bun	Yes	Full	100
CBC	long1	No	Partial	33
	cbca	Yes	Full	32
	cbcd	Yes	Full	29
	cbc	Yes	Partial	25
Chem 7	basic7	Yes	Full	67
Chemistry	Lab Test	No	Full	13
	balld5	No	Partial	44
	Chemistry Labs	Yes	Full	39
Cholesterol	chol	Yes	Full	100
Creatinine	cret	Yes	Full	100
CSF culture, gram stain	csff	Yes	Full	100
Cultures	Bacteriology Culture	Yes	Full	30
DIC Screen	dic	Yes	Full	50
Electrolytes	bmall5	No	Partial	24
	Electrolytes	No	Partial	46
	Chemistry Labs	No	Full	11
Enzymes	Chemistry Labs	No	Full	5
Enzymes	hfp	No	Full	33

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
Eosinophils	eo	Yes	Full	100
Fibrin split products	fsp	Yes	Full	100
Fibrinogen	fibr	Yes	Full	100
Gram	Bacteriology Culture	No	Full	10
HCO3	sodium bicarbonate	Yes	Full	100
HDL	high dens.lipoprot	Yes	Full	100
Hematocrit	hematocrit	Yes	Full	100
Hematology	Lab Test	No	Full	9
	Blood Counts	No	Partial	27
	balld4	No	Partial	33
	Hematology Labs	Yes	Full	27
Hemoglobin	hemoglobin	Yes	Full	100
IgG	igg	Yes	Full	100
Laboratory test	Chemistry Labs	No	Partial	27
	Lab Test	No	Full	24
LDL	ldl-cholesterol	Yes	Full	100
Lipid profile	ldlp	Yes	Full	40
Lipids	Chemistry Labs	No	Full	7
	ldlp	Yes	Full	80
Liver Function Tests	hfp	Yes	Full	43
Lymphs	lymphs	Yes	Full	100
Microbiology	bmaut2	No	Partial	7
	Bacteriology Culture	No	Partial	18
Monocytes	mono	Yes	Full	100
Other Chemistry	bili	No	Full	50
pCO2	pco2	Yes	Full	100
PCR	pcr	Yes	Full	100
pH	ph	Yes	Full	100
Platelet count	plt	Yes	Full	100
PMN	neutrophils	Yes	Full	100
	poly	Yes	Full	100
pO2	po2	Yes	Full	100
Proteins	tp	Yes	None	0
	bmauto	No	Full	6
	iepu	No	Partial	29
PT	bpt	Yes	Full	100
PTT	bptt	Yes	Full	100
Reticulocytes	ret	Yes	Full	100
Serum calcium	ca	Yes	Full	100
Serum chloride	cl	Yes	Full	100
Serum Glucose	glu	Yes	Full	100
Serum lytes	bmall5	No	Full	32
	Electrolytes	No	Full	86
Serum magnesium	mg	Yes	Full	100
Serum phosphorus	phos	Yes	Full	100
Serum potassium	k	Yes	Full	100
Serum sodium	na	Yes	Full	100
SGOT	sgot	Yes	Full	100

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
SGPT	sgpt	Yes	Full	100
Stains	Bacteriology Culture	No	Full	9
Total protein	tp	Yes	Full	100
Triglyceride	trig	Yes	Full	100
	triglyceride	Yes	Full	100
Urine culture	urnc	Yes	Full	100
Virology	Virology Labs	Yes	None	0
	bmaut2	No	Full	10
WBC	wbc count	Yes	Full	100
WBC differential	cbca	No	Full	22
	difa	Yes	Partial	46
	diff	Yes	Partial	35
	WBC differential count	Yes	Partial	32

Table 2. Generalized matches for Hospital A semantic network nodes.
Node: Hospital A node name, Matching nodes: Hospital B node names. No matching metrics are given because generalized matches are performed only on nodes where the concept exists only in one of the networks (i.e. match quality scores are always zero).

Node	***Matching Node***
Activated clotting time	Hematology Labs
	Lab Test
	Blood Counts
	balld4
Bleeding time	Hematology Labs
	Lab Test
	Blood Counts
	balld4
Creatine kinase	hfp
	Chemistry Labs
fibrin d-dimers	Hematology Labs
	Lab Test
	dic
	Blood Counts
	balld4
GGT	hfp
	Chemistry Labs
HBsAg	Virology Labs
	bmaut2
HSV Culture	Virology Labs
	bmaut2
HSV II antigen	Virology Labs
	bmaut2
KOH	Bacteriology Culture
RSV antigen	Virology Labs

Node	***Matching Node***
RSV antigen	bmaut2
RSV Culture	Virology Labs
	bmaut2
Total CO2	bmall5
	Electrolytes
	Chemistry Labs
	basic7
Viral Antigen tests	Virology Labs
	bmaut2
viral cultures	Virology Labs
	bmaut2
WB chloride	bmall5
	Electrolytes
	Chemistry Labs
WB glucose	bmall5
	Electrolytes
	Chemistry Labs
WB potassium	bmall5
	Electrolytes
	Chemistry Labs
WB sodium	bmall5
	Electrolytes
	Chemistry Labs
Whole blood lytes	bmall5
	Electrolytes
	Chemistry Labs

Table 3. Unmatched nodes for Hospital A. All unmatched nodes are either disconnected from the network, or linked only by the "attribute-of" relationship.

Accession number	Source
Comments	Specimen source
Hemogram*	Test ID
Lower Reference Range	Test name
Patient ID	Time-stamp
Result status	Units
Result value	Upper Reference Range

* Unconnected node (no links)

2. Hospital B node matches

Table 4. Direct matches for Hospital B semantic network nodes.
Node: Hospital B node name. Matching nodes: Hospital A node names. UMLS: match corresponds to UMLS link. Coverage: matching set coverage for the node from Hospital A and all its leaf nodes. Score: quality score.

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
alb	Albumin	Yes	Full	100
alk phosphatase	Alkaline phosphatase	Yes	Full	100
ap	Alkaline phosphatase	Yes	Full	100
atyps	Atypical Lymphs	Yes	Full	100
Bacteriology Culture	Cultures	Yes	Full	30
Bacteriology Labs	Bacteriology	Yes	Full	25
	Cultures	No	Full	27
balld4	Hematology	No	Full	33
balld5	Chemistry	No	Full	44
	Chem 7	No	Partial	29
band	Bands	Yes	Full	100
base excess	Base deficit	Yes	Full	100
basic7	Chem 7	Yes	Partial	67
baso	Basophils	Yes	Full	100
bg	Blood gas	Yes	Full	33
bili	Liver Function Tests	No	Full	25
	Other Chemistry	No	Full	50
bili, total	Bilirubin	Yes	Full	100
blast	Blast	Yes	Full	100
blc	Blood culture	Yes	Full	100
Blood Counts	Hematology	No	Partial	27
	CBC	No	Partial	29
BM Transplant Tests	Laboratory test	No	Partial	27
bma	WBC differential	No	Full	6
bmall2	Chem 7	No	Full	7
bmall4	Chemistry	No	Partial	4
	Laboratory test	No	Partial	15
	CBC	No	Partial	25
bmall5	Chemistry	No	Full	42
	Electrolytes	No	Partial	24
	Chem 7	No	Partial	30
bmallo	Chemistry	No	Partial	17
	Laboratory test	No	Full	27
bmaut2	Microbiology	No	Full	7
	Virology	No	Full	10
bmauto	Chemistry	No	Partial	17
	Laboratory test	No	Partial	28
bpt	PT	Yes	Full	100
bptt	PTT	Yes	Full	100

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
bun	BUN	Yes	Full	100
ca	Serum calcium	Yes	Full	100
cbc	CBC	Yes	Full	25
cbca	CBC	Yes	Partial	32
cbcd	CBC	Yes	Partial	29
Chemistry Labs	Chemistry	Yes	Partial	39
chol	Cholesterol	Yes	Full	100
cl	Serum chloride	Yes	Full	100
comp12	Chemistry	No	Full	38
	Chem 7	No	Partial	38
cret	Creatinine	Yes	Full	100
csff	CSF culture, gram stain	Yes	Full	100
dic	Hematology	No	Full	19
	DIC Screen	Yes	Partial	50
difa	WBC differential	Yes	Full	46
diff	WBC differential	Yes	Full	35
Electrolytes	Serum lytes	No	Full	86
	Electrolytes	No	Full	46
	Chem 7	No	Partial	27
eo	Eosinophils	Yes	Full	100
fibr	Fibrinogen	Yes	Full	100
fmmbmt	Chem 7	No	Full	25
frap	Enzymes	No	Full	17
fsp	Fibrin split products	Yes	Full	100
g6p	CBC	No	Full	8
glu	Serum Glucose	Yes	Full	100
hematocrit	Hematocrit	Yes	Full	100
Hematology Labs	Hematology	Yes	Partial	27
hemoglobin	Hemoglobin	Yes	Full	100
hfp	Liver Function Tests	Yes	Partial	43
	Chemistry	No	Full	18
high dens.lipoprot	HDL	Yes	Full	100
iepu	Chemistry	No	Full	6
igg	IgG	Yes	Full	100
iglb	Proteins	No	Full	20
k	Serum potassium	Yes	Full	100
Lab Test	Laboratory test	No	Partial	24
	Microbiology	No	Partial	2
ldl-cholesterol	LDL	Yes	Full	100
ldlp	Lipid profile	Yes	Partial	40
	Lipids	Yes	Full	80
long1	CBC	No	Partial	33
lymphs	Lymphs	Yes	Full	100
lyte	Serum lytes	No	Full	43
	Chem 7	No	Full	38
mg	Serum magnesium	Yes	Full	100
mono	Monocytes	Yes	Full	100
na	Serum sodium	Yes	Full	100

Node	*Matching Node*	*UMLS*	*Coverage*	*Score*
neutrophils	PMN	Yes	Full	100
newa	Chemistry	No	Full	28
pco2	pCO2	Yes	Full	100
pcr	PCR	Yes	Full	100
ph	pH	Yes	Full	100
phos	Serum phosphorus	Yes	Full	100
plt	Platelet count	Yes	Full	100
po2	pO2	Yes	Full	100
poly	PMN	Yes	Full	100
ret	Reticulocytes	Yes	Full	100
sgot	SGOT	Yes	Full	100
sgpt	SGPT	Yes	Full	100
sodium bicarbonate	HCO3	Yes	Full	100
tp	Total protein	Yes	Full	100
	Proteins	Yes	None	0
trig	Triglyceride	Yes	Full	100
triglyceride	Triglyceride	Yes	Full	100
urnc	Urine culture	Yes	Full	100
Virology Labs	Virology	Yes	None	0
wbc count	WBC	Yes	Full	100
WBC differential count	WBC differential	Yes	Partial	32

Table 5. Generalized matches for Hospital B semantic network nodes.
Node: Hospital B node name, Matching nodes: Hospital A node names. No matching metrics are given because generalized matches are performed only on nodes where the concept exists only in one of the networks (i.e. match quality scores are always zero).

Node	*Matching Node*
aat3	Chemistry
abs atyps	Hematology
	CBC
abs band	Hematology
	Chemistry
	Laboratory test
	CBC
abs basos	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs blasts	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs eos	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs fissured lymphs	Hematology
	CBC
abs lymphs	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs meta	Hematology
	CBC
abs monos	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs myelo	Hematology
	CBC
abs neutrophils	WBC differential
	Hematology
	Chemistry
	Laboratory test

Node	*Matching Node*
abs neutrophils	CBC
abs plasma cell	Hematology
	CBC
abs polys	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
abs promyel	Hematology
	CBC
acetest	Laboratory test
	Microbiology
ahav	Virology
amy	Chemistry
asp cellurity	WBC differential
	Hematology
balld3	Laboratory test
bili, direct	Liver Function Tests
	Chemistry
	Laboratory test
	Electrolytes
	Chem 7
	Other Chemistry
blood	Laboratory test
	Microbiology
bm site	WBC differential
	Laboratory test
bmall3	Laboratory test
bmaut3	Laboratory test
bone marrow comment	WBC differential
	Laboratory test
bwh type/screen	Hematology
	Chemistry
	Laboratory test
	CBC
bwh vaccine abc typing	CBC
c125	CBC
c2729	CBC
cct	Hematology
ceah	Chemistry
ceah	CBC

Node	*Matching Node*
clinitest	Laboratory test
	Microbiology
cmv ab igg	Hematology
	Chemistry
	Laboratory test
	CBC
	PCR
cmv ab igg	Viral Antigen tests
	viral cultures
cmv chemilumin assay	PCR
	Viral Antigen tests
	viral cultures
cmv enzymatic digest	PCR
	Viral Antigen tests
	viral cultures
cmv infec agent/dna-rna dir pr	PCR
	Viral Antigen tests
	viral cultures
cmv molec dx extract	PCR
	Viral Antigen tests
	viral cultures
CMV tests	Virology
cmvvla	PCR
	Viral Antigen tests
	viral cultures
co2	Chemistry
	Laboratory test
co2	Serum lytes
	Electrolytes
	Chem 7
comments	Laboratory test
	Microbiology
csf comment	Hematology
cytocentrifuge	Hematology
ebv-vca	Hematology
	Chemistry
	Laboratory test
	Chem 7
	Microbiology
	CBC
	Virology
erythroid	WBC differential

Node	*Matching Node*
	Hematology
esr	Hematology
	Chemistry
	Laboratory test
ferr	Chemistry
	Laboratory test
	CBC
fio2	Blood gas
fissured lymphs	Hematology
	CBC
fluid appearance	Hematology
fluid rbc count	Hematology
fluid wbc count	Hematology
genc	Cultures
globulin	Chemistry
	Laboratory test
	CBC
gluc.6 phos.deh.scr.	Hematology
	Chemistry
	CBC
hbc	Laboratory test
hbs	Chemistry
	Virology
hcv	Chemistry
	Virology
heart rate	Blood gas
heat stab.alk.phos.	Enzymes
	Chemistry
hep b surface ab	PCR
	Viral Antigen tests
hep b surface ab	viral cultures
hepatitis a ab	PCR
	Viral Antigen tests
	viral cultures
hepatitis a antibody	PCR
	Viral Antigen tests
	viral cultures
hepatitis be ab	PCR
	Viral Antigen tests
	viral cultures
herpes i antibody	Chemistry
	Laboratory test
	Chem 7
	Microbiology

Node	*Matching Node*
	CBC
	PCR
	Viral Antigen tests
	viral cultures
	Virology
herpes ii antibody	Chemistry
	Laboratory test
	Chem 7
	Microbiology
	CBC
	PCR
	Viral Antigen tests
	viral cultures
	Virology
hsv interpretation	Chemistry
	Laboratory test
	Chem 7
hsv interpretation	Microbiology
	CBC
	Virology
htlv1 antibody	PCR
	Viral Antigen tests
	viral cultures
iga	Chemistry
	Laboratory test
	CBC
	Proteins
igm	Chemistry
	Laboratory test
	CBC
	Proteins
immunoelectro	Chemistry
infec agent/dna-rna amp probe	PCR
	Viral Antigen tests
	viral cultures
ldh	Chemistry
	Laboratory test
	Electrolytes
	Chem 7
lymphoid	WBC differential
	Hematology
mch	Hematology
	Chemistry
	Laboratory test

Node	*Matching Node*
	CBC
mchc	Hematology
	Chemistry
	Laboratory test
	CBC
mcv	Hematology
	Chemistry
	Laboratory test
	CBC
megakaryocyte	WBC differential
	Hematology
meta	Hematology
	CBC
mpcr	Chem 7
myelo	Hematology
	CBC
myeloid	WBC differential
	Hematology
nucleated rbc's	WBC differential
	Hematology
	CBC
o2 admin. device	Blood gas
o2 liters per min.	Blood gas
oap	Cultures
oxygen saturation	Blood gas
parainfluenza 1	PCR
	Viral Antigen tests
	viral cultures
parainfluenza 2	PCR
	Viral Antigen tests
	viral cultures
parainfluenza 3	PCR
	Viral Antigen tests
	viral cultures
plasma cell	Hematology
	CBC
plt morphology	WBC differential
	Hematology
	Chemistry
	Laboratory test
	CBC
promyel	Hematology
	CBC
promyelo	WBC differential
	Hematology

Node	*Matching Node*
prot electro	Chemistry
rapid adenovirus	PCR
	Viral Antigen tests
	viral cultures
rapid hsv	PCR
	Viral Antigen tests
	viral cultures
rapid influenza a	PCR
	Viral Antigen tests
	viral cultures
rapid influenza b	PCR
	Viral Antigen tests
	viral cultures
rapid rsv	PCR
	Viral Antigen tests
	viral cultures
rbc count	Hematology
rbc count	Chemistry
	Laboratory test
	CBC
rbc morphology	WBC differential
	Hematology
	Chemistry
	Laboratory test
rbc morphology	CBC
rbcs in urine	Laboratory test
	Microbiology
rdspec	Virology
reference lab	PCR
	Viral Antigen tests
	viral cultures
resp	Cultures
respiratory rate	Blood gas
rpr	Chemistry
	Bacteriology
	Cultures
s/n ratio	Laboratory test
sample	Blood gas
samples to cell bank	Laboratory test
serum storage	Laboratory test
skin	Cultures
stlc	Cultures

Node	*Matching Node*
stlk	Cultures
t3u	Chemistry
t3u	Chem 7
t4	Chemistry
	Chem 7
temperature	Blood gas
test site	Blood gas
thsc	Cultures
tot cells counted	Hematology
	CBC
total carbon dioxide	Blood gas
	Chemistry
total globulin	Chemistry
toxoplasmosis ab igg	Chemistry
	Laboratory test
	CBC
tsh	Chemistry
	Laboratory test
	Chem 7
	CBC
tt	Hematology
	DIC Screen
ua	Laboratory test
	Microbiology
uric	Chemistry
	Electrolytes
	Chem 7
urine appearance	Laboratory test
	Microbiology
urine bacti	Laboratory test
urine bacti	Microbiology
urine bilirubin	Laboratory test
	Microbiology
urine casts	Laboratory test
	Microbiology
urine crystals	Laboratory test
	Microbiology
urine epithelial	Laboratory test
	Microbiology
urine glucose	Laboratory test
	Microbiology
urine ketones	Laboratory test
	Microbiology
urine mucus	Laboratory test
	Microbiology
urine ph	Laboratory test
urine ph	Microbiology
urine protein	Laboratory test

Node	*Matching Node*
	Microbiology
urine spec gravity	Laboratory test
	Microbiology
urobilinogen	Laboratory test
	Microbiology
varicella antibody	Chemistry
	Laboratory test
	CBC
	PCR
	Viral Antigen tests

Node	*Matching Node*
	viral cultures
very low density	Lipid profile
	Lipids
Viral serology	Virology
volume	Chemistry
wbc in urine	Laboratory test
	Microbiology
wbc morphology	Hematology
	Chemistry
	Laboratory test
	CBC

Table 6. Unmatched nodes for Hospital B. All unmatched nodes are either disconnected from the main network, or linked only by the "attribute-of" relationship.

5hia
5nuc
a1cb
Accession Number*
adcic
afp
ahbs
ahclot
aldo
ana
apad
apai
b12
b2m
bcyt
bhgbe
biopsy
blood component type
bwbm
bwbx
bwh cytology result
ca199
cal mean glucose
cdif
cglu
ch50
cmgtl4
corti
cortisol
cpk
crca
creat clearance
crpq
ctp
culture
cycl
dig
dihy
dil
estr
fa2

fa5l
fa7
fa8
fa9
factor 8 antigen
factor 8 functional
fenret
fol
fsh
fti
genp
ggtp
glur
gly a1c equivalent
granin
hapt
hbea
hcgb
hcgt
hepatitis be ag
iron
kathu
lh
lipa
Medical Record Number*
Normal range*
npbank
osfr
oxim
pap
pcp
period
pk
po
protc
prots
psa
pv
qacp
qandr
qdhea

qdheas
qestn
qpacp
qshbg
ravb
rbcu
report status
Result Type*
Result units*
rhcg
rosu
slbw
special requests
specimen description
sppb
status of unit
tacro
teg
test
Test name*
Text result*
tibc
time
tpch
trans
transfusion status
una
unit number
upr24
urine creatinine
urine protein 24hr
valp
visc
von willebrand fac.
vwfw

* Attribute concept. The node is linked to other nodes only through the "attribute-of" relationship.

APPENDIX B. LEAF MATCHES

The following tables display the detailed leaf matches of matching run #6 from the experimental results. In this matching run, the network configurations for both hospitals had all relationships instantiated, but only the leaf nodes had UMLS links instantiated. In other words, all non-leaf nodes were matched on a purely algorithmic basis utilizing concept contexts.

Table 1. Hospital A leaf matches. Node: target node. Score: leaf match quality score, or the percentage of leaf nodes matched. Matched leaves: leaf nodes which were successfully matched, with the matching Hospital B node in parentheses. Unmatched leaves: self-explanatory.

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
Bacteriology	75	CSF culture, gram stain(csff); Blood culture(blc); Urine culture(urnc)	KOH
Blood gas	100	Base deficit(base excess); HCO3(sodium bicarbonate); pO2(po2); pCO2(pco2); pH(ph)	
CBC	100	Hemoglobin(hemoglobin); Blast(blast); Monocytes(mono); Basophils(baso); Eosinophils(eo); Bands(band); PMN(neutrophils; poly); WBC(wbc count); Platelet count(plt); Hematocrit(hematocrit); Lymphs(lymphs); Atypical Lymphs(atyps)	
Chem 7	86	Serum chloride(cl); Serum Glucose(glu); Creatinine(cret); Serum potassium(k); Serum sodium(na); BUN(bun)	Total CO2
Chemistry	78	SGPT(sgpt); SGOT(sgot); Serum phosphorus(phos); Serum magnesium(mg); Serum calcium(ca); Serum chloride(cl); Serum Glucose(glu); Creatinine(cret); Serum potassium(k); Serum sodium(na); Base deficit(base excess); BUN(bun); HCO3(sodium bicarbonate); pO2(po2); pCO2(pco2); pH(ph); Bilirubin(bili, total); IgG(igg); Albumin(alb); Total protein(tp); HDL(high dens.lipoprot); LDL(ldl-cholesterol); Cholesterol(chol); Triglyceride(trig; triglyceride); Alkaline phosphatase(ap; alk phosphatase)	GGT; Creatine kinase; WB glucose; WB chloride; WB potassium; WB sodium; Total CO2
Cultures	100	CSF culture, gram stain(csff); Blood culture(blc); Urine culture(urnc)	
DIC Screen	75	PT(bpt); Fibrinogen(fibr); PTT(bptt)	fibrin d-dimers
Electrolytes	58	Serum phosphorus(phos); Serum magnesium(mg); Serum calcium(ca); Serum chloride(cl); Serum potassium(k); Serum sodium(na); HCO3(sodium bicarbonate)	WB glucose; WB chloride; WB potassium; WB sodium; Total CO2
Enzymes	60	SGPT(sgpt); SGOT(sgot); Alkaline phosphatase(ap; alk phosphatase)	GGT; Creatine kinase
Gram	100	CSF culture, gram stain(csff)	
Hematology	85	PT(bpt); Hemoglobin(hemoglobin); Reticulocytes(ret); Blast(blast); Monocytes(mono); Basophils(baso); Eosinophils(eo); Bands(band); PMN(neutrophils; poly); WBC(wbc count); Platelet count(plt); Hematocrit(hematocrit); Lymphs(lymphs); Fibrin split	Activated clotting time; fibrin d-dimers; Bleeding time

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
Hematology (cont.)		products(fsp); Fibrinogen(fibr); Atypical Lymphs(atyps); PTT(bptt)	
Laboratory test	74	SGPT(sgpt); PT(bpt); SGOT(sgot); Hemoglobin(hemoglobin); Reticulocytes(ret); Blast(blast); Monocytes(mono); Basophils(baso); Eosinophils(eo); Bands(band); PMN(neutrophils; poly); CSF culture, gram stain(csff); WBC(wbc count); Platelet count(plt); Hematocrit(hematocrit); Serum phosphorus(phos); Serum magnesium(mg); Serum calcium(ca); Serum chloride(cl); Serum Glucose(glu); Creatinine(cret); Lymphs(lymphs); Serum potassium(k); Serum sodium(na); BUN(bun); Base deficit(base excess); HCO3(sodium bicarbonate); pO2(po2); pCO2(pco2); pH(ph); Bilirubin(bili, total); IgG(igg); Albumin(alb); Fibrin split products(fsp); Total protein(tp); Fibrinogen(fibr); Atypical Lymphs(atyps); PCR(pcr); PTT(bptt); Blood culture(blc); Urine culture(urnc); HDL(high dens.lipoprot); LDL(ldl-cholesterol); Cholesterol(chol); Triglyceride(trig; triglyceride); Alkaline phosphatase(ap; alk phosphatase)	GGT; Creatine kinase; WB glucose; WB chloride; RSV antigen; WB potassium; WB sodium; HSV Culture; HSV II antigen; Total CO2; RSV Culture; HBsAg; Activated clotting time; fibrin d-dimers; Bleeding time; KOH
Lipid profile	100	Cholesterol(chol); Triglyceride(trig; triglyceride)	
Lipids	100	HDL(high dens.lipoprot); LDL(ldl-cholesterol); Cholesterol(chol); Triglyceride(trig; triglyceride)	
Liver Function Tests	100	SGPT(sgpt); SGOT(sgot); Bilirubin(bili, total)	
Microbiology	40	CSF culture, gram stain(csff); PCR(pcr); Blood culture(blc); Urine culture(urnc)	RSV antigen; HSV Culture; HSV II antigen; RSV Culture; HBsAg; KOH
Other Chemistry	100	Bilirubin(bili, total)	
Proteins	100	IgG(igg); Albumin(alb); Total protein(tp)	
Serum lytes	100	Serum phosphorus(phos); Serum magnesium(mg); Serum calcium(ca); Serum chloride(cl); Serum potassium(k); Serum sodium(na)	
Stains	50	CSF culture, gram stain(csff)	KOH
Virology	17	PCR(pcr)	RSV antigen; HSV Culture; HSV II antigen; RSV Culture; HBsAg
WBC differential	100	Blast(blast); Monocytes(mono); Basophils(baso); Eosinophils(eo); Bands(band); PMN(neutrophils; poly); Lymphs(lymphs); Atypical Lymphs(atyps)	

Table 2. Hospital B leaf matches. Node: target node. Score: leaf match quality score, or the percentage of leaf nodes matched. Matched leaves: leaf nodes which were successfully matched, with the matching Hospital A node in parentheses. Unmatched leaves: self-explanatory.

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
Bacteriology Culture	30	blc(Blood culture); urnc(Urine culture); csff(CSF culture, gram stain)	oap; skin; stlc; genc; stlk; thsc; resp
Bacteriology Labs	27	blc(Blood culture); urnc(Urine culture); csff(CSF culture, gram stain)	oap; skin; stlc; genc; stlk; rpr; thsc; resp
balld4	43	lymphs(Lymphs); bptt(PTT); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); neutrophils(PMN); plt(Platelet count); baso(Basophils); hematocrit(Hematocrit); blast(Blast); bpt(PT)	abs blasts; abs neutrophils; gluc.6 phos.deh.scr.; abs eos; plt morphology; abs lymphs; ebv-vca; bwh type/screen; mchc; mcv; rbc morphology; mch; cmv ab igg; rbc count; abs monos; abs basos
balld5	80	ca(Serum calcium); bili, total(Bilirubin); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); trig(Triglyceride); mg(Serum magnesium); tp(Total protein); alb(Albumin); bun(BUN); sgot(SGOT); cret(Creatinine); sgpt(SGPT); k(Serum potassium); cl(Serum chloride); phos(Serum phosphorus)	co2; uric; bili, direct; ldh
basic7	88	ca(Serum calcium); na(Serum sodium); glu(Serum Glucose); bun(BUN); cret(Creatinine); k(Serum potassium); cl(Serum chloride)	co2
bg	33	pco2(pCO2); sodium bicarbonate(HCO3); ph(pH); po2(pO2); base excess(Base deficit)	o2 admin. device; temperature; respiratory rate; o2 liters per min.; total carbon dioxide; sample; test site; oxygen saturation; heart rate; fio2
bili	50	bili, total(Bilirubin)	bili, direct
Blood Counts	33	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); atyps(Atypical Lymphs); ret(Reticulocytes); neutrophils(PMN); band(Bands); plt(Platelet count); poly(PMN); baso(Basophils); hematocrit(Hematocrit); blast(Blast)	abs band; abs blasts; abs myelo; nucleated rbc's; abs fissured lymphs; tot cells counted; wbc morphology; abs neutrophils; myelo; abs eos; abs plasma cell; abs atyps; abs meta; fissured lymphs; meta; abs lymphs; plt morphology; plasma cell; mchc; mcv; rbc morphology; abs promyel; promyel; mch; abs polys; rbc count; abs monos; abs basos
BM Transplant Tests	38	lymphs(Lymphs); bptt(PTT); pcr(PCR); ca(Serum calcium); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); bili, total(Bilirubin); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); trig(Triglyceride); eo(Eosinophils); ret(Reticulocytes); neutrophils(PMN); band(Bands); mg(Serum magnesium); tp(Total protein); alb(Albumin); bun(BUN); plt(Platelet count); poly(PMN); baso(Basophils); sgot(SGOT);	wbc morphology; tsh; hsv interpretation; hbs; hepatitis a ab; hbc; bwh type/screen; mcv; rbc morphology; bili, direct; mch; myeloid; erythroid; abs polys; s/n ratio; promyelo; esr; toxoplasmosis ab igg; abs basos; herpes ii antibody; hepatitis a antibody; rbc count; abs monos; varicella antibody; hep b surface ab; htlv1 antibody; co2; asp cellurity; abs neutrophils; gluc.6

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
BM Transplant Tests (cont.)		cret(Creatinine); igg(IgG); sgpt(SGPT); k(Serum potassium); hematocrit(Hematocrit); blast(Blast); cl(Serum chloride); phos(Serum phosphorus); bpt(PT)	phos.deh.scr.; uric; t4; abs eos; plt morphology; abs lymphs; rpr; ebv-vca; ldh; herpes i antibody; mchc; lymphoid; bm site; abs band; t3u; abs blasts; globulin; bone marrow comment; ferr; hcv; megakaryocyte; igm; samples to cell bank; cmv ab igg; serum storage; iga
bma	11	blast(Blast)	asp cellurity; bm site; bone marrow comment; megakaryocyte; myeloid; erythroid; promyelo; lymphoid
bmall2	13	glu(Serum Glucose)	tsh; t4; hsv interpretation; ebv-vca; herpes i antibody; t3u; herpes ii antibody
bmall4	37	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); trig(Triglyceride); neutrophils(PMN); tp(Total protein); alb(Albumin); plt(Platelet count); poly(PMN); baso(Basophils); hematocrit(Hematocrit); blast(Blast)	abs blasts; tsh; abs neutrophils; abs eos; hsv interpretation; plt morphology; abs lymphs; ebv-vca; herpes i antibody; bwh type/screen; mchc; mcv; rbc morphology; herpes ii antibody; globulin; mch; ferr; abs polys; cmv ab igg; rbc count; abs monos; varicella antibody; toxoplasmosis ab igg; abs basos
bmall5	79	ca(Serum calcium); bili, total(Bilirubin); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); mg(Serum magnesium); tp(Total protein); alb(Albumin); bun(BUN); sgot(SGOT); cret(Creatinine); sgpt(SGPT); k(Serum potassium); cl(Serum chloride); phos(Serum phosphorus)	co2; uric; bili, direct; ldh
bmallo	51	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); bili, total(Bilirubin); na(Serum sodium); ap(Alkaline phosphatase); eo(Eosinophils); ret(Reticulocytes); neutrophils(PMN); alb(Albumin); bun(BUN); plt(Platelet count); baso(Basophils); sgot(SGOT); cret(Creatinine); igg(IgG); sgpt(SGPT); k(Serum potassium); hematocrit(Hematocrit); blast(Blast); cl(Serum chloride)	abs blasts; co2; abs neutrophils; abs eos; plt morphology; abs lymphs; mchc; mcv; rbc morphology; bili, direct; mch; igm; ldh; cmv ab igg; rbc count; abs monos; esr; varicella antibody; toxoplasmosis ab igg; iga; abs basos
bmaut2	20	pcr(PCR)	hsv interpretation; ebv-vca; herpes i antibody; herpes ii antibody
bmauto bmauto (cont.)	51	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); bili, total(Bilirubin); na(Serum sodium); ap(Alkaline phosphatase); eo(Eosinophils); ret(Reticulocytes); neutrophils(PMN); band(Bands); tp(Total protein); alb(Albumin); bun(BUN); plt(Platelet count); poly(PMN); baso(Basophils); sgot(SGOT); cret(Creatinine); igg(IgG); sgpt(SGPT); k(Serum potassium); hematocrit(Hematocrit); blast(Blast); cl(Serum chloride)	abs band; abs blasts; wbc morphology; co2; abs neutrophils; abs eos; plt morphology; abs lymphs; mchc; mcv; rbc morphology; bili, direct; mch; igm; ldh; abs polys; cmv ab igg; rbc count; abs monos; esr; varicella antibody; toxoplasmosis ab igg; iga; abs basos
cbc	50	hemoglobin(Hemoglobin); wbc count(WBC);	mchc; mcv; mch; rbc count

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
		plt(Platelet count); hematocrit(Hematocrit)	
cbca	35	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); atyps(Atypical Lymphs); neutrophils(PMN); band(Bands); plt(Platelet count); poly(PMN); baso(Basophils); hematocrit(Hematocrit); blast(Blast)	abs band; abs blasts; abs myelo; nucleated rbc's; tot cells counted; wbc morphology; abs neutrophils; myelo; abs eos; abs atyps; abs meta; meta; plt morphology; abs lymphs; mchc; mcv; rbc morphology; abs promyel; promyel; mch; abs polys; rbc count; abs monos; abs basos
cbcd	32	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); atyps(Atypical Lymphs); neutrophils(PMN); band(Bands); plt(Platelet count); poly(PMN); baso(Basophils); hematocrit(Hematocrit); blast(Blast)	abs band; abs blasts; abs myelo; nucleated rbc's; abs fissured lymphs; tot cells counted; wbc morphology; abs neutrophils; myelo; abs eos; abs atyps; abs plasma cell; abs meta; fissured lymphs; meta; plt morphology; abs lymphs; plasma cell; mchc; mcv; rbc morphology; abs promyel; promyel; mch; abs polys; rbc count; abs monos; abs basos
Chemistry Labs	47	pco2(pCO2); ca(Serum calcium); ldl-cholesterol(LDL); bili, total(Bilirubin); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); chol(Cholesterol); trig(Triglyceride); sodium bicarbonate(HCO3); mg(Serum magnesium); tp(Total protein); alb(Albumin); bun(BUN); ph(pH); po2(pO2); base excess(Base deficit); sgot(SGOT); cret(Creatinine); sgpt(SGPT); igg(IgG); k(Serum potassium); high dens.lipoprot(HDL); cl(Serum chloride); phos(Serum phosphorus); triglyceride(Triglyceride); alk phosphatase(Alkaline phosphatase)	hcv; heat stab.alk.phos.; amy; temperature; o2 admin. device; tsh; co2; gluc.6 phos.deh.scr.; respiratory rate; uric; very low density lipoprotein; t4; hbs; o2 liters per min.; ceah; total carbon dioxide; sample; t3u; globulin; test site; bili, direct; rpr; ferr; oxygen saturation; igm; aat3; heart rate; ldh; fio2; iga
comp12 comp12 (cont.)	87	ca(Serum calcium); bili, total(Bilirubin); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); tp(Total protein); alb(Albumin); bun(BUN); sgot(SGOT); cret(Creatinine); sgpt(SGPT); k(Serum potassium); cl(Serum chloride)	co2; bili, direct
dic	80	bptt(PTT); fsp(Fibrin split products); fibr(Fibrinogen); bpt(PT)	tt
difa	55	lymphs(Lymphs); mono(Monocytes); eo(Eosinophils); neutrophils(PMN); baso(Basophils); blast(Blast)	abs neutrophils; abs eos; abs lymphs; abs monos; abs basos
diff	40	lymphs(Lymphs); mono(Monocytes); eo(Eosinophils); poly(PMN); baso(Basophils); blast(Blast)	abs blasts; nucleated rbc's; abs eos; plt morphology; abs lymphs; rbc morphology; abs polys; abs monos; abs basos
Electrolytes	86	ca(Serum calcium); na(Serum sodium); mg(Serum magnesium); k(Serum potassium); cl(Serum chloride); phos(Serum phosphorus)	co2
fmmbmt	67	bun(BUN); cret(Creatinine)	mpcr
frap	50	alk phosphatase(Alkaline phosphatase)	heat stab.alk.phos.
g6p	50	hemoglobin(Hemoglobin)	gluc.6 phos.deh.scr.

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
Hematology Labs	30	lymphs(Lymphs); bptt(PTT); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); fsp(Fibrin split products); eo(Eosinophils); atyps(Atypical Lymphs); ret(Reticulocytes); neutrophils(PMN); band(Bands); fibr(Fibrinogen); plt(Platelet count); poly(PMN); baso(Basophils); hematocrit(Hematocrit); blast(Blast); bpt(PT)	abs band; abs blasts; abs myelo; nucleated rbc's; fluid rbc count; abs fissured lymphs; csf comment; cytocentrifuge; tot cells counted; fluid wbc count; wbc morphology; asp cellurity; abs neutrophils; myelo; abs eos; abs atyps; abs plasma cell; abs meta; fissured lymphs; meta; plt morphology; abs lymphs; plasma cell; bwh type/screen; mchc; mcv; rbc morphology; abs promyel; promyel; fluid appearance; mch; megakaryocyte; myeloid; erythroid; abs polys; rbc count; abs monos; promyelo; esr; lymphoid; abs basos; tt
hfp	86	bili, total(Bilirubin); ap(Alkaline phosphatase); tp(Total protein); alb(Albumin); sgot(SGOT); sgpt(SGPT)	bili, direct
iepu	33	tp(Total protein); alb(Albumin)	prot electro; total globulin; volume; immunoelectro
iglb	33	igg(IgG)	igm; iga
Lab Test Lab Test (cont.)	27	lymphs(Lymphs); pco2(pCO2); bptt(PTT); ca(Serum calcium); pcr(PCR); blc(Blood culture); hemoglobin(Hemoglobin); mono(Monocytes); ldl-cholesterol(LDL); wbc count(WBC); bili, total(Bilirubin); fsp(Fibrin split products); na(Serum sodium); glu(Serum Glucose); ap(Alkaline phosphatase); chol(Cholesterol); trig(Triglyceride); eo(Eosinophils); urnc(Urine culture); sodium bicarbonate(HCO3); atyps(Atypical Lymphs); ret(Reticulocytes); neutrophils(PMN); csff(CSF culture, gram stain); band(Bands); mg(Serum magnesium); fibr(Fibrinogen); tp(Total protein); alb(Albumin); bun(BUN); plt(Platelet count); poly(PMN); ph(pH); po2(pO2); baso(Basophils); base excess(Base deficit); sgot(SGOT); cret(Creatinine); igg(IgG); sgpt(SGPT); k(Serum potassium); high dens.lipoprot(HDL); hematocrit(Hematocrit); blast(Blast); cl(Serum chloride); phos(Serum phosphorus); bpt(PT); alk phosphatase(Alkaline phosphatase); triglyceride(Triglyceride)	stlk; test site; mch; parainfluenza 3; parainfluenza 2; oxygen saturation; myeloid; heart rate; abs polys; fio2; promyelo; esr; toxoplasmosis ab igg; abs basos; herpes ii antibody; hepatitis a antibody; hepatitis be ab; rbc count; abs monos; varicella antibody; nucleated rbc's; fluid rbc count; abs fissured lymphs; hep b surface ab; htlv1 antibody; cytocentrifuge; o2 admin. device; skin; co2; myelo; abs eos; rapid rsv; o2 liters per min.; plt morphology; herpes i antibody; plasma cell; genc; mchc; globulin; bone marrow comment; ferr; igm; megakaryocyte; urine ph; urine bilirubin; cmv ab igg; iga; urobilinogen; abs blasts; oap; hcv; csf comment; urine epithelial; serum storage; wbc morphology; hbs; urine crystals; rapid influenza
ldlp	80	ldl-cholesterol(LDL); chol(Cholesterol); high dens.lipoprot(HDL); triglyceride(Triglyceride)	very low density lipoprotein
long1	38	lymphs(Lymphs); hemoglobin(Hemoglobin); mono(Monocytes); wbc count(WBC); eo(Eosinophils); atyps(Atypical Lymphs); plt(Platelet count); poly(PMN); baso(Basophils); igg(IgG); hematocrit(Hematocrit); blast(Blast)	abs blasts; nucleated rbc's; c125; abs eos; abs atyps; plt morphology; abs lymphs; ceah; mchc; mcv; c2729; rbc morphology; mch; igm; abs polys; rbc count; abs monos; bwh vaccine abc typing; iga; abs basos

Node	*Score*	*Matched leaves (matching node)*	*Unmatched leaves*
lyte	75	na(Serum sodium); k(Serum potassium); cl(Serum chloride)	co2
newa	61	ca(Serum calcium); bili, total(Bilirubin); glu(Serum Glucose); ap(Alkaline phosphatase); tp(Total protein); alb(Albumin); bun(BUN); sgot(SGOT); cret(Creatinine); sgpt(SGPT); phos(Serum phosphorus)	hcv; uric; hbs; ceah; bili, direct; rpr; ldh
WBC differential count	41	lymphs(Lymphs); mono(Monocytes); eo(Eosinophils); neutrophils(PMN); poly(PMN); baso(Basophils); blast(Blast)	abs blasts; nucleated rbc's; abs neutrophils; abs eos; plt morphology; abs lymphs; rbc morphology; abs polys; abs monos; abs basos