Aggregated Biomedical-Information Browser (ABB): A Graphical User Interface for Clinicians and Scientists to Access a Clinical Data Warehouse

Susan Maskery1, Anthony Bekhash1, Leonid Kvecher1, Mick Correll2, Jeffrey A Hooke3, Albert J Kovatich3, Craig D Shriver3, Richard J Murali and Hai Hu*+1

1Windber Research Institute, Windber, PA, USA 2InforSense, Ltd., Cambridge, MA, USA 3Walter Reed National Military Medical Center, Bethesda, MD, USA

Abstract

Clinicians have unique insight into the diseases and medical conditions they treat, and may develop their own hypotheses they wish to explore by examining existing cases in a data warehouse. To facilitate manual data mining by clinicians and scientists, we have developed an interface for our clinical data warehouse, the Aggregated Biomedical-information Browser (ABB), based on OLAP (On-Line Analytical Processing) technology. The ABB enables clinicians, researchers, and other domain experts to quickly and intuitively explore data in our data warehouse, the Data Warehouse for Translational Research (DW4TR), without needing to involve informatics staff for data extraction. The ABB is capable of handling “on the fly” queries of any data element within the DW4TR. This functionality enables researchers to use their domain knowledge to connect disparate data points as one discovery leads to another. Hypotheses generated through manual data mining combined with domain knowledge, can then be tested using more advanced statistical methods. To illustrate this process a manual data mining example comparing breast cancer pathology in African American and Caucasian American women is performed using the ABB. Analysis of several breast cancer pathology markers suggest African American women will have a worse clinical outcome than Caucasian American women, a clinically meaningful outcome well documented in scientific literature. This report demonstrates the simple yet powerful use of the ABB for manual data exploration in the initial hypothesis generation stage.

Keywords: Clinical Data Warehouse; OLAP; Data Mining; Breast Cancer

Abbreviations: ABB: Aggregated Biomedical-information Browser; OLAP: On-Line Analytical Processing; DW4TR: Data Warehouse for Translational Research; MOLAP: Multidimensional On-Line Analytical Processing; CBRC: Clinical Breast Care Project; IRB: Institutional Review Board; EAV: Entity Attribute Value; SIIV: Individual Subject Information Viewer; BMI: Body Mass Index; ER: Estrogen Receptor; PR: Progesterone Receptor

Background

Development of data warehouse technology for clinical data management has been well documented in the literature, and clinical data warehouses harboring data from electronic medical record systems to support clinical and translational research have been reported [1-8]. Some of these data warehouses allow researchers to query aggregated information for patient cohort and biospecimen selection to enable initial experimental design [7,8]. Of them, a system named Informatics for Integrating Biology and the Bedside (i2b2), is probably the best developed and most widely deployed open-source system, with a collection of functional modules (called “cells”) including ones for de-identification and natural language processing; it has an open architecture allowing integration of cells (modules) developed by independent researchers [8-10]. However, there are few reported powerful yet easy-to-use tools, beyond cohort and biospecimen selection, to allow dynamic cross-examination of multi-dimensional clinical data by non-informaticians.

User friendly interfaces that enable non-informaticians to easily query data in these often huge data repositories is a recognized need [11]. Empowered with the real-time “manual data-mining” capability, non-informatician clinicians and researchers can apply their domain knowledge at the time the results are queried and reported to allow dynamic and non-interrupted “drilling down” of the research question for efficient hypothesis generation. To fill this gap we developed the Aggregated Biomedical information Browser (ABB), an On-Line Analytical Processing (OLAP) based interface to our Data Warehouse for Translational Research (DW4TR), to enable easy access and simple analysis of stored clinical data [12,13]. The ABB is designed to query and display clinical data to enable manual data mining by researchers and clinicians. Toward this end, an ABB system capable of on-the-fly data retrieval of sparse heterogeneous clinical data was needed. The multiple commercial OLAP systems on the market are not designed for this type of data retrieval. Commercial OLAP systems, designed to quickly return query results on dense homogeneous data, are often MOLAP (Multidimensional OLAP) based. These systems require user queries that are somewhat predictable so that decisions regarding which data to aggregate for analysis and summarization can be made in advance, in the form of a data cube or data mart. However, clinical research queries are unpredictable; constantly evolving as knowledge accumulates, resulting in unanticipated queries which necessitate rebuilding the MOLAP data cube – a computationally intense process.

*Corresponding author: Hai Hu, Windber Research Institute, 620 7th Street, Windber, PA 15963, USA, Tel: 814-361-6903; Fax: 814-467-6334; E-mail: h.hu@wrwindber.org

Received November 19, 2013; Accepted December 30, 2013; Published January 06, 2014


Copyright: © 2014 Maskery S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Traditional commercial MOLAP technology was inadequate for such needs. The ABB was developed to fill this gap between the commercial OLAP platforms available and the clinical data centric retrieval needs at our institute [13].

In this paper we demonstrate the functionality of the ABB and illustrate how the ABB can be used to iteratively and manually mine data from the DW4TR. Tumor pathology differences between African American and Caucasian American women motivate an initial manual data mining exercise using the ABB. Relevant tumor pathology data is queried, simple off-line statistical calculations are performed and statistically significant differences in tumor pathology and patient characteristics are presented and interpreted. In short, we demonstrate how a non-informatician can use the ABB to successfully perform initial hypothesis exploration using clinical data collected in the DW4TR.

Methods

Study population

This study used data gathered as part of the Clinical Breast Care Project (CBCP). As of June 2012 there were 5664 participants enrolled in the CBCP, an Institutional Review Board (IRB) approved multi-institute study. The majority of the CBCP population is drawn from an adult military beneficiary population seen at the Comprehensive Breast Center at Walter Reed National Military Medical Center upon referral by a primary care doctor. Examples of conditions resulting in a referral to this clinic include: an abnormal lump in the breast, an abnormal radiological finding, at high risk for developing breast cancer, or other breast related conditions.

Upon recruitment, each enrolled and consented subject is administered a comprehensive life history questionnaire, referred to as the Core Questionnaire (427 data elements). This questionnaire covers: demographics, health history, and lifestyle practices. For those CBCP participants who require a biopsy, a Pathology Checklist (372 data elements) is completed recording the true/false occurrence of 131 breast and lymphoid conditions in addition to other standard pathology tests done on invasive breast cancer samples. Biospecimens, including blood products and tissue (when available), are also collected for genomic and proteomic studies. These questionnaires are completed by specially trained nurses (Clinical Core Questionnaire) or pathologists (Pathology Checklist).

All de-identified clinicopathologic data are stored in the CBCP data warehouse. Rigorous quality control and quality assurance measures, such as review of questionnaires for obvious mistakes, double data entry, a QAMetrics computational program to identify data consistency errors, and a quality assurance issue tracking system [14], are in place. Data collected from all participants are stored in the DW4TR and have been used for this analysis.

Data warehouse

The DW4TR is composed of a data tier, a middle tier, and an application tier [14]. In the data tier, raw clinical data are extracted from the data tracking system using a modified Entity – Attribute – Value (EAV) data model, staged, transformed, and finally loaded into the physical data tables; this whole process followed a standard practice referred to as Extract, Transform, and Load using analytical workflows developed based on the IDBS Inference platform that serves as the analytical backend of the DW4TR [15]. In the middle tier, a hierarchical patient-centric clinical data model describes a patient using 5 major modules: Medical History, Physical Examination, Diagnostic, Therapeutics, and Surveys/Consents (Figure 1). Each of these modules is composed of sub-modules and attributes. A sub-module may be further composed of its own sub-modules and attributes. An in-house developed data model defines the hierarchical relationships for all sub-modules and attributes. The middle tier also contains a specimen-centric, modularly-structured molecular data model [13] that is integrated with the patient-centric data model described above.

Application tier

The application tier is composed of two major interfaces, the OLAP-based ABB, and an Individual Subject Information Viewer (ISIV). The ISIV, not shown in the article, is for detailed analysis of individual subjects, including, the temporal relationships of data elements. The ABB allows the user to query the DW4TR by dynamically creating a two dimensional view (rows and columns). The rows are hierarchical categorical data fields of interest. The columns are either numerical or categorical data elements. Columns can be flat or hierarchical. The user can explore data in multiple dimensions by hierarchically expanding data elements in rows or columns [13].

Composing a data view in the ABB

The ABB displays a two-dimensional interactive graphical data view of variables of interest in the DW4TR. Figure 2 shows the ABB interface when CBCP participants were stratified by invasive breast cancer and ethnicity. Additional columns are added to compare local percentages and average age between ethnicities and cancer/non-cancer cases.

---

Figure 1: Hierarchical Structure of the Patient-Centric Clinical Data Model. The patient is composed of 5 major modules including Medical History, Medical History is composed of 8 submodules including Personal Information. Personal Information is composed of 9 attributes or submodules, including Ethnicity.
Custom Binning in the ABB

Multiple binning strategies can be either automatically or manually created to explore data elements that contain numerical or categorical values. For example, customized binning was used to define a variable into four ethnicity categories from the original 11 possible ethnicity categories in the DW4TR. The ABB custom binning process is shown in Figure 3.

Hierarchical Data Structures of the ABB (Drilling Down)

The ABB enables the user to drill down through the population, expanding a two dimensional view into multiple dimensions. In Figure 2, two hierarchical levels, Invasive Breast Cancer (INV Yes/no) and Ethnicity, are expanded. Expanded rows can be contracted to hide less important variables for a cleaner display.

Analysis capability of the ABB

Simple analyses can be directly performed using the ABB by selecting a function when defining the columns of the data view. Functions for patient counts, global percentages relative to the whole data view, local percentages relative to the subset, etc., are available. Similar analyses can be done for biospecimens. For non-categorical data (e.g. age, tumor size, number of live births, etc.) functions for calculating mean, standard deviation, median, sum, and minimal or maximal values, etc. are available.

View saving, printing, and exporting of the ABB

The data view can be printed as shown or saved for future use. It can also be exported to a flat file in comma-separated values. The flat file can be read into an Excel spreadsheet or imported into a statistical package for further analysis.

Cohort saving in the ABB

Subject groups of interest can be saved as a cohort of subjects or corresponding biospecimens for future analysis using the Cohort View feature. Cohort View is an application for analyzing a pre-saved cohort, with members of the cohort shown as rows. The user can introduce variables of interest in columns to further explore the properties of the cohort. If a user is applying the same analysis to cohorts of choice, the analysis can be developed into an application using the data warehouse platform and launched from the interface of Cohort View.

Uni-Variate Statistical Analysis

Variables selected for further analysis are exported from the ABB.

---

Figure 2: Screenshot of the ABB interface for Analyzing the Variables "INV: Yes/no" and "Ethnicity_Binned". These two variables are shown in rows. "INV: Yes/no" is a patient’s invasive breast cancer status and Ethnicity_Binned is a patient’s self reported ethnicity. The columns are patient count, local percentage and average age. The hierarchical view shows the drill down option: from "INV: Yes/no" to "Ethnicity_Binned".
Figure 3: Screenshots of the ABB Custom Binning Feature.

A) Initial view of binning window, all Ethnicity values are in the “Value not assigned” box to the right.

B) New bins (Caucasian American, African American, and other) have been created.

As part of the custom binning feature, the user can view a pie or bar chart of the counts for each variable in their new binning. Shown is a pie chart of the newly binned Ethnicity variable.
A 2x2 contingency table is created for the frequency of each exported variable. Significance is assessed using the Pearson Chi-Squared ($\chi^2$) statistic calculated in Intercooled Stata 10.0 (College Station, TX). Significance is recorded when the $\chi^2$ probability p value falls below 0.05.

**Results**

**Population Summary**

The study population is the June 2012 CBCP population of 5664 participants. Table 1 is a summary of this population’s characteristics including age, ethnicity, education, body mass index (BMI), and biopsy results (if any). This population is ethnically and socio-economically mixed.

**ABB Utilization – Iterative Manual Data Mining**

**Step 1 - Initial Hypothesis:** The scenario motivating this manual data mining exercise was the following: a physician noticed their African American breast cancer patients were more likely to present with high grade breast cancer compared to their Caucasian American patients. In tumor pathology, tumor grade reports how similar tumor cells are to normal cells in the body. Low grade tumor cells appear similar to normal cells and do not rapidly divide. High grade tumor cells appear very different than normal cells and tend toward rapid growth. In general, the higher the tumor grade, the worse the prognosis for the patient [16]. The initial data mining iteration retrieved breast cancer grade data for African American and Caucasian American women to confirm that African American women did present with a higher grade breast cancer than Caucasian American women.

**Step 2 - Manual Data Mining Iteration 1:** The first step was to create a data view in the ABB. First, the variable, “INV: Yes/no” was selected. Next, the variable, “Ethnicity_Bin1” was selected. Last, the variable, “INV: Grade” was selected into the hierarchical row structure. “Patient Count” and the local percentage were selected as columns to complete the view (Figure 4). Drilling down through the hierarchical structure of the view it quickly became apparent by inspection that the percentage of high grade tumors in African Americans (46%) was greater than the percentage of high grade tumors in Caucasian Americans (28%).

**Step 3 - Statistical Analysis Iteration 1:** The resulting view was then exported to a flat file for a cross tabulation off-line statistical analysis (Table 2). The analysis generated $p$ ($\chi^2$)<0.001, a highly significant result. The clinician’s initial hypothesis was correct – in this population, African American women were significantly more likely to present with a higher grade breast cancer when compared to Caucasian American women.

**Step 4 – Hypothesis Refinement Iteration 2:** In addition to high grade, several other pathology variables indicate a worse breast cancer prognosis. These indicators are: tumor size >2 cm, negative estrogen receptor (ER) status, negative progesterone receptor (PR) status, positive her2/neu receptor status, and high cell proliferation (Ki67 presence). Tumors that are both ER and PR positive respond well to adjuvant therapies, thus increasing long term patient survival rate [17,18]. Her2/neu positive breast cancers are typically very aggressive [19,20]. However, these cancers do have a targeted treatment that increases long term disease free survival [21]. Tumor size is an independent marker for prognosis, and is used in staging breast cancer for treatment. Breast tumors 2 cm and under are in the smallest size class (T1) and generally have the highest (88%) five year survival rate.

**Table 1: Study Population Characteristics.**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Participant Count (N=5664)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Under 35</td>
<td>560</td>
<td>10%</td>
</tr>
<tr>
<td>35-45</td>
<td>1240</td>
<td>22%</td>
</tr>
<tr>
<td>45-55</td>
<td>1409</td>
<td>25%</td>
</tr>
<tr>
<td>55-65</td>
<td>1109</td>
<td>20%</td>
</tr>
<tr>
<td>Over 65</td>
<td>1048</td>
<td>19%</td>
</tr>
<tr>
<td>Unknown</td>
<td>298</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up to and including high school diploma</td>
<td>1311</td>
<td>23%</td>
</tr>
<tr>
<td>Some college or Associate Degree</td>
<td>1615</td>
<td>29%</td>
</tr>
<tr>
<td>Bachelor Degree</td>
<td>1200</td>
<td>21%</td>
</tr>
<tr>
<td>Graduate Degree</td>
<td>888</td>
<td>16%</td>
</tr>
<tr>
<td>Unknown</td>
<td>709</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caucasian American</td>
<td>3799</td>
<td>67%</td>
</tr>
<tr>
<td>African American</td>
<td>1082</td>
<td>19%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>191</td>
<td>3%</td>
</tr>
<tr>
<td>Asian</td>
<td>156</td>
<td>3%</td>
</tr>
<tr>
<td>Other</td>
<td>99</td>
<td>2%</td>
</tr>
<tr>
<td>Unknown</td>
<td>337</td>
<td>6%</td>
</tr>
<tr>
<td><strong>Body Mass Index (BMI)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underweight - BMI &lt;20 kg/m²</td>
<td>265</td>
<td>5%</td>
</tr>
<tr>
<td>Normal - 20kg/m²&lt;BMI&lt;25kg/m²</td>
<td>1831</td>
<td>32%</td>
</tr>
<tr>
<td>Overweight - 25kg/m²&lt;BMI&lt;30kg/m²</td>
<td>1649</td>
<td>29%</td>
</tr>
<tr>
<td>Obese - BMI&gt;30 kg/m²</td>
<td>1463</td>
<td>26%</td>
</tr>
<tr>
<td>Unknown</td>
<td>456</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Pathology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benign</td>
<td>1734</td>
<td>31%</td>
</tr>
<tr>
<td>Atypical</td>
<td>138</td>
<td>2%</td>
</tr>
<tr>
<td>In Situ</td>
<td>368</td>
<td>6%</td>
</tr>
<tr>
<td>Invasive</td>
<td>1651</td>
<td>29%</td>
</tr>
<tr>
<td>Malignant/Malignant NOS</td>
<td>12</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>No Pathology Report</td>
<td>1816</td>
<td>32%</td>
</tr>
</tbody>
</table>

Table 2: Invasive Grade by Ethnicity.

<table>
<thead>
<tr>
<th>Category</th>
<th>Caucasian American (N=1207)</th>
<th>African American (N=235)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Well Differentiated</td>
<td>383 (32%)</td>
<td>46 (20%)</td>
</tr>
<tr>
<td>Moderately Differentiated</td>
<td>478 (40%)</td>
<td>79 (34%)</td>
</tr>
<tr>
<td>Poorly Differentiated</td>
<td>334 (28%)</td>
<td>109 (46%)</td>
</tr>
</tbody>
</table>

\( p(\chi^2) < 0.001 \)

Figure 5: Ethnicity Stratification of Pathology Characteristics. The result of the ethnicity binning in Figure 3 and drill down capability in Figure 4 is shown. The data exported from this data view is used to generate Table 3.

Ki67 is a marker for high rates of cell division. Tumors with a high rate of cell division are typically more aggressive, and consequently have a worse prognosis [23]. The next data mining iteration retrieved these breast cancer prognostic factors stratified by ethnicity.

Step 5 – Manual Data Mining Iteration 2: Additional fields were added into columns of the view, tumor size and the biomarker assay results of ER, PR, HER2/neu, and Ki67. These results are shown in Figure 5. By inspection it was apparent that the distribution of ER positive, PR positive and Ki67 present cases in Caucasian American women was different from those in African American women. Further analysis was needed to see if these differences were statistically significant.

Step 6 - Statistical Analysis Iteration 2: Using the exported flat file, a new cross-tabulation analysis was performed as illustrated in Table 3. From these analyses, it was concluded that in this study population, African American women present with tumors that were significantly more likely to be ER negative, PR negative and Ki67 positive when compared to the tumors of Caucasian American women. Her2/neu receptor presence and tumor size were not significantly different between the two ethnicity groups.

Step 7 - Cohort Selection for Hypothesis Testing: Breast tumor pathology characteristics found significantly more often in African American women - ER negative, PR negative, Ki67 positive, and high grade - suggest a worse prognosis. A survival analysis is needed to verify, in this study population, that African American breast cancer patients do have a worse outcome compared Caucasian American breast cancer. Unfortunately, outcome information is currently incomplete in the CBCP. However, as shown in Figure 6, once outcome data is available the ABB is capable of exporting a cohort of African American and Caucasian American breast cancer patients for further analysis.

Discussion

With a hypothetical example case, we demonstrated how a physician, who is not an informatics specialist, can use the ABB interface of the DW4TR to perform manual data mining based on their own and their colleagues’ clinical observations. The example case started with comparing the pathology characteristics of tumors in African American women to tumors in Caucasian American women. Similar to results in literature, African American women in the CBCP present with breast cancer that is of higher grade and ER and PR negative [24]. Although not testing Ki67 specifically, other groups have seen greater cell proliferation in African American breast cancer tumor samples when compared to Caucasian American samples [24]. All four of these tumor pathology characteristics are indicative of a more aggressive cancer and worse outcome compared to low grade, ER/PR positive, Ki67 negative tumors [16,23].

The ABB enables manual data mining by non-informatician users. The display and data query options are specifically designed for clinicians and scientists working in clinical research. Features such as population drill down, hierarchical data display, and cohort selection enable a quick perusal and export of data in the data warehouse. The data exportation functionality enables subsequent statistical analysis when interesting observations are made in manual data mining. Unlike off-the-shelf commercial OLAP applications, the ABB is designed to retrieve sparse heterogeneous clinical data and dynamically and quickly handle unanticipated queries. Previous work in the development of user friendly tools to access clinical data warehouse data primarily focused on extracting cohorts (e.g. list of subjects and their attributes) [11,25-27]. The ABB has cohort retrieval functionality. In addition, the ABB empowers physicians and scientists to directly manually mine the data in a clinical data warehouse, thus offering a highly desired service to this group of users.

While the DW4TR was initially designed for a breast cancer study, it has been expanded to a gynecological cancer research program, the Gynecologic Cancer Center of Excellence (GYN-COE) [13,14]. This study currently involves over 1,000 data elements including demographic, surgicopathologic, biospecimen, and patient outcome data. The data were collected using both data forms and electronic data capturing systems from five clinical sites. One of the main additional tasks we faced was to standardize the different formats of the legacy data collected from different clinical sites. New data elements specific to
gynecological cancer were also managed, as well as the clinical outcome data that were not available in the CBCP when this study was started. In addition to breast and gynecological cancers, the principle presented here for the use of the ABB interface also applies to other disease studies. We expect that as more cases and follow-up information are added to the CBCP, and as the DW4TR continues to expand to cover other disease types, the system we present here will find wider use by clinicians and scientists studying breast cancer and other human diseases.

Acknowledgments

We thank the patients whom participated in this study. This work was supported by the Clinical Breast Care Project with funds from the US Department of Defense through the Henry Jackson Foundation for the Advancement of Military Medicine, Rockville, MD. The views expressed in this article are those of the author and do not reflect the official policy of the Department of Defense, or U.S. Government.

References


