MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1299

The Clinical Research Data Repository of the US National Institutes of Health

James J. Cimino, Elaine J. Ayres

Laboratory for Informatics Development, National Institutes of Health Clinical Center, Bethesda, Maryland, USA

Abstract

The US National Institutes of Health (NIH) includes 27 institutes and centers, many of which conduct clinical research. Previously, data collected in research trials has existed in multiple, disparate databases. This paper describes the design, implementation and experience to date with the Biomedical Translational Research Information System (BTRIS), being developed at NIH to consolidate clinical research data. BTRIS is intended to simplify data access and analysis of data from active clinical trials and to facilitate reuse of existing data to answer new questions. Unique aspects of the system includes a Research Entities Dictionary that unifies all controlled terminologies used by source systems and a hybrid data model that unifies parts of the source data models and include other data in entity-attribute-value tables. BTRIS currently includes over 300 million rows of data, from three institutes, ranging from 1976 to present. Users are able to retrieve data on their own research subjects in identified form as well as deidentified data on all subjects.

Keywords:

Clinical research, Data repositories, Controlled terminologies and ontologies, Data reuse

Introduction

The United States National Institutes of Health (NIH) consists of 27 institutes and centers (ICs) dedicated to biomedical research for improving the health of the public. Most ICs are located wholly or in part at the main NIH campus in Bethesda, Maryland, just north of Washington, DC. All clinical research on the Bethesda campus is coordinated through the Clinical Center (CC), the 242-bed, 90-day-station hospital of the NIH.

Much of the information collected on human subjects at the CC exists in the Clinical Research Information System (CRIS). Many researchers also collect data in other locations, including IC systems, laboratory systems within the ICs, and even individual researchers' computers and notebooks. This data distribution causes two problems for researchers. First, CRIS is primarily an electronic medical record system, concerned with the tasks involved in patient care. Although it can support tasks related to research protocols, it is not designed to support research data analysis (e.g, queries for data across subjects in a clinical trial). Second, distribution of data across multiple sources complicates the ability of researchers to use the data to answer their research questions.

The clinical research data at NIH are also of interest to researchers besides those who are collecting them in the course of active trials. The US government mandates the sharing of clinical data that have been collected with federal funding, yet there is no mechanism at NIH to share the data that have been collected here for over half a century.

This paper describes the new NIH Biomedical Translational Research Information System (BTRIS), which has been providing investigators with access to clinical research data since July of 2009. Although still in evolution, BTRIS contains a substantial database and makes use of unique data models and terminology management techniques to merge data from multiple, disparate sources, to support active clinical trials and reuse of data.

Background

NIH Efforts to Consolidate and Reuse Research Data

NIH initiatives have recognized the need for a clinical research data repository to support reuse and sharing of data for clinical research, including the NIH Roadmap NECTAR project,[1] the CABIG project through the National Cancer Institute[2] and the current Clinical and Translational Science Awards (CTSA) program.[3] Based on researcher requirements as well as a business case, the NIH endorsed the concept of a clinical research data repository for aggregation and re-use of data collected at the NIH itself (as opposed to data collected by NIH-funded projects at other institutions). Initial funding for the project was received in 2007 and the development of BTRIS began in earnest in 2008.

The Columbia University Clinical Data Repository

The initial design of BTRIS has been based on experience with the creation of the Clinical Data Repository (CDR) at the Columbia University Medical Center in New York.[4] That system has accrued patient care data since 1988 from many different sources, including laboratories, pharmacies, radiology departments, order entry, and clinician documentation. Over the years, the repository has supported a number of systems for clinical care[5] and clinical research.[6]

All data in the Columbia CDR have been merged using a single, common relational data model that simplifies representation of disparate data while maintaining important distinctions and details.[7] The model makes extensive use of the "Entity-Attribute-Value" (EAV) approach, which allows specifications about the meanings of data to be stored with the data themselves, rather than being modeled as tables or columns in the data model. This method provides great flexibility for accommodating changes in data sources.[8]

The data in the Columbia CDR are represented with a single coding system, called the Medical Entities Dictionary (MED),[9] that unifies terminologies from all the sources providing data to the CDR. The MED provides a one-to-one mapping of individual concepts from each source and organizes them into a multiple-hierarchy ontology that provides definitional information about the concepts and supports data aggregation and inferencing functions.

Design Considerations

System Architecture

The first issue to be addressed in the BTRIS design was whether we should attempt to create a single, centralized repository (as was done at Columbia) or seek to create a federated system in which individual sources systems could be queried to provide data on demand. Although there are potential advantages to the federated model,[10] we quickly realized that most of the sources we would be dealing with (including archived repositories from defunct systems) would be incapable of participating in a federated design. We therefore proceeded to design a centralized repository.

Data Model

In designing the BTRIS data model, we considered the various advantages and disadvantages of traditional modeling approaches and the EAV modeling approach. We chose to take a hybrid approach in which data from disparate sources (for example laboratory test results from CRIS, from archives of the system that preceded CRIS, and from various IC systems) are analyzed and commonalities (such as the fact that all laboratory tests have primary times and results) are represented with columns in tables, while distinct source-specific differences are captured in EAV tables.

In addition to data collected from research subjects, we recognized that we would need a repository of information about the subjects themselves, including the protocols with which they are affiliated and the dates of those affiliations. While some of this information is available from CRIS, much is missing and some individual data may be "tagged" with protocol affiliations that do not match the CRIS database.

Data Acquisition, Extraction, Translation and Loading

The approach to adding data to the BTRIS database is a fairly typical *extraction, translation and loading* (ETL) process. Acquired data are dissected into their component elements and converted into a form compatible with the BTRIS database. They are then stored in the appropriate tables, rows and columns, according to a set of mapping rules. Sources include archived files, copies of active databases, and collections of transaction messages (typically in HL7 format). Sources may provide data on a one-time basis (from archives) or on a periodic basis (typically, daily or weekly). Mapping rules are created manually for each source, based on careful analysis of the source systems' documentation and the actual data provided (which do not always match the documentation).

Data Coding

Early in the project, we established a repository of controlled terminologies used by source systems to represent their data (for example, laboratory codes for tests and unique names for medications). This "Research Entities Dictionary" (RED) is based on the experience with Columbia's MED: each source term corresponds to a unique concept in the dictionary, with additional knowledge about the terms represented in hierarchical and non-hierarchical semantic relationships between concepts. The ETL process maps individual data elements to their corresponding entries in the RED so that the RED Codes can be stored along with the original data. Although the source systems do not use standard terminologies, concepts in the RED are being mapped to international standards to facilitate data sharing, including those contained in the US National Library of Medicine's Unified Medical Language System (UMLS).[11]

Data Reporting

Another key early decision in the system architecture was to determine that BTRIS users would perform retrievals themselves, using predetermined queries that could be tailored by the users for their specific needs. Given that a number of mature commercial "business intelligence" tools currently exist to support such capabilities, we evaluated several options and ultimate chose one to be our user interface to the database. System developers create query templates with general retrieval strategies (for example, to obtain demographic information or laboratory test results) and search filters (e.g., an age range, date range, type of laboratory test, or type of medication). Users provide values for the search filters when running the query to limit retrieval to specific subsets of data.

Queries were developed in response to a variety of perceived information needs. Some of these were identified in the original requirements gathering process for CRIS (see above), while others were developed through interactions with a BTRIS user group composed of interested NIH investigators.

Progress to Date

The BTRIS project officially began in January of 2008, with assembly of the development team in March, acquisition of sample data from several systems in May, and demonstration of a proof-of-concept prototype in July. The initial prototype used SQL Server (Microsoft, Redmond, WA) as the database management system, Terminology Development Editor (Apelon, Mountainview CA) for the RED, and Business Objects (SAP, Newton Square, PA) as the reporting tool.

Experience with the BTRIS prototype informed a number of changes in database design and user requirements, which led to the selection of Cognos (IBM, Armonk, NY) as the reporting tool. Based on the performance and user feedback with the prototype, approval for the project was secured in October. The first version of the actual BTRIS system was released on July 30, 2009 to PIs with active clinical protocols.

The Research Entities Dictionary

Each data source incorporated into BTRIS has one or more controlled terminologies that have been added to the RED. For example, the radiology system has a list of codes for procedures, while the laboratory has codes for tests, panels, organisms, antibiotics, specimens, and results. The RED currently contains 120,636 concepts with 155,321 hierarchical relationships (i.e., each concept has, on average, 1.3 parents).

Database Design

The BTRIS database contains five general sections, with information about investigators, protocols, subjects, the RED, and subject data. Investigator, protocol, and subject tables are related to each other in a typical manner.

Subject data are considered *measurable* (for data with normal ranges, such as laboratory tests), *substance* (for data with routes of administration, such as medications), and *general* (everything else). Data stored in *event* tables (for "things that happen", such as orders and procedures) and *observations* (for "things that report something", such as results and dosages given). Most events are associated with one or more observations. Each table has an associated EAV table (Figure 1); thus, there are a total of 12 tables for subject data (three event tables, three observation tables, and an EAV for each).

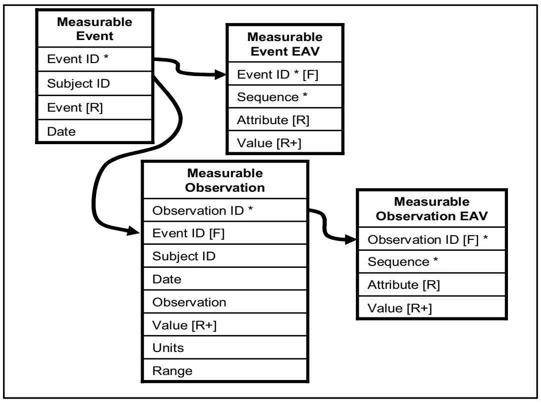


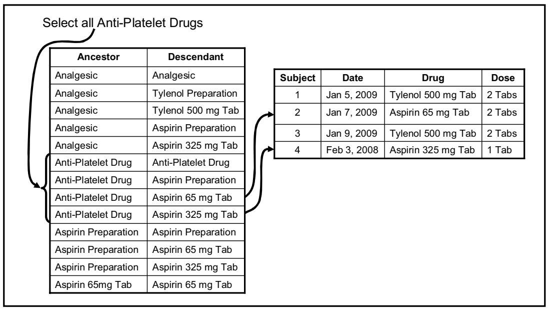
*Figure 1- Simplified view of part of the BTRIS data model. * = primary keys, [F]=foreign keys, [R]=elements coded in RED, [R+]=multiple column elements*

The six main tables include information common to all data from all the sources, such as subject ID, source, name, etc. These tables also include data elements that have been judged to be similar enough across sources that they form part of the merged data model. For example, each source provides one or more times (including date) for observations; for each source, we choose one of these to be the primary time. For radiology procedures, it is the time the procedures were performed. For laboratory tests, it may be the specimen collection time, or it might be specimen analysis time. Other times for each source are retained in the EAV tables.

Another example of the merged data model can be found in the way results of observations are treated. In the BTRIS model, results of observations that have normal ranges (such as laboratory test results and vital sign measurements) are all stored in the same table. Separate columns are used to store the result (or parts of the result) that are numeric, text, controlled terms, and comments. Controlled term results are stored as they appear in the data and as the corresponding RED Codes. Observations without normal ranges, whether nurse's notes, radiology reports, or discharge summaries, are all included as text results in the general observations table.

The RED is represented in two particularly interesting tables. One table relates every concept in the RED to one or more data sources, such that an identifier (such as a laboratory test code from the laboratory system or a medication name from the order entry system) can be uniquely mapped to a particular RED Code. This information is managed in the RED (see below) and exported to this table for use by the ETL process.

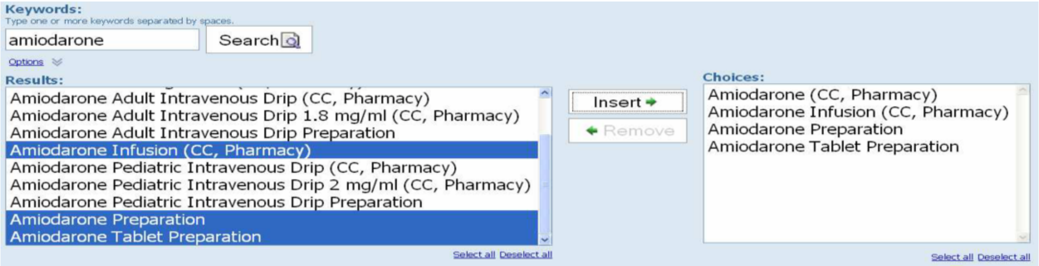
Another important RED table is the ancestor-descendant table. This table is derived from the RED hierarchy and supports class-based queries of the subject data. Figure 2 shows a simplified example of the use of the ancestor-descendant table for class-based queries. As of this writing, there are 1,214,646 ancestor-descendant relationships (that is, each concept subsumes, on average, ten concepts including itself).



Ancestor	Descendant
Analgesic	Analgesic
Analgesic	Tylenol Preparation
Analgesic	Tylenol 500 mg Tab
Analgesic	Aspirin Preparation
Analgesic	Aspirin 325 mg Tab
Anti-Platelet Drug	Anti-Platelet Drug
Anti-Platelet Drug	Aspirin Preparation
Anti-Platelet Drug	Aspirin 65 mg Tab
Anti-Platelet Drug	Aspirin 325 mg Tab
Aspirin Preparation	Aspirin Preparation
Aspirin Preparation	Aspirin 65 mg Tab
Aspirin Preparation	Aspirin 325 mg Tab
Aspirin 65mg Tab	Aspirin 65 mg Tab

Subject	Date	Drug	Dose
1	Jan 5, 2009	Tylenol 500 mg Tab	2 Tabs
2	Jan 7, 2009	Aspirin 65 mg Tab	2 Tabs
3	Jan 9, 2009	Tylenol 500 mg Tab	2 Tabs
4	Feb 3, 2008	Aspirin 325 mg Tab	1 Tab

Figure 2- Class-based query for "Anti-Platelet Drugs" using ancestor-descendant table.



Keywords:
Type one or more keywords separated by spaces.
amiodarone
Search

Options

Results:
Amiodarone Adult Intravenous Drip (CC, Pharmacy)
Amiodarone Adult Intravenous Drip 1.8 mg/ml (CC, Pharmacy)
Amiodarone Adult Intravenous Drip Preparation
Amiodarone Infusion (CC, Pharmacy)
Amiodarone Pediatric Intravenous Drip (CC, Pharmacy)
Amiodarone Pediatric Intravenous Drip 2 mg/ml (CC, Pharmacy)
Amiodarone Pediatric Intravenous Drip Preparation
Amiodarone Preparation
Amiodarone Tablet Preparation

Select all Deselect all

Insert

Remove

Choices:
Amiodarone (CC, Pharmacy)
Amiodarone Infusion (CC, Pharmacy)
Amiodarone Preparation
Amiodarone Tablet Preparation

Select all Deselect all

Figure 3- Example of text-based terminology searches in BTRIS. Note that each of the selected terms will be used to query against the Ancestor-Descendant table.

Database Content

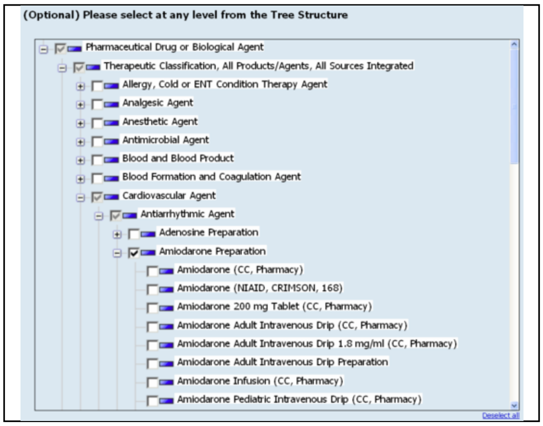
As of this writing, data have been accrued from three ICs: the Clinical Center (including CRIS and archived tapes of CRIS's predecessor), the National Institute of Allergy and Infectious Diseases (NIAID), and the National Institute of Alcohol Abuse and Alcoholism (NIAAA). Data types include demographics, vital signs, laboratory test results, medication orders, medication administrations, medication lists, and problem lists. Additional data to be added in the near-term include clinical documents (e.g., progress notes and discharge summaries) and procedure notes (radiology, pathology, etc.). Next steps include obtaining radiology image data and gene sequence and expression data from the National Cancer Institute (NCI). Thus far, there are over 86 million rows in event tables, 180 million rows in observation tables, and 855 million rows in EAV tables. Data are derived from 436,422 subjects, 196,036 of whom have been affiliated with one or more of the 9,055 protocols involving 3180 investigators, for a total of 393,447 protocol-subject affiliations.

Reports

Thus far, we have created three types of reports for identifiable data: summary reports (enrollment inclusion report for the Institutional Review Boards (IRBs)), detailed data reports (demographics, vital signs, laboratory test results, and medication data) and "list" reports (which create lists of patients, tests and medications that can be used as filters for other reports). Summary reports and detailed reports for all of these data have been created for de-identified access as well. When running reports, users interact with the RED in one of two ways. A text-based search produces a list of terms from which users may select terms (Figure 3). The users may also browse the RED hierarchy to select terms (Figure 4).

User Experience

Currently, BTRIS has been providing access to identified data for 30 weeks; 111 users have logged on to date. Of the 4,349 reports run thus far, 3,015 have been to retrieve laboratory test results, 303 to retrieve medication information, 275 to retrieve vital signs, and 310 to create summary reports for IRBs. De-identified data have recently been made available. User feedback been extremely positive. Additional reports have been requested; the current BTRIS model appears to be capable of supporting these requests.



(Optional) Please select at any level from the Tree Structure

- ☑ Pharmaceutical Drug or Biological Agent
  - ☑ Therapeutic Classification, All Products/Agents, All Sources Integrated
    - □ Allergy, Cold or ENT Condition Therapy Agent
    - □ Analgesic Agent
    - □ Anesthetic Agent
    - □ Antimicrobial Agent
    - □ Blood and Blood Product
    - □ Blood Formation and Coagulation Agent
    - ☑ Cardiovascular Agent
      - ☑ Antiarrhythmic Agent
        - □ Adenosine Preparation
        - ☑ Amiodarone Preparation
          - □ Amiodarone (CC, Pharmacy)
          - □ Amiodarone (NIAID, CRIMSON, 168)
          - □ Amiodarone 200 mg Tablet (CC, Pharmacy)
          - □ Amiodarone Adult Intravenous Drip (CC, Pharmacy)
          - □ Amiodarone Adult Intravenous Drip 1.8 mg/ml (CC, Pharmacy)
          - □ Amiodarone Adult Intravenous Drip Preparation
          - □ Amiodarone Infusion (CC, Pharmacy)
          - □ Amiodarone Pediatric Intravenous Drip (CC, Pharmacy)

Deselect all

Figure 4- Example of tree-based terminology search in BTRIS. A subsequent query will use the ancestor-descendant table to select all data with any of the 12 amiodarone medications.

Discussion

BTRIS is intended to encompass all clinical research data collected on subjects at the NIH. We began with some initial assumptions about design requirements and desired functionality and have proceeded rapidly through design, construction, and deployment. Our hybridization between column-oriented and EAV data models has allowed us to accommodate diverse data from disparate source in a way that supports aggregation across multiple data sources. The use of the RED and the EAV tables allows us to maintain the distinct aspects of data that are unique to their sources. The combination of the hybrid data modeling and the rich terminology representation provides a novel approach to the creation of a multi-purpose clinical data repository.

Thus far, BTRIS appears to be meeting the needs for researchers to obtain identified data on active clinical protocols. BTRIS is poised to provide access to de-identified

NIH data, across protocols, to analyze old data in new ways and ask new questions. We do not yet know how our researchers will make use of such functionality, but we believe that it will be in creative, unforeseen ways. BTRIS is designed to be flexible enough to meet a wide variety of such needs.

In particular, coding data with the RED supports queries that can aggregate or distinguish data as needed for the users' purposes. For example, instances of the administration of a 325mg aspirin tablet will be retrieved when a user requests that specific information, or all instances of the administration of any aspirin, any analgesic, any antipyretic, any platelet inhibitor, or simply any drug of any kind.

Elements of different data sources that have been stored in common columns in our six main tables have been carefully chosen to support what we believe will be the kinds of data aggregation that researchers are likely to want. For example, a user interested in the use of aspirin in a set of research subjects can request all instances of aspirin administration from the CRIS system, all instances of aspirin orders from the CRIS system, all instances of aspirin on a subject's medication list from the NIAID system, or a combination of any of these. Together with the flexible class-based queries supported by the RED, users have a range of ways to retrieve desired data.

The commercial reporting tool we have chosen (Cognos), allows us to create a variety of reports that appear to meet many of the users' needs, while giving them the power to tailor their queries and immediately obtain results. However, we fully expect that there will be information needs that will not be easily met with this approach. For example, a user may require a complex query that makes use of data from several main tables and EAV tables. In these cases, we may create specialized reports within Cognos, or we may perform retrievals, directly against the database, on the user's behalf.

In addition to the technical challenges, the development of BTRIS has required addressing a variety of policy issues that are beyond the scope of this paper. We have been successful at overcoming these issues in ways that address human subjects protection, privacy, data ownership, data access, and data sharing concerns. Solutions have required combinations of administrative and technical methods.

As with any system, BTRIS is faced with a number of potential limitations, particularly with regard to scaling (as we add image and genomic data), scope (as we add new data from institutes) and performance. Thus far, however, we have been able to address these issues and are not yet close to reaching capacity. BTRIS is still very much in development, as we add new reports for identified data, explore creative ways to reuse de-identified data, and expand to include new sources and types of data. The NIH has demonstrated deep commitment to creating a repository that serves all of the NIH community and eventually the research community at large, for the betterment of the health of humankind.

Conclusion

BTRIS addresses a long-standing need to consolidate clinical research data across the NIH for a variety of purposes. Our design includes a combination of novel approaches, development has been rapid, and it is already successfully addressing the information needs of NIH researchers.

Acknowledgments

This research was supported by the Intramural Research Program of the NIH.

References

[1] http://nihroadmap.nih.gov/clinicalresearch/overview-networks.asp

[2] von Eschenbach AC, Buetow K. Cancer Informatics Vision: caBIG. Cancer Inform. 2007 Feb 6;2:22-4.

[3] Zerhouni EA. Translational research: moving discovery to practice. Clin Pharmacol Ther. 2007 Jan;81(1):126-8.

[4] Johnson S, Friedman C, Cimino JJ, Clark T, Hripcsak G, Clayton PD. Conceptual data model for a central patient database. Proc Annu Symp Comput Appl Med Care. 1991:381-5.

[5] Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proc AMIA Symp. 1999:804-8.

[6] Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, Friedman C, Calman NS, Mostashari F. Syndromic surveillance using ambulatory electronic health records. J Am Med Inform Assoc. 2009 May-Jun;16(3):354-61.

[7] Johnson SB, Hripcsak G, Chen J, Clayton P. Accessing the Columbia Clinical Repository. Proc Annu Symp Comput Appl Med Care. 1994:281-5.

[8] Nadkarni PM. QAV: querying entity-attribute-value metadata in a biomedical database. Comput Methods Programs Biomed. 1997 Jun;53(2):93-103.

[9] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994 Jan-Feb;1(1):35-50.

[10] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):624-30.

[11] http://www.nlm.nih.gov/research/umls

Address for correspondence:

James J. Cimino, M.D.
Laboratory for Informatics Development
US National Institutes Clinical Center
10 Center Drive, Room 6-2551
Bethesda, Maryland, 20892 USA
James.Cimino@nih.gov