Journal of
Computer Science & Systems Biology

Maskery et al., J Comput Sci Syst Biol 2014, 7:1
DOI: 10.4172/jcsb.1000134

Research Article Open Access

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

Susan Maskery[1], Anthony Bekhash[1], Leonid Kvecher[1], Mick Correll[2], Jeffrey A Hooke[3], Albert J Kovatich[3], Craig D Shriver[3], Richard J Mural[1] and Hai Hu[1]*

[1]Windber Research Institute, Windber, PA, USA
[2]InforSense, Ltd., Cambridge, MA, USA
[3]Walter 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; DW4TRL: Data Warehouse for Translational Research; MOLAP: Multidimensional On-Line Analytical Processing; CBCP: Clinical Breast Care Project; IRB: Institutional Review Board; EAV: Entity Attribute Value; ISIV: 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 OLAP 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@wriwindber.org

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

Citation: Maskery S, Bekhash A, Kvecher L, Correll M, Hooke JA, et al. (2014) Aggregated Biomedical-Information Browser (ABB): A Graphical User Interface for Clinicians and Scientists to Access a Clinical Data Warehouse. J Comput Sci Syst Biol 7: 020-027. doi:10.4172/jcsb.1000134

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 incidence of the DW4TR. 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.

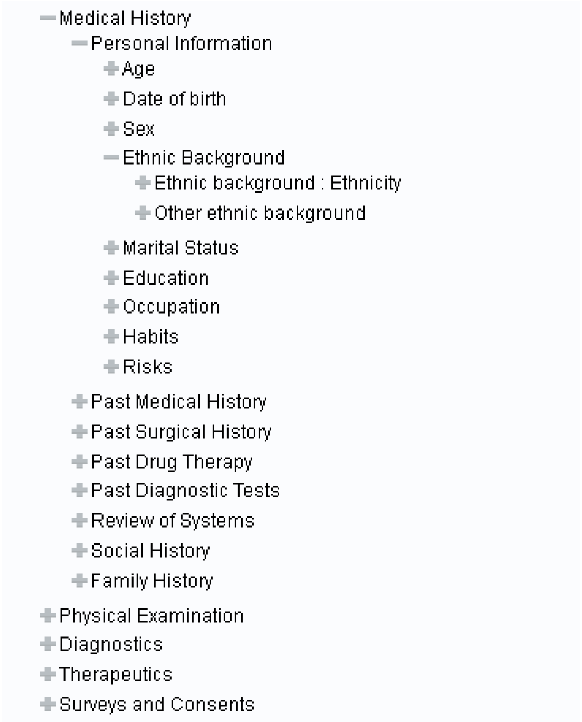


Figure 1: Hierarchical Structure of the Patient-Centric Clinical Data Model. The patient is composed of 5 major modules including Medical History. Medical History is composed of 8 submodules including Personal Information. Personal Information is composed of 9 attributes or submodules, including Ethnicity.

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

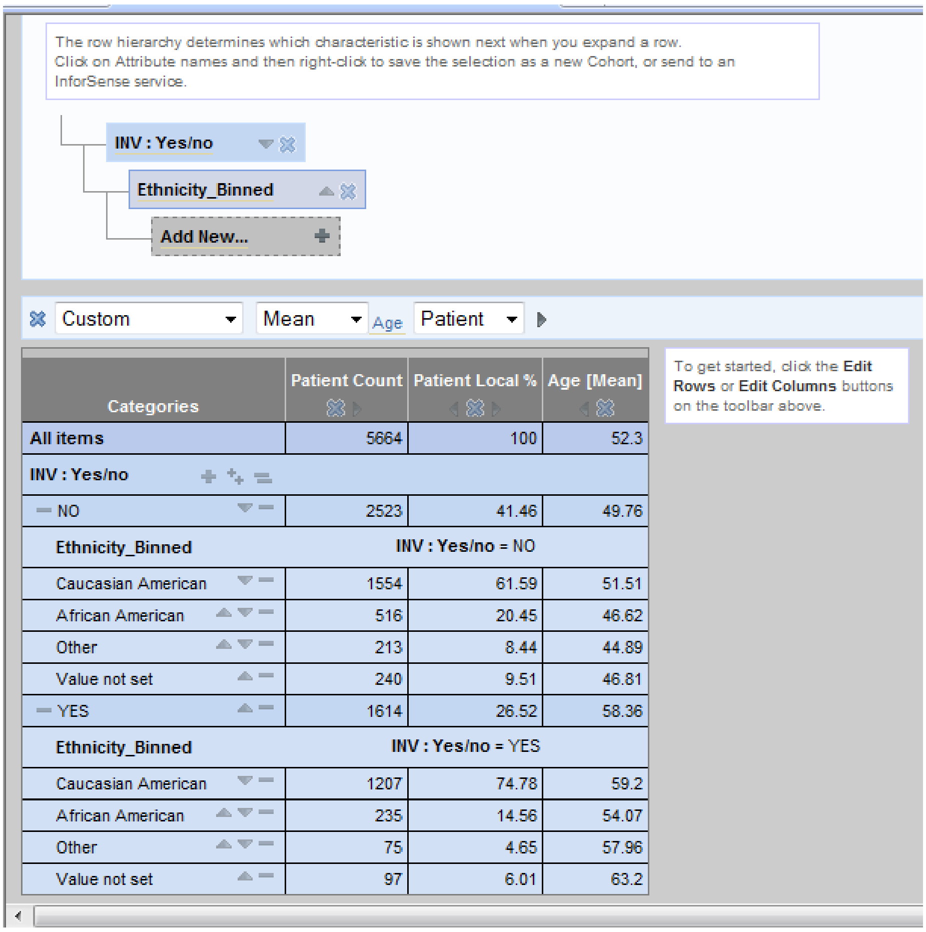
The data view can be printed as shown or saved for future use. It can also be exported to a flat file in comma-separated values. The flat file can be read into an Excel spreadsheet or imported into a statistical package for further analysis.

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.

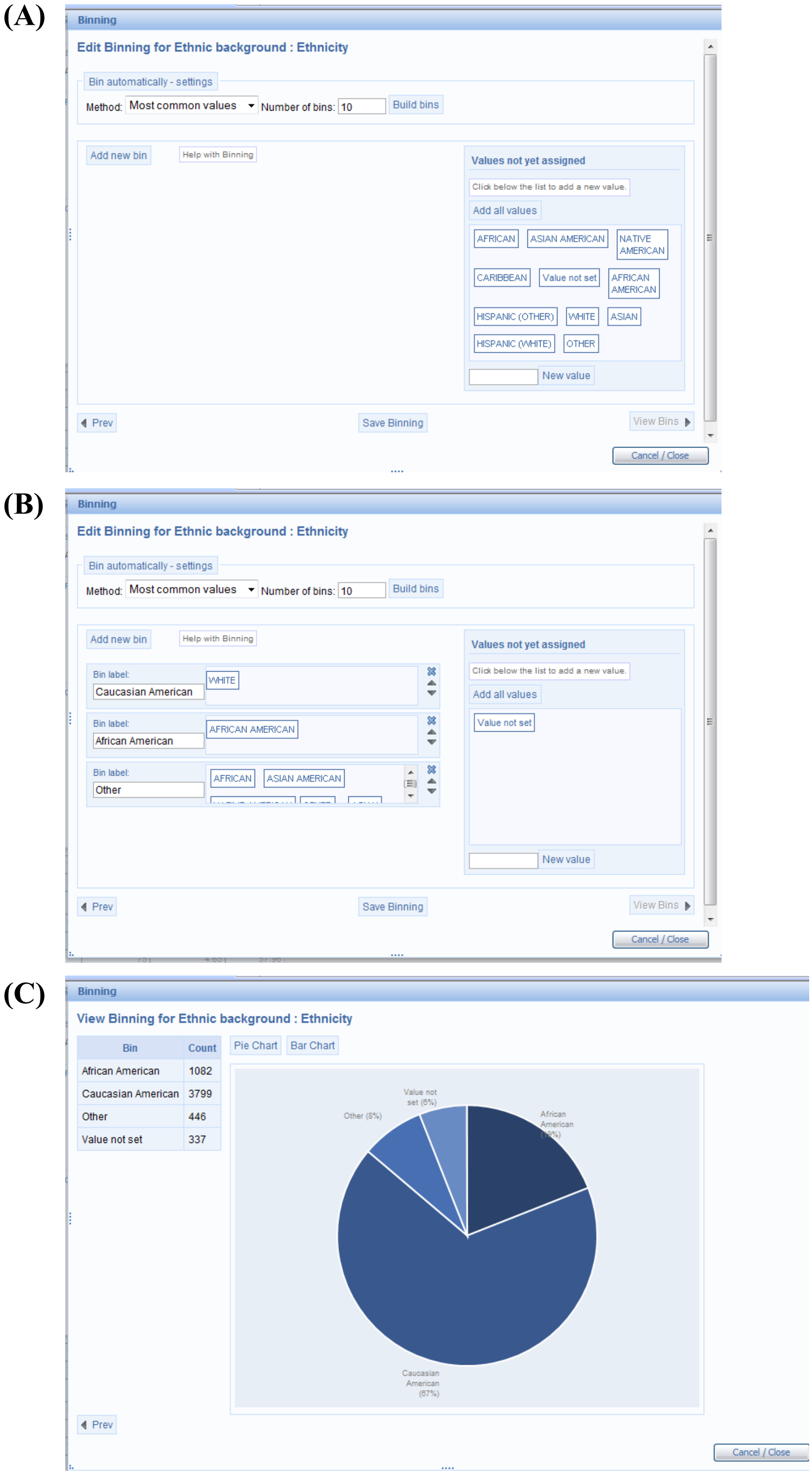


The row hierarchy determines which characteristic is shown next when you expand a row.
Click on Attribute names and then right-click to save the selection as a new Cohort, or send to an InforSense service.

To get started, click the **Edit Rows** or **Edit Columns** buttons on the toolbar above.

Categories	Patient Count	Patient Local %	Age [Mean]
All items	5664	100	52.3
INV : Yes/no			
NO	2523	41.46	49.76
Ethnicity_Binned	INV : Yes/no = NO		
Caucasian American	1554	61.59	51.51
African American	516	20.45	46.62
Other	213	8.44	44.89
Value not set	240	9.51	46.81
YES	1614	26.52	58.36
Ethnicity_Binned	INV : Yes/no = YES		
Caucasian American	1207	74.78	59.2
African American	235	14.56	54.07
Other	75	4.65	57.96
Value not set	97	6.01	63.2

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".



(A)

Edit Binning for Ethnic background : Ethnicity

Bin automatically - settings

Method: Most common values Number of bins: 10 Build bins

Values not yet assigned

AFRICAN, ASIAN AMERICAN, NATIVE AMERICAN, CARIBBEAN, Value not set, AFRICAN AMERICAN, HISPANIC (OTHER), WHITE, ASIAN, HISPANIC (WHITE), OTHER

(B)

Edit Binning for Ethnic background : Ethnicity

Bin automatically - settings

Method: Most common values Number of bins: 10 Build bins

Bin label: Caucasian American — WHITE

Bin label: African American — AFRICAN AMERICAN

Bin label: Other — AFRICAN, ASIAN AMERICAN

Values not yet assigned

Value not set

(C)

View Binning for Ethnic background : Ethnicity

Bin	Count
African American	1082
Caucasian American	3799
Other	446
Value not set	337

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 (χ^2) statistic calculated in Intercooled Stata 10.0 (College Station, TX). Significance is recorded when the χ^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.

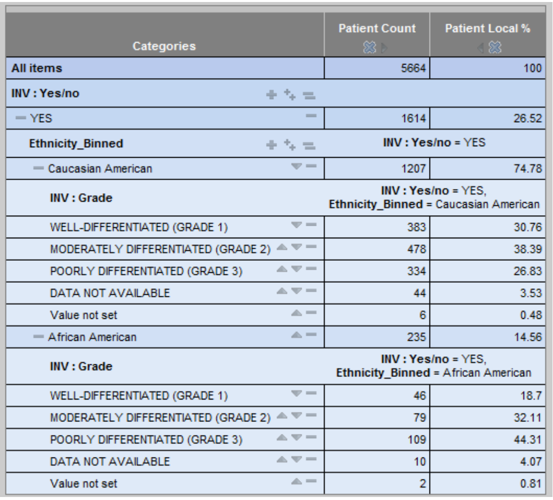
Characteristic	Participant Count (N=5664)	Percentage
Age		
Under 35	560	10%
35-45	1240	22%
45-55	1409	25%
55-65	1109	20%
Over 65	1048	19%
Unknown	298	5%
Education		
Up to and including high school diploma	1311	23%
Some college or Associate Degree	1615	29%
Bachelor Degree	1200	21%
Graduate Degree	888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

Table 1: Study Population Characteristics.

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%).



Categories	Patient Count	Patient Local %
All items	5664	100
INV : Yes/no		
YES	1614	26.52
Ethnicity_Binned	INV : Yes/no = YES	
Caucasian American	1207	74.78
INV : Grade	INV : Yes/no = YES, Ethnicity_Binned = Caucasian American	
WELL-DIFFERENTIATED (GRADE 1)	383	30.76
MODERATELY DIFFERENTIATED (GRADE 2)	478	38.39
POORLY DIFFERENTIATED (GRADE 3)	334	26.83
DATA NOT AVAILABLE	44	3.53
Value not set	6	0.48
African American	235	14.56
INV : Grade	INV : Yes/no = YES, Ethnicity_Binned = African American	
WELL-DIFFERENTIATED (GRADE 1)	46	18.7
MODERATELY DIFFERENTIATED (GRADE 2)	79	32.11
POORLY DIFFERENTIATED (GRADE 3)	109	44.31
DATA NOT AVAILABLE	10	4.07
Value not set	2	0.81

Figure 4: Screenshot of ABB Interface Drill Down Data View of Invasive Breast Cancer Cases. Cases are stratified by ethnicity and grade. To drill down through the view, first, invasive cancer cases are selected. The population is further divided into ethnicity and grade.

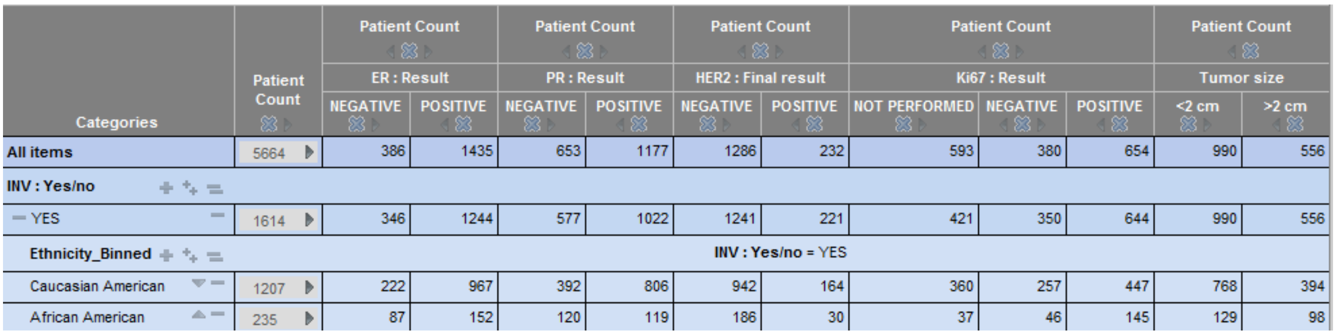
Step 3 - Statistical Analysis Iteration 1: The resulting view was then exported to a flat file for a cross tabulation off-line statistical analysis (Table 2). The analysis generated p (χ^2)<0.001, a highly significant result. The clinician's initial hypothesis was correct – in this population, African 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

	Caucasian American (N=1207)	African American (N=235)
Well Differentiated	383 (32%)	46 (20%)
Moderately Differentiated	478 (40%)	79 (34%)
Poorly Differentiated	334 (28%)	109 (46%)

$p(\chi^2)<0.001$

Table 2: Invasive Grade by Ethnicity.



Categories	Patient Count	Patient Count		Patient Count		Patient Count		Patient Count			Patient Count	
		ER : Result		PR : Result		HER2 : Final result		Ki67 : Result			Tumor size	
		NEGATIVE	POSITIVE	NEGATIVE	POSITIVE	NEGATIVE	POSITIVE	NOT PERFORMED	NEGATIVE	POSITIVE	<2 cm	>2 cm
All items	5664	386	1435	653	1177	1286	232	593	380	654	990	556
INV : Yes/no												
YES	1614	346	1244	577	1022	1241	221	421	350	644	990	556
Ethnicity_Binned	INV : Yes/no = YES											
Caucasian American	1207	222	967	392	806	942	164	360	257	447	768	394
African American	235	87	152	120	119	186	30	37	46	145	129	98

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.

[22]. 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

	ER (-\+)	PR (-\+)	HER2 (-\+)	Ki67* (-\+)	Tumor Size ≤ 2 cm\ >2 cm
Caucasian American	222\967	392\806	942\164	257\447	768\394
African American	87/152	119\120	186\30	46\145	129\98
p(χ^2)	p(χ^2)<0.001	p(χ^2)<0.001	p(χ^2)<0.7	p(χ^2)<0.001	p(χ^2)<0.25

*test for ER, PR, HER2 and Ki67 were not run on all tumour samples

Table 3: Cross-tabulation analysis of tumour pathology and patient characteristics stratified by ethnicity.

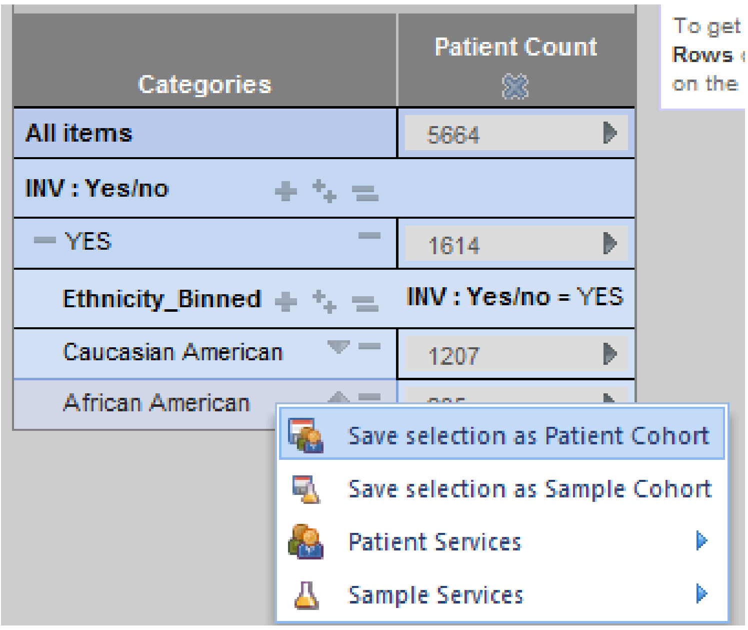


Figure 6: Cohort Export Option Screenshot. The ABB cohort export utility is illustrated. Once again, the ethnicity binning and drill down utilities are used to set up the desired data view. Once the cohort of interest is identified, in this case through manual data mining, it can be saved for further analysis.

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

1. Kerkri E, Quantin C, Yetongnon K, Allaert FA, Dusserre L (1999) Application of the medical data warehousing architecture EPIDWARE to epidemiological follow-up: data extraction and transformation. Stud Health Technol Inform 68: 414-418.
2. Ramick DC (2001) Data warehousing in disease management programs. J Healthc Inf Manag 15: 99-105.
3. Koprowski SP Jr, Barrett JS (2002) Data warehouse implementation with clinical pharmacokinetic/pharmacodynamic data. Int J Clin Pharmacol Ther 40: S14-29.
4. Kamal J, Silvey SA, Buskirk J, Ostrander M, Erdal S, et al. (2008) Innovative applications of an enterprise-wide information warehouse. AMIA Annu Symp Proc 1134.
5. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, et al. (2010) Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 17: 124-130.
6. Prather JC, Lobach DF, Goodwin LK, Hales JW, Hage ML, et al. (1997) Medical data mining: knowledge discovery in a clinical data warehouse. Proc AMIA Annu Fall Symp 101-105.
7. Lowe HJ, Ferris TA, Hernandez PM, Weber SC (2009) STRIDE--An integrated standards-based translational research informatics platform. AMIA Annu Symp Proc 2009: 391-395.
8. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, et al. (2007) Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc 548-552.
9. Mendis M, Phillips LC, Kuttan R, Pan W, Gainer V, et al. (2008) Integrating outside modules into the i2b2 architecture. AMIA Annu Symp Proc 1054.
10. Mendis M, Wattanasin N, Kuttan R, Pan W, Philips L, et al. (2007) Integration of Hive and cell software in the i2b2 architecture. AMIA Annu Symp Proc 1048.
11. Nigrin DJ, Kohane IS (1998) Data mining by clinicians. Proc AMIA Symp 957-961.
12. Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, et al. (2004) Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. Pharmacogenomics 5: 933-941.
13. Hu H, Correll M, Kvecher L, Osmond M, Clark J, et al. (2011) DW4TR: A Data Warehouse for Translational Research. J Biomed Inform 44: 1004-1019.
14. Zhang Y, Sun W, Gutchell EM, Kvecher L, Kohr J, et al. (2013) QAIT: a quality assurance issue tracking tool to facilitate the improvement of clinical data quality. Comput Methods Programs Biomed 109: 86-91.
15. Beaulah SA, Correll MA, Munro RE, Sheldon JG (2008) Addressing informatics

challenges in Translational Research with workflow technology. Drug Discov Today 13: 771-777.

16. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, et al. (2010) Breast cancer prognostic classification in the molecular era: the role of histological grade. Breast Cancer Res 12: 207.

17. [No authors listed] (1998) Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet 351: 1451-1467.

18. Bardou VJ, Arpino G, Elledge RM, Osborne CK, Clark GM (2003) Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. J Clin Oncol 21: 1973-1979.

19. Borg A, Tandon AK, Sigurdsson H, Clark GM, Fernö M, et al. (1990) HER-2/neu amplification predicts poor survival in node-positive breast cancer. Cancer Res 50: 4332-4337.

20. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, et al. (1987) Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 235: 177-182.

21. Viani GA, Afonso SL, Stefano EJ, De Fendi LI, Soares FV (2007) Adjuvant trastuzumab in the treatment of her-2-positive early breast cancer: a meta-analysis of published randomized trials. BMC Cancer 7: 153.

22. (2012) American Cancer Society. Detailed Guide to Breast Cancer. American Cancer Society [updated 2012].

23. Goodson WH 3rd, Moore DH 2nd, Ljung BM, Chew K, Mayall B, et al. (2000) The prognostic value of proliferation indices: a study with in vivo bromodeoxyuridine and Ki-67. Breast Cancer Res Treat 59: 113-123.

24. Newman LA (2005) Breast cancer in African-American women. Oncologist 10: 1-14.

25. Nadkarni PM, Brandt C (1998) Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc 5: 511-527.

26. McDonald CJ, Dexter P, Schadow G, Chueh HC, Abernathy G, et al. (2005) SPIN query tools for de-identified research on a humongous database. AMIA Annu Symp Proc 515-519.

27. Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, et al. (2009) The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc 16: 624-630.