DW4TR: A Data Warehouse for Translational Research


Windber Research Institute, Windber, PA, USA
InforSense Ltd., London, UK
InforSense Ltd., Boston, MA, USA
InforSense Ltd., Shanghai, China
Walter Reed Army Medical Center, Washington, DC, USA
MDR Global, Windber, PA, USA

Abstract

The linkage between the clinical and laboratory research domains is a key issue in translational research. Integration of clinicopathologic data alone is a major task given the number of data elements involved. For a translational research environment, it is critical to make these data usable at the point of need. Individual systems have been developed to meet the needs of particular projects though the need for a generalizable system has been recognized. Increased use of Electronic Medical Record data in translational research will demand generalizing the system for integrating clinical data to support the study of a broad range of human diseases. To ultimately satisfy these needs, we have developed a system to support multiple translational research projects. This system, the Data Warehouse for Translational Research (DW4TR), is based on a light-weight, patient-centric modularly-structured clinical data model and a specimen-centric molecular data model. The temporal relationships of the data are also part of the model. The data are accessed through an interface composed of an Aggregated Biomedical-Information Browser (ABB) and an Individual Subject Information Viewer (ISIV) which target general users. The system was developed to support a breast cancer translational research program and has been extended to support a gynecological disease program. Further extensions of the DW4TR are underway. We believe that the DW4TR will play an important role in translational research across multiple disease types.

1. Introduction

Translational research requires integration of clinicopathologic data, biospecimen data, and molecular data, across multiple data collection and generation platforms [1–4]. It is also critical to make these data usable at the point of need. From the data management perspective, clinicopathologic data and molecular data have distinct features. The former are characterized by a large number of data fields, often with missing values due to failure to collect, non-existence or inaccessibility of the data. Molecular data, on the other hand, are characterized by a limited number of data fields, large number of records, and ever-evolving data types associated with the development of new technologies. While a translational research project typically collects and generates its own clinicopathologic and molecular data, public domain data are often used as well. Furthermore, temporal information needs to be properly managed and presented.

Many clinical and translational research projects use questionnaires as a clinical data collection instrument, and currently using data from Electronic Medical Records (EMRs) for translational research is gaining momentum as EMR systems replace paper-based medical records. Compared to EMRs, questionnaires are simpler as only questions useful to the specific research program of the disease would be included, thus managing such data does not require a comprehensive system covering all the human diseases. An EMR system cannot be used as a data management system for translational research, due to the transactional nature of the former and the reporting and analysis requirements of the latter. The requirements from the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) for protection of patients’ Protected Health Information need to be addressed.
Data warehousing as an important component of biomedical informatics, provides a way for data integration in clinical and translational research [2,5–13]. A Data Warehouse (DW) is an information repository for data analysis and reporting [14,15]. It is important that a data centralization system be developed, each serving a specific purpose. In the clinical field, data warehousing, as one of the major technological solutions in the area of translational research, relies on pre-computed data cubes whereas the latter directly queries a relational database.

Developing a comprehensive DW system to meet the needs of translational research faces many challenges. Based on our own experience, such challenges include; management of operationally important information, such as HIPAA compliance and PHI, data input from questionnaires with version control, data input from EMRs, biospecimen collection and banking, data de-identification, molecular images, multiple-platforms of molecular data, temporal information, and data ownership, etc. From the system development perspective challenges include; development and application of ontologies including controlled vocabulary and modeling of clinical data, molecular data, image data, and temporal data. In addition, development of graphical user interfaces for different levels of users is important, and data security and accessibility should be properly addressed.

Ontology development is very important to data integration and exchange; for the purpose of this paper we focus on its physical artifacts of terminologies/controlled vocabulary and information model/data model [20]. Many ontologies have been developed, each serving a specific purpose. In the clinical field,
commonly accepted standards/ontologies include; Health Level 7 (HL7) for exchange of clinical messages to support clinical practice and healthcare service [21]; Medical Subject Heading (MeSH) for medical literature indexing which has also proven useful for classification of biomedical entities [22]; Systemized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), as a comprehensive clinical healthcare terminology organized hierarchically [23]; and Digital Imaging Communications in Medicine (DICOM) for medical image annotations [24]. For biological research, Gene Ontology (GO), arguably the most successful biological ontology, is aimed to standardize the representation of gene and gene product attributes across species and databases [25]; Minimum Information About a Microarray Experiment (MIAME) for microarray-based technologies is generally accepted by the community [26]; Biological Pathway Exchange (BioPAX) is for pathway data exchange [27]. NCI Thesaurus provides reference terminology for many NCI and other systems covering vocabulary for clinical care and translational and basic research [28]. Systems have been developed to unify existing ontologies or provide ontology library services, for example the BioPortal by the US National Center for Biomedical Ontology [29], and the Ontology Lookup Service by European Bioinformatics Institute based on the Open Biomedical Ontology format [30]. The Unified Medical Language System (UMLS) developed by the US National Library of Medicine, aims to provide ad hoc linkages across life science ontologies [31]. A number of reviews are available including ones discussing development, implementation, and use of ontologies [20,32–34].

Temporal information is also very important in modeling health and disease development. We consider disease as a process, not a state. Breast cancer, for example, takes many years to develop before a lesion can be detected by mammography [35]. First considered by artificial intelligence researchers in the 1980s, temporal reasoning and temporal data query became a topic of interest in biomedical/medical informatics in the 1990s [36–38]. To enable "controlled language use" of the temporal information in healthcare, a European Prestandard CEN ENV 12381 has been developed [39]. Temporal information needs to be collected, modeled, and presented. Collection of temporal information can be done explicitly through questionnaires or data forms which record historical information, or abstracted through the time-stamps of data collected and analyzed over time. For the latter, temporal-abstraction and temporal-querying systems have been developed which have proven clinically useful in monitoring clinical disorders [40–42].

A data model comprising Parameters, Events, and Constants has been defined for a temporal query language [43,44], where Parameters are basically entities with values bearing a time-stamp, Events are external acts upon a patient which could be time-stamped or occur over a time interval, and Constants are non-temporal measures. There have been reports of a data management or DW system incorporating temporal information for clinical trials, time-series gene expression microarray experiment, and medical information events in a hospital setting [38,45,46].

The many challenges faced in developing a comprehensive DW system for translational research make this a difficult undertaking. In practice, different systems have been developed focusing on different challenges based on specific needs of supported program(s). For example, I2B2, Informatics for Integrating Biology and the Bedside [47], took a top-down "Enterprise" approach with an open architecture enabling integration of outside modules performing specialized functions that are needed for translational research [48,49]. It addresses a very important and specialized niche in the clinical informatics space, namely, providing researchers with sufficient access to clinical data to enable study design and cohort selection, while minimizing many of the patient privacy risks and concerns [50–54]. A central conundrum of translational research is that access to detailed clinical data typically requires proper informed consent and IRB approval; however, in order to design a cohort study or to determine if a particular project is even feasible, a researcher often needs access to aggregate statistics for key clinical attributes of interest at an early stage of the process. I2B2 overcomes these challenges by enabling researchers to design and execute queries against a de-identified data mart that return only aggregate counts; furthermore, results are subtly obfuscated to avoid identification by more sophisticated combinations of queries. Having been deployed for a number of projects across multiple sites, I2B2 is a mature software offering. However, it does not currently provide the level of detailed analysis necessary for clinical and translational research beyond cohort selection and study design [51–54]. STRIDE, The Stanford Translational Research Integrated Database Environment [55], focuses on a Clinical Data Warehouse (CDW) and supports a virtual biospecimen model for accessing several specialized tissue banks [13,56,57]. It also has a Research Database management system supporting multiple logically separate research databases. An EAV model is used for data storage and HL7 Reference Information Model is used for data representation. Its semantic layer consists of a framework supporting multiple terminologies. Some of the Clinical and Translational Science Award (CTSA) centers [61] also have data warehousing efforts.

There are also data warehousing efforts in the cancer Biomedical Informatics Grids (caBIG®) [62]. caBIG® is an information network enabling members of the cancer community to share data and knowledge based on a core Grid architecture. It was launched in 2003 by the NCI Center for Bioinformatics (NCICB) of the US NIH [63,64]. One of its data warehousing efforts is caBIO, a major component of caCORE (cancer Common Ontological Reference Environment) that was developed before the launch of caBIG® [65], enabling access of biomedical annotations from curated data sources in an integrated view. As one of the released products of caBIG®, the current version caCORE 3.x includes Enterprise Vocabulary Services for hosting and managing vocabulary, cancer Data Standards Registry and Repository (caDSR) for hosting and managing metadata, and the caCORE Software Development Kit [66]. Another data warehousing approach for caGrid is a semantic web-based data warehouse for creating relationships among caGrid models, accomplished through the transformation of semantically-annotated caBIG® Unified Modeling Language (UML) information models into Web Ontology Language (OWL) ontologies that preserve those semantics [67]. Still another caBIG effort, calintegrator [58], is a framework for data integration currently capable of integrating microarray data in caArray and imaging data from the National Biomedical Imaging Archive (NBIA). It has been applied to several projects, e.g., REMBRANDT [59] and CGEMS [60]. These efforts are best described as development projects using the calintegrator framework. None of these efforts in caBIG, however, enables users to dynamically interrogate the clinicopathologic data in a multidimensional manner by non-informatics specialists using an intuitive interface.

We began our DW development focused on a questionnaire-based system, but later realized that a patient-centric and expandable data model would be a better solution. We envisioned developing a system to support translational research in general and began by developing a system to support the Clinical Breast Care Project (CBCP) [68] but designed it to be expandable to support additional programs. This approach has been validated by our successful expansion of the system to support a second translational research program, the Gynecological Disease Program (GDP) [69]. It is currently under further expansion to support a third translational research program, a consortium effort headed by Thomas Jefferson University involving five organizations which is characterizing a set of 5000 invasive breast cancer cases. All these projects have multiple platforms for clinical and molecular studies.
Development of our system, the Data Warehouse for Translational Research (DW4TR), focused on three challenges in DW development. The first was the development of a pragmatic data model to satisfy the needs of questionnaire-based clinical data input. Despite the existence of a large number of clinical and biological ontologies, there are often gaps, in supporting a research project, between the available ontologies and the specific project needs [70,71]. Many groups therefore create their own ontologies [72,73]. Given the characteristics of clinical and molecular data, it is more challenging to develop a general and flexible system for clinical data; therefore we have made this our first priority [9]. In our case, both CBCP and GDP utilizes not only data elements covered by existing ontologies, but also detailed and specific data elements that are currently not covered. Comparing what we need to what is available, we concluded that completely adopting one or two existing ontologies for our DW development is not feasible (c.f. Section 5). Therefore, we developed our own framework ontology but referred to existing ones, to satisfy our immediate needs first, planning to map our ontology to existing ontologies at a later time to enable transportability.

The second challenge, we focused on was interface development. A user-friendly interface tailored to the non-informatics specialist is needed to ensure both the utility and adoption of this system. The interface that we developed is composed of an Aggregated Biomedical-Information Browser (ABB) and an Individual Subject Information Viewer (ISIV). ABB is RLAP-based, thus all the calculations and queries are performed at the time of use rather than relying on a pre-calculated data cube, enabling clinicians and scientists who are not informatics experts to dynamically interrogate any combination of the myriad of data elements stored in the DW4TR. The third challenge we focused on was the management of temporal information. We defined three temporal data types and applied them to data attributes to enable the use of the temporal information in the DW4TR from both interfaces. In addition to focusing on these three challenges, we also addressed a number of other challenges described earlier. For example, in managing de-identified clinical information of CBCP and GDP, we developed solutions to satisfy different requirements for data security and accessibility that is also described.

2. Methods

2.1. System design and development

Initially developed to support the immediate needs of managing the clinical, biospecimen, and molecular data for the CBCP, our design of the DW4TR to be flexible and extensible enabled subsequent expansion to support GDP. Both programs use questionnaires for collecting clinicopathologic information, making expansion of the light-weight production system straightforward.

2.1.1. The data model

The EAV model is used for the original raw clinical data which are subsequently hosted in the DW in an extensible data model reflecting the ontology, where the relationships among the attributes are presented in a meaningful way to the users through a hierarchical organization. The extensible data model is currently composed of the following components:

2.1.1.1. Patient-centric clinical data model. The clinical data model was developed to reflect the physician–patient interaction. The ontology framework was developed by a multi-disciplinary team composed of research physicians, biomedical informaticians, IT developers, surgeons, pathologists, and glycobiologists and was based on the standard clinical workflows and clinical experiences. The development not only referred to existing standards and ontologies mainly MeSH and SNOMED-CT but also to the NCI Thesaurus and caBIG VCDE [22,23,28,74], reflecting other reported methods and efforts [32,33,75]. This ontology is reflected on the physical data model, which is hierarchical and composed of a number of primary modules, e.g., “Diagnostic”, which are made up of secondary modules (e.g., “Biopsy”) and attributes. A secondary module is further composed of tertiary modules and attributes, and so on. An attribute is a fine-grained object that is composed of the name, the data type, and the temporal characteristics (see below). A simple attribute is often called a data element, e.g., sex. A complex attribute is composed of multiple facets, e.g., exercise is composed of frequency, intensity, duration, and type. This way, each data element of a study is represented in the model either by an attribute or by a sub-module. The sub-modules are structured hierarchically to reflect the ontology.

2.1.1.2. Specimen-centric molecular study data model. Molecular (biochemical, genomic and proteomic) analysis are performed on collected specimens using multiple experimental platforms. Every experiment is done using biospecimens, including body fluids, solid tissues, and their derivatives. Thus, a specimen-centric data model is a natural choice for these types of data, which are connected to the clinical data model hierarchy as a series of sub-modules. We initially focused on immunohistochemistry (IHC), Fluorescence In Situ Hybridization (FISH), and gene expression microarray data and adopted existing clinical or molecular data standards and guidelines [26,76,77]. The storage and retrieval of high throughput molecular data poses a number of IT challenges. A result file from a single microarray can be tens of meabytes, and it is not uncommon for a single experiment to make use of hundreds of such arrays. Common approaches for handling this type of data include: organizing output files into a hierarchical file system, loading of raw or partially transformed data into a DW, or a hybrid approach of maintaining certain metadata in a database but referring back to original output files with file pointers. Given the large number of molecular modalities used in our current and future research, no single method was considered suitable, therefore our system was designed to take an “all of the above approach” that makes use of flexible data processing pipelines. With this approach distinct pipelines can be created across or even within assay types and all of the storage methods discussed above can be supported. This approach has enabled us to optimize storage of each assay group individually, as well as provide a mechanism for handling changing technology and legacy data.

2.1.1.3. Temporal data model. Critical to human disease studies is the proper representation of temporal information. We define three types of temporal data: (1) Static, which is data with no temporal dimension, e.g., ethnicity. (2) Event, which is associated with a specific time point, e.g., a surgical procedure. (3) Interval, which is associated with a starting and ending time point, e.g., a course of medication. Each attribute in the data model is tagged with a property that indicates the proper representation of temporal information and analysis.

2.1.1.4. Medical image data model. Medical imaging is an important part of clinical practice and a source of data for various studies. We initially focus on digitized and digital mammograms, which follows the DICOM standards [24] and we de-identified the images using DICOM Anonymizer Pro. The data files associated with these images are large and they impose a challenge in data management, e.g., storing those image files in the database of the DW will drastically impede the system performance. Besides, analyzing images typically requires specialized software. Thus, in the DW4TR, we store the image files on a file server; the file locations are then...
stored in the DW and logically linked to the patient. The annotations and a thumb nail of the image are also stored in the database to enable efficient search of relevant features. We intend to apply the same principle to other images.

2.1.1.5. Relationships between questionnaires and the data model. The data for both CBCP and GDP studies are collected using questionnaires each having multiple questions. Each question has one or more data elements. The data elements could contain values and temporal information. If a data element contains a value, it is mapped to an attribute(s). Some questions ask about the same event at different time points, and these questions are mapped to the same attribute with corresponding temporal information. There are also different questionnaires collecting the same information, e.g. for different studies, and they have been mapped to the same attributes as well. Finally each attribute is mapped to the data model as described above.

2.1.2. User interface development

The requirement to develop an intuitive, user friendly system that could be used directly by clinicians and researchers as opposed to being a specialist tool useful only to the technically savvy, led to our decision to take the dimensional modeling approach. While traditional query interfaces are considered intuitive by IT professionals, we found that the dimensional modeling approach was generally preferred by our target demographic. A number of criteria were considered in designing the physical data structure to support the clinical data warehouse, the first of which was the ability to handle a large number of dimensions. A custom binning capability that is convenient to use, was an important feature requested by the users. As is the case with any longitudinal study, the temporal dimension of data is critical, and therefore proper handling of the temporal dimension was another important concern. Efficient handling of sparsely distributed data and support for multiple terminology sets were also important criteria. During such comprehensive communications with the end users, we concluded that all the desired analyses could be categorized as either an aggregate data analysis or an individualized data analysis. Thus, we designed a ROLAP-based ABB for the former, and an ISIV for the latter.

2.2. Implementation

2.2.1. Development environment

Two DELL Windows server systems, one for the application and the other for the database, were used. Both servers have two Quad Core Intel® Xeon® processors and 8 GB RAM. The database server hard drive is 1 TB and for the application server 0.5 TB, both are in the RAID 5 configuration. Development is based on the Oracle RDBMS 10 g, using the InforSense platform that enables a series of analyses to be performed in an analytical workflow environment with each analysis step represented as a node in the workflow [78]. The user-interface is Java based and web accessible. The production system has the same configuration. The DW4TR also supports a one-server system configuration.

2.2.2. Data model and interface implementation

Our physical data model is composed of two distinct elements, an attribute repository, and an attribute ontology. The attribute repository stores data in the EAV model. The attribute ontology is a separate hierarchical data model capturing and providing structure to the relationships among the attributes that have been collected so that the ontology (logical patient-centric or specimen-centric data model) are physically positioned in a hierarchical organization of attributes for presentation in a meaningful way to the users of the system.

In implementing the data model, first the source questionnaires are reviewed and a logical mapping of the data attributes onto the newly designed patient modules is developed. This process requires the definition of the data variable type and other metadata for different data attributes, and often requires splitting or merging data elements in order to accommodate the difference in requirements of capturing the data versus analyzing the data.

Parallel to the development of the physical and logical data model was the development of the end user interfaces. Our system provides users with two distinct applications: the ABB, and the ISIV. These Java tools are entirely web based with zero client side foot print, and do not require any additional browser plug-ins. They provide users with a highly interactive and dynamic experience, making use of Asynchronous JavaScript and XML (AJAX) techniques for asynchronous rendering and updating. Both tools are deployed within the InforSense web portal environment, which provides the services used for data source connection pooling, user authentication and authorization, as well as web session management.

2.2.3. Data loading

Multiple sources of data in different format are loaded into the DW4TR, including direct Oracle–to–Oracle load from the Laboratory Information Management System (LIMS) and flat files [12,79]. Data loading is performed through a standard process referred to as Extract, Transform, and Load (ETL) [14,15].

2.3. Data sources

The current DW4TR is host to clinical data from CBCP and GDP. Subjects are enrolled into the programs via HIPAA compliant, Institutional Review Board (IRB) approved protocols at multiple participating clinical sites, and these protocols explicitly specified WRI as the data integration center for the programs. Clinical data are collected through multiple questionnaires, and biological specimens are collected, processed, banked, and analyzed using genomic and proteomic experimental technologies.

CBCP and GDP happen to be two distinct types of programs in administrative structure, clinical dataset, and tissue banking. They have different data security and access requirements, commanding distinct solutions in the DW4TR. CBCP is a centralized program using one set of the questionnaires, one LIMS, one Tissue Bank, and one data warehouse (the DW4TR). All the participating organizations use approved master IRB protocols from the Walter Reed Army Medical Center (WRAMC), with minor adaptations to the requirements of the local institution for local IRB approval, which is subsequently approved by the IRB of the US Army Medical Research and Materiel Command that contracted the Henry Jackson Foundation to manage the CBCP. Proper paperwork was done to enable centralized tissue banking and data warehousing at WRI. The IRB approved questionnaires contain certain date information such as date of birth which patients are consented to, thus the CBCP clinical data is a “Limited Data Set” per HIPAA definition. The clinical data and biospecimens are de-identified, and each subject is represented by a CBCP number. The link table is securely maintained at the office of the CBCP site PI which is strictly accessible only to the PI or his/her designee at that site. When clinical data Quality Assurance (QA) problems are identified in any participating institutions, they are all reported to WRAMC and the WRAMC clinical data team coordinates the QA resolving process. As of February 2011, over 5000 subjects have been enrolled in the study using two major questionnaires to collect up to 799 elements of clinicopathology data per subject. More than 43,000 specimens (including aliquots) have been collected and genomic and proteomic experiments have been conducted using these specimens.
GDP, on the other hand, is a consortium program constituting of nine clinical and research organizations with WRAMC serving as the lead institution. Each participating organization follows its own IRB-approved protocol allowing only “Safe Harbor Data Set”, and each site manages and owns its own questionnaire-specific datasets. All the captured data are in a de-identified form, and each subject is represented by a GDP study ID. Each site PI maintains the link key to the identity of the subject enrolled at that site. All the de-identified data are then sent to WRI, and eventually loaded into the DW4TR. Aggregated subject information can be shared across the consortium members, but the detailed individual subject information in the DW4TR is only accessible to the originating clinical site. In GDP, about 500 subjects have been enrolled in the study. A dozen questionnaires were used to collect more than 7600 elements of data focusing on surgical procedures, pathology, family history, psychology, food intake, etc. Currently GDP is consolidating the questionnaires being used reducing the data elements to less than 1400.

3. Results

3.1. Overview of the DW4TR and the relationships between involved data types

Key to the development of the DW4TR is the understanding of the complex relationships between data collected or generated with multiple platforms. Fig. 1 illustrates our understanding of such relationships in a 3-dimensional patient data space which we can use to represent our data models.

Fig. 2 is a conceptual view of the DW4TR, as a hybrid system, that integrates the data collected or generated by the supported programs, and federates most of other data needed in the studies. Blocks with solid lines are the areas already developed or currently under development, and blocks with dashed lines are areas for future development. The whole DW is composed of a data tier, a middle tier, and an application tier. In the data tier, an extended EAV model is used for the raw clinical data. The middle tier is based on a patient-centric, modularly-structured clinical data model (including a medical image data model), integrating a specimen-centric molecular data model, and a temporal data model. The clinical and molecular data models are designed to be extensible. The application tier is composed of the ABB and the ISIV, and other utilities. Additional applications will be developed to federate and present other data needed but not generated by the supported translational research programs.

3.2. The ontology and the extensible data model

As a bottom-up approach, we developed a pragmatic lightweight ontology using controlled terminology both from existing ontologies and supplemental ones we developed to support the detailed program needs. The relationships between these terminologies in the ontology are reflected on the physical data models described below.

The patient-centric clinical data model was developed to reflect the physician-patient interaction. As illustrated in Fig. 3, the model is composed of six major modules: Medical History, Physical Exam, Diagnostic, Treatment, Outcome, and Scheduling and Consent. Each of these modules is composed of attributes and sub-modules; e.g., the Diagnostic module is composed of Biopsy, Imaging, and...
3.2.3. Example data module definition

Table 1 is a simplified illustration of three defined data modules. Thus, “Ethnic Group” is a simple (single attribute) data module allowing multiple values for a given subject, and is “Static”. “Alcohol Use” is a complex data module with three facets, and covers an “Interval”. “Surgical Proc(edure)” is a module similar to “Alcohol Use” but is an “Event”.

3.3. DW4TR interfaces—ABB and ISIV

3.3.1. ABB

This interface allows the user to analyze the data by dynamically creating a multidimensional pivot view. The rows and columns of the table can be data elements of either a numerical or categorical nature, in a flat or hierarchical structure. Although the view is a two-dimensional table, the user can effectively explore data of theoretically unlimited dimensions by simply building multiple levels of a hierarchical data structure in both the rows and columns, and expanding them. The number of dimensions (levels of hierarchy) is only practically limited by the performance of the system (see Section 3.6).

One type of molecular data can be directly accessed through ABB. Table 1 is a simplified illustration of three defined data modules. Thus, “Ethnic Group” is a simple (single attribute) data module allowing multiple values for a given subject, and is “Static”. “Alcohol Use” is a complex data module with three facets, and covers an “Interval”. “Surgical Proc(edure)” is a module similar to “Alcohol Use” but is an “Event”.

3.3.2. Temporal information definition and presentation

All attributes are tagged with one of the three defined temporal properties. All temporal information, including static information, can be viewed in the ISIV (c.f. Fig. 7). Within the ABB users have the ability to define “time filters” for the different columns in the display (c.f. Fig. 6).

3.3.3. Example data module definition

Table 1 is a simplified illustration of three defined data modules. Thus, “Ethnic Group” is a simple (single attribute) data module allowing multiple values for a given subject, and is “Static”. “Alcohol Use” is a complex data module with three facets, and covers an “Interval”. “Surgical Proc(edure)” is a module similar to “Alcohol Use” but is an “Event”.

3.3. DW4TR interfaces—ABB and ISIV

3.3.1. ABB

This interface allows the user to analyze the data by dynamically creating a multidimensional pivot view. The rows and columns of the table can be data elements of either a numerical or categorical nature, in a flat or hierarchical structure. Although the view is a two-dimensional table, the user can effectively explore data of theoretically unlimited dimensions by simply building multiple levels of a hierarchical data structure in both the rows and columns, and expanding them. The number of dimensions (levels of hierarchy) is only practically limited by the performance of the system (see Section 3.6).

Using the ABB, simple analyses such as mean, standard deviation, counts and percentages can be performed. Views on subject counts or specimen counts can be created. The results of interest can be printed, or exported to an Excel spreadsheet for additional analysis using other specialized software. A subject set of interest can also be saved as a cohort, for example the triple-negative Caucasian American breast cancer group shown in Fig. 4B, and used in a subsequent study or analyzed with a different application. The cohort can be analyzed using the ISIV described in the next section, or rendered to additional complicated analysis that the InforSense Analytical Workflow platform supports (not shown). The cohort data can also be exported to an Excel spreadsheet for analysis by specialized software outside of the DW4TR.

ABB offers many other features including a unique custom data binning capability. Multiple binning methods can be used, either automatically or manually, to create bins to explore data elements that contain either numerical or categorical values. For different studies, a user may need to use the same information in different ways. As shown in Fig. 4 (Panels C and D), the ethnic group attribute contains more than a dozen different possible values from CBCP and GDP. If the user wants to concentrate on Caucasians, African Americans, and Asian Americans, he/she can create three bins for them, and place the rest in “Other”. A histogram of the binning can then be viewed. Custom data binning can be done with categorical data as well as numerical data, and in our practice viewing histograms helps to adjust the binning strategy to ensure that there are a sufficient number of records available in each bin.

One type of molecular data can be directly accessed through ABB is IHC. Fig. 4B shows data for three molecules (ER, PR and HER2) which are routinely assayed in the CBCP, for which there are established clinical standards for clinical IHC [76,77]. Two other molecules, Ki67 and p53, are also routinely assayed though there are no current clinical standards for these molecules. In CBCP, all five molecules are characterized by protein expression assays by IHC, and HER2 is further characterized with gene copy number ratio assessed using FISH as needed. Both IHC and FISH data types are supported in the production system. The IHC submodule contains Result, Percentage, Intensity, and Staining Pattern for nuclear proteins (ER and PR) or Result and Score for membrane-bound proteins (HER2). The FISH submodule contains Result and Ratio. Both submodules belong to “Pathologic Results” of “Biopsy” of “Diagnostic” in the data model. Specific business rules are applied in
the ETL process, for example the final HER2 expression result is determined by IHC Score (0 or 1+ as negative, and 3+ as positive), but when IHC Score is ambiguous (2+) the Ratio of the FISH assay is used, with \( \leq 1.8 \) as negative and \( \geq 2.2 \) as positive, and an intermediate ratio is considered as ambiguous.

A proof-of-principle model of microarray-based molecular data is in our test system; this feature is the only feature described in Section 3 that has not yet been applied to the production system. The MIAME standards [26] are followed with some modification, for example, our “sample ID” eliminates the need for subject and...
specimen information. The metadata are modeled, including sample ID, bio-molecule information (lab protocol, 260/280 ratio, RIN, etc.), array information (platform, array ID/batch number, etc.), experimental information (project name, PI, lab operator, date, etc.), and results (call rate, 3'5' ratio, Pass/Fail flag, data file location, etc.). Metadata are accessible via ABB for selecting microarray data of interest in a cohort defined by clinicopathologic characteristics, but we saved the array raw data files in a file server due to their size and the fact that their analysis typically requires specialized software that read in the raw data file directly. Fig. 5 shows an example how a set of microarray data can be identified from previous experiments performed on samples from a cohort of interest. This information may lead to a secondary use of the microarray data, or help the researcher to design a new experiment.

ABB also contains a utility “Time Filter” to apply temporal information in the analysis of aggregated information. The time filters are used to limit or constrain the data that are returned much like any other type of filter criteria. Time filters can be created based on absolute dates, patient age, or relative to other events. Fig. 6 shows one example how “Alcohol Usage” was analyzed for CBCP subjects when they were at different age ranges. “Alcohol Usage” is an Interval temporal data type with values self-reported by subjects covering different periods of time in patient’s life. Note that direct
3.3.2. ISIV

This interface is for viewing and analysis of detailed subject information, most importantly the temporal-related information. It takes subject IDs as the input either by direct entry, or selecting from a pre-defined cohort. Events and Intervals of interest can be selected, and applied to the whole subject set. Information can be viewed independently for every subject, or aligned across the subject set on Events of interest to enable comparison between subjects. Static information can be displayed as well, so all the information about the subject can be studied. An example screenshot of ISIV is shown in Fig. 7.

Both the ABB and ISIV require minimal training before one can start to use them. The average training time for a clinician or scientist is 15 min, and the user can learn more features by using the system. Interestingly, for a high school student the average training time is only 10 min.

3.4. Direct data extraction from the database

The raw data in the DW4TR can be directly queried from the database, as a complete or partial dataset depending on the research need. The data can be presented in either a flat file structure or in Excel spreadsheet, for use with specialized data analysis software such as SAS, SPSS, or R. Such queries have been developed for and applied to both CBCP and GDP data. Direct data extraction requires a database administrator and is not the focus of this paper, thus is only briefly described here.

3.5. Extensibility of the DW4TR

The DW4TR was developed to support the CBCP with a vision to support translational research in general. This extensibility was tested and validated by expanding the system to support the GDP. Its flexibility was tested by our successful revision of the model commanded by the consolidation of GDP questionnaires and data elements. A number of disease-independent attributes and submodules developed for the CBCP implementation were re-used for the GDP instance, for example demographics and elements of past medical history. Many other attributes/submodules were modified to satisfy the requirements of GDP, for example alcohol usage. For GDP-specific data elements, additional modules were developed. In addition, the GDP contains a psychology questionnaire and a food questionnaire that are program-specific but not disease-specific. These were developed and structured to the overall patient-centric clinical data model, and could potentially be re-used for future programs. In addition, the GDP has different data security and accessibility requirements, and we have developed additional mechanism for it as detailed in Section 4. With the experience gained in supporting the GDP, we are confident that we will succeed in the current further extension of the system to support the third translational research program for the consortium led by Thomas Jefferson University.

3.6. System performance

To evaluate the performance potential of the DW4TR, we conducted a stress test using the CBCP data in a one-server system configuration. The test was done from a user’s perspective, by
measuring how long it takes for the data to completely display in a hierarchical view for the first time—note that subsequent display of the same data is much faster. The view is composed of three levels of Pathology Category, Ethnicity, and Body-Mass Index (BMI, with custom-binning) in the rows and record counts in the columns. Starting from the then patient number of 4234, we artificially replicated the records in the database to 8468, 16,936, 33,872, and 67,744 without performing a database tune-up other than Oracle table analysis. It took one second or less, for the level 1 (Pathology Category) and level 2 (Ethnicity) data to be displayed in all these tests. For level 3 (BMI), it took 2, 4, 12, 39, and 134 s respectively to display the data when the number of records doubled stepwise from 4234 to 67,744. Given that the current CBCP annual subject enrollment is about 600, it will take another 20 years before 16,936 subjects are enrolled when 12 s are needed to display 3 levels of data. We expect that the system performance will be further dramatically improved with hardware upgrades and Oracle tuning.

Fig. 6. Use of temporal information in ABB to study data of interests. (A) Screenshot showing the use of the ‘Time Filter’ to define the starting and ending time for study, with starting time already being configured but not the ending time. (B) Configuring the ending time based on Age, but it can also be based on Absolute Date or an Attribute. (C) Example showing the alcohol usage history across the CBCP subjects. The first column shows the usage frequencies of subjects when they were between 18–25 years of age, and the second column shows the usage frequencies across the whole population. Additional age ranges can be configured to show the corresponding alcohol usage habit.
3.7. Example use cases

The DW4TR has been playing an important role in supporting translational research across user groups in both WRI and WRAMC. Here we report four example use cases to demonstrate the utility of the system.

3.7.1. The DW4TR enables clinicians to study the integrated data directly

When first exposed to the DW4TR, CBCP clinical users appreciated the simple interface which gave them their first opportunity to directly study the integrated clinicopathologic and biospecimen data. They were happy to see that the results shown on the ABB confirmed their qualitative clinical observations. In one short session of less than 1 h, the program PI identified several lines of evidence which subsequently led to an abstract accepted to the San Antonio Breast Cancer Symposiums [80]. The system has since been further improved in both the data model structure and the user interface.

3.7.2. The DW4TR enables enhanced cohort and biospecimen selection

Before the DW4TR was developed, cohort and biospecimen selection for any research project was a major task involving multiple data sources including: Core Questionnaires, Pathology Checklists, and IHC assay reports, in the form of an Oracle database, a Microsoft Access database, Excel spreadsheets, and occasionally hard copy reports. Many manual steps were involved, and the procedure was not standardized. The selection of biospecimens was especially challenging when there were temporal restrictions, for example identifying blood samples drawn in a subject cohort before certain kinds of invasive procedures were performed. This requirement alone could take an experienced researcher several days since multiple questionnaires are involved and a number of surgical procedures could be performed on one subject and that there were 15
different kinds of such procedures. Using the DW4TR interface and additional utilities, the time needed to perform such tasks has now been reduced to minutes. Currently any cohort and biospecimen selection can be done in hours instead of days or weeks.

3.7.3. The DW4TR serves as a research environment for risk factor assessment

The DW4TR provides an effective research environment for cancer risk factor analysis. Several studies examining breast cancer risk factors have been done using this system [81–85], and several new studies are underway.

3.7.4. The DW4TR facilitates virtual experimental studies

Virtual experimental studies can be performed as integrated information associated with completed research projects is available. Using the information stored in the DW4TR and the file server, we identified the gene expression microarray data from three previous projects, on two types of specimens from three subject groups. From these datasets we were able to create two virtual experiments, and both studies were accepted to leading scientific conferences [86,87].

4. Security and accessibility

Data security and access control are important factors in the design of the DW4TR, particularly as our system can be expanded to include multiple research studies. The use of clinical data in translational research is governed by HIPAA compliant IRB approved protocols, and typically cannot be unconditionally shared. Thus we have designed and developed a set of data security and access solutions to satisfy such requirements.

The first layer of security for the system is data de-identification. Prior to being loaded into the DW4TR, PHI is stripped away from the collected data. At this point subjects will have been assigned a unique research subject ID for linking together subject information across multiple sources. The next level of security makes use of the “Virtual Private Database” (VPD) feature of Oracle, which allows for row level access control within the DW. While one of the strengths of the system is the ability to perform cross-study analysis, in many instances it is necessary to restrict data visibility among different users of the system. In our security model, all users are assigned roles which dictate the data they are permitted to access. The VPD system is designed such that every data element in the DW is tagged with a study identifier that is associated with the different user roles, enabling data access control. The third level of security is at the application tier. All access to the DW occurs through a password protected web portal that is served by an application server. The final layer of security is a network firewall and Virtual Private Network (VPN) for which the 168-bit encryption Secure Sockets Layer VPN from Juniper Networks Inc. is deployed. All components of the system, including the DW4TR and the web portal, are protected behind the local network firewall and are only accessible from either within the Institute itself, or by first establishing a VPN connection to the network.

These security and access control measures were developed based on the needs of the supported programs, and can be selectively applied to individual programs. For the two programs currently supported by the DW4TR, CBCP protocols allow data access by all CBP researchers, and VPN access to the DW4TR is granted with written permission of the CBP program PI. For GDP, aggregated biomedical information is allowed for use by all the GDP researchers, but the detailed subject information is only allowed to the originating clinical sites unless additional paperwork is completed. Thus, we have designed and developed a role-based site-specific privilege system to enable GDP users to use ABB and the ISIV. DW4TR can also be configured to disable functions and utilities of choice. Currently at the request of the GDP PI, we have disabled the ISIV functions for GDP users so that the GDP data are only available through ABB to GDP users for aggregated information. When a GDP researcher identifies a cohort of interest, additional data use agreements will be completed to request individual subjects’ information from the corresponding clinical sites, for approval by the involved sites and the GDP program PI. With a written request from the GDP program PI, WRI will supply the detailed data from the DW4TR to the researcher.

5. Discussion

We have taken a bottom-up approach to develop a DW4TR to support translational research programs. The system is extensible, integrating internally generated clinicopathologic and molecular data based on a light-weight ontology, with a patient-centric, modularly structured clinical data model supplemented with a specimen-centric molecular data model. Two interfaces have been developed targeting non-informatician end users, with the ABB supporting aggregated data analysis and ISIV supporting single subject information analysis. The system enables study cohort and biospecimen selection and is capable of handling temporal relationships between clinical data and biospecimen data collected at multiple points in time.

The development of our data model relies on using or creating ontologies for clinical and molecular data types. We will focus this discussion on clinical data ontology. In fact, for molecular data types, the scale of work is much smaller and the existing ontologies are mostly adequate. For example, we found that the MIAME standard met our needs for modeling gene expression microarray metadata. Our analysis indicated that existing ontologies for clinical data cannot be readily adapted to satisfy our specific research needs, due to gaps between existing ontologies and our needs as well as overlaps and inconsistencies between existing standards and ontologies [70,71]. For example, the CBP Pathology Checklist contains 372 data elements including 131 breast pathology conditions. When the data form was developed and revised, the AJCC (American Joint Committee on Cancer) guidelines and ASCO/CAP (American Society of Clinical Oncology/College of American Pathologists) guidelines were followed; these guidelines, like many other guidelines, have evolved over the years [88–91]. Many data elements present in the CBP lack standard descriptions, for example
for many benign diseases. Some biomarkers used in the CBCP (Ki67 and p53) do not even have a standard clinical IHC protocol. Therefore, descriptions or definitions of such data elements were developed based on medical textbooks and pathologist’s clinical practice for the Pathology Checklist, which we modeled ourselves.

There are classifications in existing standards/ontologies that we find cumbersome or which do not meet our needs [92]. We show as an example in Table 2 the classifications of “Breast Neoplasms” by two existing ontologies. In MeSH, it is classified under “Skins Diseases”. This classification (named “MeSH1” in Table 2) is not available in the SNOMED-CT classifications under either Skin Diseases or Connective Tissue Diseases. In MeSH, “Breast Neoplasms” is also classified under “Neoplasms” (in a way we agree with, “MeSH2” in Table 2). However, in the two classifications, “Breast Neoplasms” have different codes, which may cause a problem in the practical applications of data modeling. In SNOMED-CT, the term “Neoplasm of Breast” is used, and there are several classifications leading to it with relatively long paths. We show two of them, “SNOMED1” and “SNOMED2” in Table 2. In our pragmatic approach to support the CBCP, applying such a complete classification is cumbersome. In addition, reconciliation of discrepancies among the available ontologies and our understanding of the problem, for all the data elements in the CBCP, is a major task and not one of our immediate project needs. Therefore, we have developed a classification path for “Breast Cancer”, which is more lightweight but adequate to our needs (see “DW4TR” in Table 2).

Finally, some questions in the questionnaires of the two programs we currently support do not comply with existing standards, but we are still required to model them to enable proper storage and presentation of the data. Such problems have been reported to program PIs and were taken into consideration, for example in the consolidation of the GDP questionnaires. In addition, existing ontologies are not static, for example MeSH is currently updated weekly and SNOMED-CT is updated monthly. Thus, after analyzing what is available and what we need to support, we determined that our best approach was to design a light-weight ontology incorporating existing standards and ontologies where possible and to map our classification system to existing ones at a later time. The mapping table(s) will serve as the buffer between the two moving targets which will ensure transportability and un-interrupted use of the system. We will also supplement those standards with new ontologies and common data elements that we define.

The DW4TR currently supports limited number of molecular data types and we will continue to develop data models to support other molecular study platforms, including SNP microarray and array CGH data. We expect that the method developed for gene expression microarray will in general be applicable to these array-based technologies. We will explore support for RT-PCR, tissue microarrays, mass spectrometry, and Next Generation sequencing technologies. These will depend on the needs of the users and programs that we support.

Currently a prototype has been developed to enable access to medical and molecular images through the ISIV, including digital/digitized mammograms and gene expression microarray images. A thumbnail of the image is stored in the database of the DW with searchable annotations. The full-size image is stored in a file server instead of the DW4TR database for performance considerations. We plan to follow the DICOM standard for medical images and will explore how to best annotate molecular images. Additional images we need to support include, pathology H&E slide images, biomarker IHC images, and tissue microarray images. Note that all the medical images have to be de-identified for research purpose.

Currently the DW4TR does not support EMR for data input. We have a stepwise plan to begin using EMR records for CBCP subjects and we are looking at subject de-identification, data extraction, and data model expansion. Although we do not plan to take a systematic “Enterprise” approach, we will learn from such approaches as used by I2B2 and others. When EMRs are used as an input, a number of functions need to be integrated into the system including subject de-identification and information extraction, for which the I2B2 development teams have made good progresses [50,54,93]. While I2B2 clearly plays an important role in the study design phase including addressing de-identification and natural language processing, it was not designed to be a comprehensive solution that can provide the level of detailed access necessary for later stage research [50–54]. The ABB described here provides the same ability to generate aggregate statistics for population stratification and cohort selection, but by doing this in a web-based pivot table interface it also facilitates trend identification and discovery in a way traditional query interfaces cannot. From a data modeling perspective, the ABB also allows for more sophisticated handling of complex, multi-faceted attributes whereas I2B2 only offers aggregation around the single pivot point of the patient. I2B2 makes good use of ontologies, but has not yet developed an ontology with the level of detail described here [54,94]. Furthermore, by design I2B2 does not enable access to row level clinical data and detailed clinical timeline provided by the ISIV [54,94].

In conclusion, we have developed the DW4TR to support a breast cancer translational research program and then expanded it to support a gynecological disease translational research program. It is based on a light-weight ontology and equipped with an interface capable of handling temporal information, designed for use by non-informatician specialists including clinicians and laboratory scientists. The system supports both in silico and in vitro studies. With its proven extensibility, we believe that the DW4TR will play an important role in translational research across multiple disease studies.

Acknowledgments

We thank Mr. Anton Oleynikov, Mr. Faruk Cay, Mr. Raveen Sharma, Dr. Csaba Mihaly, Mr. Eric Babyak, and Dr. Yonghong Zhang, for their help in this project. We thank Ms. Joni Kohr, Dr. Hallgeir Rui, and Mr. John Eberhardt for their help in manuscript preparation and revision. We also thank the anonymous reviewers for their constructive comments that resulted in enhanced quality of this paper. This work was supported by the Clinical Breast Care Project, and the Gynecological Disease Program, with funds from the US Department of Defense through Henry Jackson Foundation for the Advancement of Military Medicine, Rockville, MD. The views expressed in this paper are those of the authors and do not reflect the official policy of the Department of the Army, Department of Defense, or the government of the United States.

References


