ARTICLE IN PRESS

ELSEVIER

Computers in Biology and Medicine ▮▮▮ (▮▮▮) ▮▮▮–▮▮▮

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Data mining and clinical data repositories: Insights from a 667,000 patient data set

Irene M. Mullins[a], Mir S. Siadaty[a], Jason Lyman[a], Ken Scully[a], Carleton T. Garrett[b], W. Greg Miller[b], Rudy Muller[b], Barry Robson[c], Chid Apte[c], Sholom Weiss[c], Isidore Rigoutsos[c], Daniel Platt[c], Simona Cohen[d], William A. Knaus[a,*,1]

[a]*Department of Public Health Sciences, University of Virginia Health System, Charlottesville, VA, USA*
[b]*Department of Pathology, Virginia Commonwealth University, Richmond, VA, USA*
[c]*IBM T.J. Watson Research Center, IBM Life Sciences, Yorktown Heights, New York, USA*
[d]*IBM Research, Haifa, Israel*

Received 28 January 2005; accepted 22 August 2005

Abstract

Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives. To investigate the potential value of searching these databases for novel insights, we applied a new data mining approach, HealthMiner®, to a large cohort of 667,000 inpatient and outpatient digital records from an academic medical system. HealthMiner® approaches knowledge discovery using three unsupervised methods: CliniMiner®, Predictive Analysis, and Pattern Discovery. The initial results from this study suggest that these approaches have the potential to expand research capabilities through identification of potentially novel clinical disease associations.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Clinical data repository; Complex data sets; Large patient cohort; FANO; HealthMiner®; Search tools

* Corresponding author.
E-mail address: wak4b@virginia.edu (W.A. Knaus).
[1] P.O. Box 800717, University of Virginia, Charlottesville, VA, 22908, USA.

0010-4825/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2005.08.003

1. Introduction

Like many academic health centers, the University of Virginia and its partner Virginia Commonwealth University Health System have established, or are developing, Clinical Data Repositories (CDRs). CDRs are large, usually relational, databases that receive a variety of clinical and administrative data from primary electronic sources. These repositories collect comprehensive data on large patient cohorts, assembled and stored over time, which not only permit these institutions to examine trends in utilization and outcomes, but also to perform sophisticated quality assurance and medical management queries independent from the systems that collect the data (laboratory, management systems, etc.) [1,2]. Despite the breadth of stored information, which increasingly includes long-term outcome and associated biological and genetic data, mining for potentially novel and useful biomedical associations in CDRs is a relatively recent approach [3–6].

The term "data mining" often refers to search tools that originated in statistics, computer science, and other non-biomedical disciplines [7]. Currently, the major use for data mining is to find associations among variables that may be useful in future managerial decision making. For example, data mining approaches have been applied extensively within the commercial and defense sectors where they have reported associations as divergent as consumer marketing preferences [8] and corrosion potential for civilian and military aircraft [9].

The application of non-hypothesis driven data mining approaches to high-dimensional medical information may give rise to several problems. First, as with the data mining method chosen for this project, undirected or unsupervised queries (meaning that no, or few, prior assumptions are made about the variables that will correlate) may result in the creation of a combinatorial explosion. However, because this method assumes no prior knowledge, it therefore has the potential to uncover previously unknown relationships.

In many problems outside of medicine, one can avoid the difficulty of unwieldy numbers of solutions by deduction of correlations from just $N(N-1)/2$ pairwise correlations or distance metrics. Applications of this alternative approach depend on the nature of the system being investigated and its underlying constraints and mechanisms. For example, the fact that A and B, B and C, or A and C are often associated together does not allow one to deduce, on statistical grounds, that A, B, and C are never simultaneously seen together. A degree of non-reducibility may hold for at least some of the 50 genomic and 10 lifestyle and clinical history factors responsible for complex disease states, such as cardiovascular disease. Thus, detection of meaningful biomedical correlations from CDRs will require the development of special techniques and heuristics.

The second difficulty in mining CDR data is also a consequence of high dimensionality. Data for complex relationships are usually sparse because they are thinly spread across many dimensions, and extensive data are required to alleviate this problem. However, until quite recently, robust clinical record data have not been available. Large electronic data repositories were not frequently housed at individual institutions [10], much less across institutions in data-sharing consortiums [11]. It also has not been traditional for biomedical research to be driven by the highly structured analyses that are typically attributed to data mining approaches. There is, however, beginning support for the use of larger clinical data resources and, more recently, non-hypothesis-driven research in the biomedical information sciences [12]. This interest is generated both by the increasing availability of large clinical and integrated databases created by the collection of data from routine patient encounters.

Previous analyses using large clinical data sets have typically focused on specific treatment or disease entities. Most have examined targeted treatment procedures: cesarean delivery rate (270,774 women) [13], coronary artery bypass graft (CABG) surgery volume (267,089 procedures) [14], routine chemistry panel testing (438,180 people) [15], and patient care: cancer risk for non-aspirin NSAID users (172,057 individuals) [16], preoperative beta-blocker use and mortality and morbidity following CABG surgery (629,877 patients) [17], and incidence and mortality rate of acute (adult) respiratory distress syndrome (ARDS) (2,501,147 screened discharges) [18], to name a few. These studies have several factors in common: large sample size, clinical information source, and they support or build upon pre-established hypotheses or defined research paradigms that use specific procedure or disease data.

Clinical outcomes algorithms have also been applied to harness large health information databases in order to generate models directly applicable to clinical treatment. These models have been used successfully to create mortality risk assessments for adult [19–21] and pediatric [22] intensive care units. Recently, however, knowledge discovery algorithms have been utilized [4,23,24] in an effort to limit the inherent bias in a priori hypothesis assumptions that can be found in traditional clinical data analysis. In addition, Bayesian networks, which use a graphical diagram to represent probabilistic knowledge [25], have been used in healthcare as a method for pattern recognition and classification for disease management [26–28]. Emerging from Bayesian integration, Robson recently formulated a more generalized theory of expected information (or "Zeta Theory") and application to the development of tools for the analysis and mining of large clinical data sets [29,30].

The University of Virginia, Virginia Commonwealth University, and IBM Life Sciences formed a collaboration designed to test and evaluate data mining approaches in large repositories of clinical, and eventually integrated, biomedical data. As a first step, a 667,000 de-identified patient data set was mined using unsupervised techniques from IBM's HealthMiner® suite, which comprises (i) Association Analysis using FANO (now typically known as CliniMiner®), (ii) Predictive Analysis (PA) using decision rule induction methods [31], and (iii) Pattern Discovery (PD) using THOTH. All three approaches can be considered as distinct types of data mining based on separate data mining philosophies.

FANO/CliniMiner® has been extensively revised for clinical applications, though general in approach, and has "plug-in" components that address specific subject domains previously developed for the clinical and biomedical domains. For example, CliniMiner® contains security features to maintain patient privacy. Also, laboratory data values can be automatically converted to low, normal, and high ranges, while times and dates are converted to universal decimal year time (e.g. 2003.4752827) to facilitate time-stamping of clinical events and time series analysis. Because techniques (ii) and (iii) had not yet been fully completed at the time of this study, the initial cleansing and preparation were performed with CliniMiner® and the results for PA and PD are preliminary.

Our initial and limited goal was to test whether or not it is possible to search a large database of electronic patient records and find novel correlations. This was done without prior selection or bias toward the inclusion or exclusion of particular patient records so as to maximize the potential to lead to novel and useful research hypotheses. In order to accomplish this, we also created an infrastructure that complies with all Health Insurance, Portability, and Accountability (HIPAA) regulations, which were designed to protect the privacy of personal health information [32].

2. Materials and methods

2.1. Theoretical basis of data mining techniques

We have brought, for the first time, three related, but distinct, knowledge discovery tools from the HealthMiner® suite to bear on a remarkably large data set of patient records. HealthMiner® is comprised of three knowledge discovery tools designed to analyze a large dataset of patient records. The methods used by each tool are *related* in that they are all unsupervised "Rule Discovery" techniques. Namely, interesting relationships are sought and discovered without prior knowledge of what those relationships might be, as opposed to directed queries or classical statistical tests of hypotheses.

The methods used in this analysis *differ* in that they pursue different goals in the construction and treatment of the rules they discover. They may reasonably be described as representing three major types of approaches used in the knowledge discovery field, excluding specialist areas, such as time series analysis and cannot be further integrated at this time. None of these three should be considered as more correct than the others.

2.1.1. Pattern Discovery/ THOTH

In the first step, THOTH (named after an Egyptian god who was credited with inventing writing, record keeping, and medicine) begins with PD. Pattern Discovery seeks to enumerate all of the associations that occur at least k times in the data. In the second step, the patterns are clustered based on distances computed from the comparison of the lists of the individual patient records that match the patterns. These clusters find patterns that identify the same lists of patients, and reflect underlying relationships between the parameters shared by all of the patients marked by the patterns in each of the clusters. From the patterns in each of the clusters, the third step constructs all of the possible *enthymemes* (if-then statements) consistent with valid pattern pairs. These take a form such as IF A & B & C . . . & Y THEN Z, and are scored according to the conditional probabilities $P(\mathrm{Z}|\mathrm{A, B, C, \ldots Y, Z}) = P(\mathrm{A, B, C, \ldots Y, Z})/P(\mathrm{A, B, C, \ldots Y})$, which are estimated on a test or trial set for rules that were generated on a training set. Here, as in all three methods, an event such as (A, B, C, . . ., Y, Z) is sometimes called a "complex," "compound" or "conjoint" event and is made up of (e.g. is a simultaneous occurrence in a record of) simple events (items, entries, observations); events such as (A, B, C, . . ., Z) constitute the individual patterns from which enthymemes are constructed. Since each cluster may have associated with it a number of enthymemes or rules, all of the rules are related to each other in that they apply to the same patients and are common to the pathologies the patients share.

2.1.2. Predictive analysis

Predictive analysis learns or generates decision rules from medical data using logical operations (in disjunctive normal form) such as "Diastolic Blood > 100 AND Overweight IMPLIES High Risk of Heart Attack". When applied to a patient record, the terms of the rules are evaluated as true or false, using the operators AND, OR, greater-than, and less-than. As a part of generating the rules, PA searches the full universe of thresholds for numerical variables. Predictive Analysis designates each one of the variables in the patient record as a goal for prediction. Using the remaining variables, it learns rules for each of the goal variables from the sample training data, separating those patients who have the label from those who do not (for example, cancer patients versus normals). The procedure for learning the decision rules is "lightweight rule induction" [31]. Predictive Analysis evaluates, or scores, its decisions by testing on

a completely independent set of patient records. For this analysis, 100,000 patient records were used solely for evaluation. Predictive analysis solves a prediction problem (its rules must predict an outcome on new data with a likelihood significantly greater than chance). It discriminates between the positive and negative outcomes by rules that minimize false positive and false negative errors. Only rules that can potentially predict the outcome are included in the search space. The method searches through many possibilities, attempting to find the best ones in terms of predictive value, sensitivity, and specificity [33]. In this study, 112 variables existed and 112 problems were solved. When solvable, each solution resulted in a small set of predictive rules for each outcome.

2.1.3. FANO/CliniMiner®

Association mining is concerned with whether the conjoint event (A, B, C,. . .) occurs more, or less, than would be expected on a chance basis. If it occurs as much (within a pre-specified margin), then it is not considered an interesting rule. The particular "Zeta Theory" approach used in CliniMiner® is both recent and novel; Zeta Theory seeks to be a self-consistent theory of observations and data which has deep roots in information theory, quantum mechanics, thermodynamics and, most importantly, number theory. It focuses on expectations of (Fano mutual) information measures, these measures being related to the natural log of the probability ratio $P(\mathrm{A, B, C}, \ldots)/[P(\mathrm{A})P(\mathrm{B})P(\mathrm{C})\ldots]$ (and hence measured in natural units or "nats"). More precisely, the "estimate" used is $\zeta\,(\mathrm{s}, o[\mathrm{A, B, C}, \ldots]) - \zeta(s, e[\mathrm{A, B, C}, \ldots])$, where ζ is the Incomplete Riemann Zeta function summed up to the value of the second (o or e) parameter, and o and e are the observed and expected number of observations about conjoint event (A, B, C,. . .). For increasing amounts of data, and $s = 1$, it converges to the log probability ratio; "estimate" is placed in quotes not to indicate any poorness in estimation of this convergence, rather that this expression, in terms of Zeta Functions, is more fundamental than the log probability ratio form. Importantly, at the other extreme, information values for extreme zero occurrence cases of $o = 0$ and/or $e = 0$ are also calculable and meaningful, so that a conjoint event which is not observed, but which statistically should have been, is reported. The parameter s has considerable importance in the theory and method. Varying s values provides both the ability to pre-estimate the chances of a hit while searching a database, and the ability to detect and isolate the influence of errors, noise, approximations, and any probabilistic sampling component. The above applies to qualitative data, but by taking a fuzzy set approach, multivariances between quantitative data can also be processed and expressed as analogous rules by FANO/CliniMiner®.

2.1.4. Comparisons between the HealthMiner® methods

A comparison of the HealthMiner® methods highlights the differences in the types of questions addressed, and their relative strengths and weaknesses. One might argue that in some ways the CliniMiner® mutual association measures, as used here, are more "atomic" in that, given the extensive output from several rules, the other measures (PD, PA) can be estimated from them (by subtracting information for simpler rules from more complex rules containing the simpler rules). If so, this might help to compare output. In practice, however, this comparison is difficult because of the different concepts of reliability and use of negative evidence built into the methods.

Pattern Discovery is built on a traditional pattern discovery foundation, and seeks patterns that exceed a threshold. CliniMiner® seeks to identify relationships between variables through correlation, and then computes a FANO mutual information index for the rules. CliniMiner® can deliver complicated rules (of complexity greater than 4) if (a) Monte Carlo rather than exact sampling is used, or (b) provided that data is numeric *and* has meaningful multivariance. In the latter case, it starts with the assumption that

rules are so complex as to involve every parameter, and then removes poorly contributing parameters in a data fitting process involving global minimization. If approach (a) is to be accurate, however, it requires enormous amounts of data that increase dramatically with rule complexity.

Typically for qualitative data, PD tends to identify more complicated rules economically, and exhaustively enumerates all of those that exceed the support threshold. However, PD suffers from a combinatorial explosion in different ways than CliniMiner®. For example, the combinatorial effects of abundantly strong correlations, such as in therapeutic drug cocktails, are difficult for CliniMiner® to efficiently compute in that they lead to massive output and require additional set theory pruning algorithms, but are relatively easy for PD. An advantage of CliniMiner® is that it is capable of identifying relationships that occur with rates less than would be expected by chance, even if they never occur at all. Pattern Discovery would require tracking not only conditions for particular values, but also all of their complement sets. This would lead to combinatorial problems for PD. Otherwise, while PD can potentially pick up longer, more complicated rules, this advantage is offset in the loss of the more rare events that score below the threshold.

Unlike CliniMiner® and PD, PA is a form of outcome analysis. The rules predict the outcome of a column from the conditions in all of the other columns with measures of false positives AND false negatives, together with other joint measures of confidence. The algorithms that learn the rules are therefore more constrained than CliniMiner® and PD. While all three methods produce rules that can be evaluated as true or false, PA also constructs thresholds from the entire possible space of values. It also shares the use of the training and test set methodology with PD.

2.2. Data assembly

The University of Virginia Department of Public Health Sciences built and compiled 667,000 individual patient records (Human Investigation Committee protocol 10932) into a spreadsheet form (dating from 1993-present), one row of 208 core columns per patient (query required 80 h for data extraction; data compilation partially represented in Table 1). The UVA CDR is a comprehensive clinical and administrative relational (MySQL) data warehouse (30GB in size) that uses the Linux (Red Hat 9.0) operating system on a Dell 400 MHz dual processor server. It contains laboratory, microbiologic, and other electronic data for over one million in- and outpatient visits at the University of Virginia Health System from 1992 forward, from admission to discharge [1,34].

Prior to inclusion, each record was de-identified according to HIPAA regulations. This required the removal of 18 unique identifiers [32]. Thirty conditions (based on the ICD9-CM codes of [35]) (Table 2), 24 laboratory test categories (Table 3), 23 procedure groupings (Table 4), and 32 distinct medications types (Table 4) were included in the analysis. Due to formatting requirements, time was omitted as a variable in the patient records. For each laboratory test, the "First", "Last", "Average (Avg)", and "Total Count (Cnt)" values were initially extracted for each patient, however, because all four values were highly consistent the first values were used in the analysis. These data were transferred (via file transfer protocol [FTP]) to IBM researchers located in New York and Israel for processing.

2.3. Preparation

A previously assembled and experienced team of IBM researchers (the IBM HealthMiner® and MED-II teams) explored and performed initial processing of the data for IBM, resulting in a lengthy spreadsheet of

Table 1
Representative patient record compilation for analysis

PtID	YOB	G	R	S	YOD	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18
1	1994	M	W	A	2000	N	Y	N	Y	N	Y	N	N	Y	N	N	N	N	N	N	N	N	N
2	1923	F	W	D	1995	N	Y	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N	Y

D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
N	N	N	N	N	Y	N	N	N	N	N	N	Y	N	N	N	N	N	N	N	N	N	N	N
Y	N	Y	N	Y	Y	N	Y	N	N	N	N	N	Y	Y	N	N	N	N	Y	N	Y	N	N

P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	HCTFirst	HCTLast	HCTAvg	HCTCnt	PLTFirst	PLTLast
N	N	N	N	N	N	N	N	N	N	N	0	0	0	0	0	0
N	N	N	N	N	N	N	N	N	N	N	0	0	0	0	0	0

PLTAvg	PLTCnt	WBCFirst	WBCLast	WBCAvg	WBCCnt	GLUCFirst	GLUCLast	GLUCAvg	GLUCCnt	BUNFirst
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

BUNLast	BUNAvg	BUNCrit	CALCMFirst	Med1	Med2	Med3	Med4	Med5	Med6	Med7	Med8	Med9	Med10
0	0	0	0	1	2	0	26	9	3	10	3	1	6
0	0	0	0	1	1	0	0	3	1	1	10	1	1

Key: PtID = Patient identification number (randomized); YOB = Year of birth; YOD = Year of death; D1 = Diagnosis 1 (Comorbid condition, see Table 1); Px=Procedure grouping (see Table 4); HCT=Hematocrit; PLT=Platelet count; WBC=White blood cell; GLUC = Blood glucose; BUN = Blood urea, nitrogen; CALCM = Calcium, Med = Medication (see Table 5).

triplet comparisons (representative example, Table 5). CliniMiner® was extensively involved in preparing the data for use by all the data mining methods. All data, which were predominately in three states such as yes/don't know/no were converted to $-1/0/+1$. Laboratory data were converted to low, normal, and high ranges, which were then converted to $-1/0/+1$, respectively.

The CliniMiner® program was run on a variety of Unix, Linux, and Windows systems. Substantial progress could be made on a T40 1.6 GHz laptop with 1 gigabyte of RAM running for 24 h +. The query mechanism for CliniMiner® was a full "seek all interesting rules" without bias. The PA program was run on an Intel XEON 2.2 GHz CPU (512MB RAM) and took 90 min to complete. The PD program was executed on 24 CPUs (450 MHz processors) and was completed in 45 min.

2.4. Formal rule and pattern extraction

As noted in Section 2.1, CliniMiner® was the primary tool used in these initial studies to cleanse the data for the other two methods. The "*rule*" is the particular association A, B, C,. . .. FANO assesses the extent to which *Events* (items, entries, properties) occur together more, or less, than would be expected on a chance basis; rules were reported by CliniMiner® when there was mutual information content greater than +0.5 nat or less than −0.5 nat (this threshold is adjustable). This means that reported rules occurred $e^{0.5} = 1.6487\ldots$ times more than expected or $e^{0.5} = 1.6487\ldots$ times less than expected, where

Table 2
Comorbid conditions included in the analysis

Description	Number of patients
Congestive heart failure	28,054
Cardiac arrhythmias	43,795
Valvular disease	29,628
Pulmonary circulation disorder	7878
Peripheral vascular disorder	22,799
Hypertension	90,457
Paralysis	11,630
Other neurological	36,935
Chronic pulmonary disease	50,771
Diabetes, uncomplicated	37,135
Diabetes, complicated	12,456
Hypothyroidism	21,916
Renal failure	5323
Liver disease	12,985
Peptic ulcer disease (excluding bleeding)	8581
AIDS	1747
Lymphoma	6313
Metastatic cancer	11,873
Solid tumor without metastasis	49,000
Rheumatoid arthritis/ collagen vascular diseases	10,657
Coagulopathy	13,946
Obesity	21,772
Weight loss	20,689
Fluid and electrolyte disorders	37,227
Blood loss anemia	5157
Deficiency anemias	29,977
Alcohol abuse	15,240
Drug abuse	6331
Psychoses	27,109
Depression	30,116

e is the base of the natural logarithm (i.e. 2.718 . . .). In other words, the observed frequency differed from expected by some 60%. However, attention focused on rules of approximately +1 nat and −1 nat and stronger, which is a ratio of approximately 3:1. The *Complexity* of each such determined "rule", which is also reported (Tables 5A, 6), is the number of associating properties or simple events, such as 5 for the conjoint event (A, B, C, D, E) there being in that example 5 symbols. CliniMiner® also reports the observed and expected frequencies of abundance, from which the *Information* measures are calculated.

Predictive Analysis produces measures of the significance of, and support for, each rule (Tables 5B, 7). The *Predictive Value* (tp/[tp+fp]) represents the percentage that is correct when the rule is true. *Sensitivity* (tp/[tp+fn]) is the percentage of total disease patients found when the rule is true. The *Specificity* (tn/[tn+fp]) is defined as the percentage of total non-disease patients found when the rule is false.

Table 3
Laboratory codes and description used in the analysis

Lab code	Lab description	Units
ALKP	Alkaline phosphatase	U/l
ALT	Alanine aminotransferase (GPT)	U/l
AST	Aspartate aminotransferase (GOT)	U/l
BUN	Urea, nitrogen, blood	mg/dl
CALCM	Calcium	mg/dl
CREAT	Creatinine, blood serum	mg/dl
GLUC	Glucose, blood	mg/dl
HCT	Hematocrit	%
K	Potassium	mmol/l
LDH	Lactate dehydrogenase	U/l
MG	Magnesium	mg/dl
NA	Sodium	mmol/l
PCO2	Carbon dioxide, partial pressure	mmHg
PH	Blood PH	
PHOS	Phosphorous	mg/dl
PLT	Platelets	k/ul
PO2	Oxygen pressure	mmHg
PTAV	Prothrombin time	s
PTINR	Prothrombin time, INR	s
PTTAV	Partial thrombolplastin time	s
TBIL	Bilirubin-total, blood	mg/dl
TP	Protein, total	g/dl
TSH	Thyroid stimulating hormone	uIU.ml
WBC	White blood cell	k/ul

Accuracy ([tp+tn]/[tp+tn+fp+fn]) is the percentage of correct decisions if the disease is predicted when the rule is true and a non-disease is predicted when the rule is false. Finally, *Prevalence* ([tp+fp]/[tp+tn+fp+fn]) indicates the percentage of diseased patients in the total population.

Pattern discovery/THOTH quotes the observed number of times the rule is seen, the *Fraction of consequent given antecedent* as a measure of $P(\text{A\&B\&C})/P(\text{A\&B})$ as a weight of the rule "If A & B then C" (Tables 5C, 8). The validation of the data, described later, involves searches of PUBMED and other sources for the occurrence of studies that include the simple events in relationship with each other. The output of PD was filtered to restrict the number of items that enthymemes could contain in order to facilitate database mining.

2.5. Rationality check

The data forms were mined by CliniMiner®, PA, and PD and the results were then examined manually in order to locate less expected relationships and any apparent anomalies. We then attempted to verify the resulting associations with existing medical knowledge in order to determine those that may be novel. This was done using published standards (PUBMED®, Web of Science®, and PsycINFO®). PubMed® was developed by the National Center for Biotechnology Information (NCBI) to provide access to

Table 4
Procedure groupings (Px) and patient medications (Med) used in the analysis

Code	Description	Code	Description
Px1	Diagnostic bronchoscopy and biopsy of bronchus	Med 1	Lidocaine
Px2	Blood transfusion	Med 2	Magnesium
Px3	Physical therapy exercises, manipulation, and other procedures	Med 3	Famotidine
Px4	Upper gastrointestinal endoscopy, biopsy	Med 4	Midazolam
Px5	Tracheoscopy and laryngoscopy with biopsy	Med 5	Furosemide
Px6	Diagnostic cardiac catheterization, coronary arteriography	Med 6	Morphine
Px7	Electrocardiogram	Med 7	Heparin
Px8	Cancer chemotherapy	Med 8	Dextrose
Px9	Lobectomy or pneumonectomy	Med 9	Cefazolin
Px10	Enteral and parenteral nutrition	Med 10	Dexamethasone
Px11	Respiratory intubation and mechanical ventilation	Med 11	Albuterol
Px12	Hemodialysis	Med 12	Ondansetron
Px13	Magnetic resonance imaging	Med 13	Prednisone
Px14	Computerized axial tomography (CT) scan head	Med 14	Diltiazem
Px15	Skin graft	Med 15	Propofol
Px16	CT scan chest	Med 16	Nitroglycerin
Px17	Diagnostic ultrasound of heart (echocardiogram)	Med 17	Clindamycin
Px18	Colonoscopy and biopsy	Med 18	Insulin
Px19	Diagnostic procedures on nose, mouth, pharynx	Med 19	Cyclosporine
Px20	Tracheostomy, temporary and permanent	Med 20	Omeprazole
Px21	Therapeutic radiology	Med 21	Ciprofloxacin
Px22	Coronary artery bypass graft (CABG)	Med 22	Metoprolol
Px23	Biopsy of liver	Med 23	Warfarin
		Med 24	Chemo-infusion
		Med 25	Cortrimoxazole
		Med 26	Chemo
		Med 27	Digoxin
		Med 28	Methylprednisolone
		Med 29	Gentamicin
		Med 30	Acyclovir
		Med 31	Any Antibiotic
		Med 32	Epo

biomedical literature citations, and includes MEDLINE® (dating 1966-present) and OLDMEDLINE® (dating 1951–1965). MEDLINE® is the National Library of Medicine's (USA) premier database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. MEDLINE® contains bibliographic information from over 4,800 biomedical journals published in the United States and over 70 countries. The ISI Web of Science® (The Thomson Corporation) is a multidisciplinary collection of bibliographic material from over 8,600 scholarly journals (dating 1981–2004). It is comprised of five databases: Science Citation Index Expanded™, Social Sciences Citation Index®, Arts & Humanities Citation Index®, Index Chemicus®, and Current Chemical Reactions®. PsycINFO® is a database produced by the American Psychological Association that contains more than 1900 titles of psychological relevance (dating 1894-present).

Table 5
Representative output from the three HealthMiner® algorithms

A. Representative FANO triplet data output

Info.	Complexity	Saw	Expected	Event 1	Event 2	Event 3
3.4	3	1565	51.54	BUNFirst(Urea_Nitrogen _Blood_mg/dl: = > 0.7)	CREATFirst(CREATINE _BLOOD_SERUM _mg/dl: = > 0.06	Renal failure: = > − 0.85
2.91	3	871	46.6	CREATFirst(CREATINE _BLOOD_SERUM_mg/dl: = > 0.06	Diabetes_Complicated: = > − 0.76	Renal failure: = > − 0.85
2.84	3	774	44.61	BUNFirst(Urea_Nitrogen _Blood_mg/dl: = > 0.7	Diabetes_Complicated: = > − 0.76	Renal failure: = > − 0.85

B. Representative Predictive Analysis output

Cardiac arrhythmias	
[Congestive heart failure & age at diagnosis > 7.500]	
OR [Rx:Digoxin & Rx:Nitroglycerin < 6.500]	
Predictive value	68.04%
Sensitivity	52.46%
Specificity	95.78%
Accuracy	89.44%
Prevalence	14.62%

C. Representative Pattern Discovery output

% Cluster 30			
0.830986	Gender = Male	AND	Cardiac_arrhythmias = Positive
	AND Valvular_disease = Positive	IMPLIES	Race = White
0.741784	Gender = Male	AND	Cardiac arrhythmias = Positive
	AND Valvular_disease = Positive	IMPLIES	Hypertension = Pos

Search strategies were conducted as directed by the instructions for each database: PubMed® http://www.(ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html), Web of Science® (http://www.isinet.com/tutorials/wos6/wos6tut5.html), and PsycINFO® (http://www.apa.org/psycinfo/training/apa.pdf). For most of the searches, the Boolean operator "AND" was used to combine search terms (Tables 6–8). For the PubMed® searches, we started with a three-term phrase, for example (cardiac arrhythmias) (respiratory tract diseases) (heart valve diseases) and used PubMed®'s Automatic Term Mapping to convert it to (("arrhythmia"[TIAB] NOT Medline[SB]) OR "arrhythmia" [MeSH Terms] OR cardiac arrhythmias[Text Word]) AND ("respiratory tract diseases" [MeSH Terms] OR respiratory tract diseases[Text Word]) AND ("heart valve diseases" [MeSH Terms] OR heart valve diseases [Text Word]) in order to simultaneously increase the sensitivity and specificity of the search. We also chose key words that would match the MeSH terms to the ICD-9 codes used in this study. For the PubMed searches, the search phrases were enclosed in parentheses () in order to instruct processing as a unit and then incorporated into the overall strategy (Tables 6–8).

3. Results

3.1. CliniMiner® data trend characterization

Estimation of the percentages of "unknown", "less well known", and "established" biomedical knowledge from the data rules was calculated using a representative equal probability sampling method (EPSEM), Simple Random Sampling, with a sampling ratio of approximately one percent, hence 280 associations out of the total 27,764 triplets from the CliniMiner® output. Of that fraction, rules with negative *Information* values and "<" *Event* signs were removed, leaving a total of 75 rules. Each of the remaining triplet *Event* terms was submitted to PubMed® as previously described, and the results were tabulated. Triplets with six or more citations were considered to be "well established", one to five were "less well known", and zero were potentially "unknown". Eighty-six percent of the rules (53% well-established, 33% less well-known) were found in the scientific literature using PubMed®. Fourteen percent of the triplet associations had zero citations in PubMed®, and were then further queried in the Web of Science® and PsycINFO® databases (resulting in 0 citations).

3.2. CliniMiner®: medically-known correlations

A number of well-published medical correlations were found within the dataset, and a selected subset is summarized in Table 6 . These triplet combinations include: alcohol abuse+drug abuse+AIDS [36]; alcohol abuse+depression+drug abuse [37,38]; and fluid and electrolyte disorders+AIDS+other neurological [39,40].

3.3. CliniMiner®: data anomalies

We developed a string-matching code to find triplets with similar structure, where the first and second components were exactly the same, while the third *Event* was slightly different in the direction of the $\langle$ or $\rangle$ sign associated with each *Event* (Table 5A). For example, a strong correlation between age at diagnosis ($=<56.46$) and blood loss anemia ($=>-0.87$) was associated with both elevated ($=>-0.37$) and low ($=<0.37$) deficiency anemias. There existed 4085 pairs with such similarities, however, in approximately 970 pairs their *Information* scores indicated that one of them occurred more than expected (+) while the other was less (−) than expected. Because of this, we believe that these potential "conflicts" are resolved, leaving 3115 triplet rules (22% of the results) that remain unresolved under this scenario.

In addition, there was a tendency for the occurrence of peptic ulcer disease ($=>-0.82$) and psychoses ($=>-0.62$) with *both* obesity ($=>-0.69$) and weight loss ($=>-0.52$). These data may, however, represent a real bifurcation in the patient population for these two disease profiles and were not considered to be in conflict. For example, some patients with peptic ulcer disease and psychoses may respond to their diseases by eating excessively, while others may consume too little.

3.4. CliniMiner®: medically "unknown" correlations- 3 examples

The following sets of triplet associations were manually crosschecked with the entire data set for internal repetitions or conflicts and were verified not to be among the 22% of the previously described results that were unresolved for conflict.

Table 6
Selected CliniMiner results vs. search engine literature publications (— = no information)

CliniMiner® rule	Search terms	PubMed results	Web of science results	Psych INFO results
Alcohol_abuse: = > − 0.75;	(alcohol-related disorders) (drug abuse) (AIDS)	343	—	—
Drug_abuse: = > − 0.88;	(TS = alcohol-related disorders) AND	—	0	—
AIDS: = > − 0.97	(TS = drug abuse) AND (TS = AIDS)			
Expected: 3;	(exp ALCOHOLISM/ OR exp Alcoholic Psychosis/	—	—	1
Saw: 47;	OR exp Alcohol Intoxication/) AND (exp Drug			
Complexity: 3;	Dependency/ OR exp DRUGS/ OR drug dependence.mp.			
Information: 2.6	OR exp OPIATES/) AND (exp Acquired Immune			
	Deficiency Syndrome/)			
Alcohol_abuse: = > − 0.75;	(Alcohol-related disorders) (Drug abuse) (Depression)	3467	—	—
Depression: = > − 0.56;	(TS = alcohol-related disorders)	—	0	—
	AND (TS = drug abuse) AND (TS = depression)			
Drug_abuse: = > − 0.88	(Alcoholism/ or alcoholic psychosis/ or	—	—	12
Expected:61.85; Saw: 721;	alcohol intoxication/) AND (exp Drug Dependency/ OR drug			
Complexity:3,	dependence.mp. OR exp OPIATES/			
Information: 2.44	OR exp DRUGS/) AND (exp Dysthymic Disorder/ OR neurotic			
	depression.mp. OR depressive reaction.mp.)			
Fluid_and_elctrolyte_disorders	(Water-electrolyte imbalance OR acid-base	40	—	—
: = > − 0.18;	imbalance) (AIDS) (Neurological Disorders)			
AIDs: = > − 0.97;	(TS = water-electrolyte imbalance OR acid-base imbalance)	—	0	—
Other_neurological: = > − 0.55	AND (TS = AIDS) AND (TS = Neurological Disorders)			
Expected:39.15;	(exp DEHYDRATION/ OR acidosis.mp. OR alkalosis.mp.	—	—	0
Saw: 99;	OR exp POTASSIUM/ OR hyperkalemia.mp. OR			
Complexity: 3;	hypokalemia.mp. OR exp ELECTROLYTES/)			
Information: 0.91	AND (exp PARKINSONISM/ OR huntington's chorea.mp. or exp			
	Huntingtons Disease/ OR multiple sclerosis.mp. or exp			
	Multiple Sclerosis/ OR schilder's disease.mp. OR exp			
	EPILEPSY/ OR nonconvulsive epilepsy.mp.)			
	AND (exp Acquired Immune Deficiency Syndrome/)			
Paralysis: = > − 0.83;	(Paralysis) (Peptic ulcer disease) (Renal failure)	3[a]	—	—
Peptic_ulcer_disease: = > −0.82;	(TS = paralysis) AND (TS = peptic	—	0	—
Renal_failure: = > − 0.85	ulcer disease) AND (TS = renal failure)			
Expected = 21.07;	(exp Gastrointestinal Ulcers/ or peptic ulcer disease.mp.)	—	—	0
Saw = 76,	AND (exp Organ Transplantation/ or exp			
Information = 1.26,	Hemodialysis/ or renal failure.mp.)			
Complexity = 3	AND (exp PARALYSIS/ or paralysis.mp.)			
Paralysis: = > − 0.83;	(Paralysis) (Peptic ulcer disease)	0	—	—
Peptic_ulcer_disease: = > −0.82;	(Rheumatoid arthritis)			
Rheumatoid_arthritis_collagen	(TS = paralysis) AND (TS = peptic ulcer disease)	—	0	—
_vascular_disease: >−0.87	AND (TS = rheumatoid arthritis)			

Table 6 (*continued*)

CliniMiner® rule	Search terms	PubMed results	Web of science results	Psych INFO results
Expected = 18.61; Saw = 48, Information = 0.93, Complexity = 3	(exp Gastrointestinal Ulcers/ or peptic ulcer disease.mp.) AND (exp Organ Transplantation/ or exp Hemodialysis/ or renal failure.mp.) AND (exp Rheumatoid Arthritis/ OR lupus.mp. or exp LUPUS/)	—	—	0
Paralysis: = > − 0.83; Peptic_ulcer_disease: => −0.82; Psychoses: = > − 0.62 Expected = 55.42;	(paralysis) (peptic ulcer disease) (psychotic disorders OR bipolar disorder OR schizophrenia OR paranoid disorders)	0	—	—
Saw = 166, Information = 1.09, Complexity = 3	(TS = paralysis) AND (TS = peptic ulcer disease) AND (TS = psychotic disorders OR biopolar disorder OR schizophrenia OR paranoid disorders)	—	0	—
	(exp Gastrointestinal Ulcers/ or peptic ulcer disease.mp.) AND (exp Organ Transplantation/ or exp Hemodialysis/ or renal failure.mp.) AND (schizophrenia.mp. OR exp Schizophrenia/ or exp Psychosis/ or psychotic disorders.mp. OR paranoid disorders.mp. or exp "Paranoia (Psychosis)"/ OR bipolar disorder.mp. or exp Bipolar Disorder/)	—	—	0

[a]Upon review of the manuscripts, these articles were unrelated to the ICD-9 codes used in this study.

Paralysis/peptic ulcer disease/renal failure

A strong correlation (expected = 21.07, saw = 76, information = 1.26, complexity = 3) was observed between paralysis (= > − 0.83), peptic ulcer disease (= > − 0.82), and renal failure (= > − 0.85). A search of these three combined terms (paralysis) (peptic ulcer disease) (renal failure) using PubMed® yielded three Refs. [41–43]; however, upon closer inspection these sources examined the impact of surgical procedures on one or more of the three terms, but did not directly correlate the three together. The Web of Science® did not yield any references. It should be noted that the clinical manifestations of chronic renal failure are known to include congestive heart failure, weak bones, stomach ulcers, and damage to the central nervous system (among a lengthy list of other symptoms) [44].

Paralysis/peptic ulcer disease/rheumatoid arthritis

The correlation between paralysis (= > − 0.83), peptic ulcer disease (= > − 0.82), and rheumatoid arthritis (= > − 0.87) was strong (expected = 18.61, saw = 48, information = 0.93, complexity = 3). No publications were found using the PubMed®, Web of Science®, or PsycINFO® databases (Table 6). The association between peptic ulcer disease and rheumatoid arthritis alone is unremarkable given that the risk of peptic ulcer formation with the use of NSAIDs for the relief of pain and inflammation of rheumatoid arthritis [45] is well known. In addition, cervical spinal involvement in patients with rheumatoid arthritis can result in quadriplegia [46].

Paralysis/peptic ulcer disease/psychoses

A strong correlation (expected = 55.42, saw = 166, information = 1.09, complexity = 3) was observed between paralysis (= > −0.83), peptic ulcer disease (= > −0.82), and psychoses (= > −0.62). A search of these three combined terms using the PubMed®, Web of Science®, and PsycINFO® databases did not yield any supporting references (Table 6). Previous work has reported an association between peptic ulcer disease and organic psychoses as a result of drug therapy [47,48]. Alternatively, work examining the effects of corticotropin therapy in multiple sclerosis (a disease that can lead to paresis and plegia) found that both psychosis and ulcers were potential side effects of treatment [49].

3.5. Predictive analysis trend characterization

Estimation of the percentages of "unknown", "less well known", and "established" biomedical knowledge for the PA algorithm was calculated as previously described. Given the small number of rules generated using this method, a random sampling was unnecessary. Of the 120 rules examined, 73 (61%) were established, 18 (15%) were less well known, and 29 (24%) were unkown in the published biomedical literature.

3.6. Predictive analysis medically known correlations

A selected subset of PA rules that were found to be well known in the PubMed®, Web of Science®, and PsycINFO® databases are included in Table 7 . They include: hypertension+renal failure+age at diagnosis [50,51], liver disease+biopsy of liver+total protein [52,53], and psychoses+drug abuse+depression [54,55].

3.7. Predictive analysis medically unknown correlations—3 examples

Famotidine/midazolam/magnesium/any antibiotic

The correlation between prescription famotidine, prescription midazolam > 2.5, prescription magnesium, and any antibiotic > 1.500 was strong (predictive value: 73%, sensitivity: 52.25%, prevalence: 14.16%). Zero references were found (Table 7) using PubMed, Web of Science®, or PsycINFO®. It is possible, however, that these terms are associated with the management of cancer pain [56]. For example, antibiotics are used to relieve the pain associated with infections, famotidine for the prevention of NSAID-related peptic ulceration, and midazolam for relief of anxiety accompanying pain [56].

Omeprazole/magnesium/liver disease

Prescription of both omeprazole and magnesium was associated with liver disease (predictive value: 65.41%, sensitivity: 7.34%, prevalence: 5.69%). No references were found for this association using PubMed®, the Web of Science®, or PsychINFO® databases (Table 7). The association between omeprazole and liver disease is not entirely surprising, however, given that in rare instances liver disease has been associated with omeprazole usage [57].

Albuterol/tracheostomy temporary and permanent/magnesium

A strong correlation between the prescription of albuterol and magnesium was associated with temporary and permanent tracheostomies (predictive value: 67.31%, sensitivity: 19.71%, prevalence: 11.82%).

Table 7
Selected Predictive Analysis (PA) results vs. search engine literature publications (— = no information)

PA rule	Search terms	PubMed results	Web of science results	Psych INFO results
10. Hypertension	(Hypertension) (Renal failure) (Age)	3382	—	—
[Renal Failure & Age at diagnosis > 12.000]	(TS = hypertension) AND	—	1094	—
Predictive Value: 75.17%;	(TS = renal failure) AND (TS = age)			
Sensitivity: 55.40%;	(exp HYPERTENSION/ or exp	—	—	3
Specificity: 94.48%;	ESSENTIAL HYPERTENSION/ or hypertension.mp.)			
Accuracy: 85.42%;	AND (exp Organ Transplantation/			
Prevalence: 23.18%	OR exp HEMODIALYSIS/ OR renal failure.mp.)			
	AND (age.mp.)			
18. Liver Disease	(Liver disease) (Biopsy of liver) (Total protein)	1299	—	—
[Biopsy of liver & TPFirst (PROTEIN TOTAL g/dL) > − 0.500]	(TS = liver disease)	—	0	—
Predictive Value: 77.70%;	AND (TS = biopsy of liver) AND (TS = total protein)			
Sensitivity: 16.54%;	(exp "Cirrhosis (Liver)"/ or exp Hepatitis/	—	—	0
Specificity: 99.84%;	or liver disease.mp.) AND (biopsy of liver.mp.)			
Accuracy: 97.16%;	AND (total protein.mp.)			
Prevalence: 3.22%				
33. Psychoses	(Psychotic disorders OR bipolar disorder	1690	—	—
[Drug abuse & Depression]	OR schizophrenia OR paranoid disorders)			
Predictive Value: 71.39%;	(Drug abuse) (Depression)			
Sensitivity: 9.14%;	(TS = psychotic disorders OR TS = bipolar disorder	—	129	—
Specificity: 99.67%;	OR TS = schizophrenia OR TS = paranoid			
Accuracy: 92.28%;	disorders) AND (TS = drug abuse)			
Prevalence: 8.16%	AND (TS = depression)			
	(schizophrenia.mp. or exp SCHIZOPHRENIA/	—	—	44
	OR exp PSYCHOSIS/ OR psychotic			
	disorders.mp. OR paranoid disorders.mp.			
	or exp "Paranoia (Psychosis)"/ OR bipolar			
	disorder.mp. or exp Bipolar Disorder/)			
	AND (exp Opiates/ or exp Drug Dependency/			
	or exp Drugs/ or drug dependence.mp.)			
	AND (exp Dysthymic Disorder/ OR neurotic			
	depression.mp. ORdepressive reaction.mp.)			
83. Rx: Famotidine	(Famotidine) (Midazolam) (Magnesium) (Antibiotic)	0	—	—
[Rx: Midazolam > 2.500 & Rx: Magnesium & Rx: Any Antibiotic > 1.500]				
Predictive Value: 73.00%,	(TS = midazolam) AND (TS = antibiotic)	—	0	—
Sensitivity: 52.25%,	AND (TS = magnesium) AND (TS = famotidine)			
Accuracy: 90.50%,	(exp MIDAZOLAM/ or midazolam.mp.)	—	—	0
Prevalence: 14.16%;	AND (antibiotic.mp. or exp ANTIBIOTICS/)			
Specificity: 96.81%	AND (famotidine.mp.) AND			
	(exp MAGNESIUM/ or magnesium.mp.)			

Table 7 (*continued*)

PA rule	Search terms	PubMed results	Web of science results	Psych INFO results
91. Rx: Albuterol	(Albuterol) (Tracheostomy) (Magnesium)	0	—	—
[Tracheostomy temporary and permanent	(TS = magnesium) AND (TS = albuterol)	—	0	—
& Rx: Magnesium]	AND (TS = tracheostomy)			
Predictive Value: 67.31%;	(albuterol.mp.) AND (tracheostomy.mp.)	—	—	0
Sensitivity: 19.71%;	AND (exp MAGNESIUM/ or magnesium.mp.)			
Specificity: 98.72%;				
Accuracy: 89.38%;				
Prevalence: 11.82%				
100. Rx: Omeprazole	(Omeprazole) (Magnesium) (Liver disease)	0	—	—
[Rx: Magnesium > 13.500 & Liver Disease]	(TS = magnesium) AND (TS = liver	—	0	—
Predictive Value: 65.41%,	disease) AND (TS = omeprazole)			
Sensitivity: 7.34%,	(exp MAGNESIUM/ or magnesium.mp.)	—	—	0
Specificity: 99.77%,	AND (exp "Cirrhosis (Liver)"/ or exp Hepatitis/			
Accuracy: 94.50%,	or liver disease.mp.) AND (omeprazole.mp.)			
Prevalence: 5.69%				

This association has a potential clinical rationale given that albuterol is frequently used as a treatment for patients with chronic pulmonary disease (who may also be candidates for tracheotomies). The association of magnesium with these two conditions may be related to an underlying impact on strength of the respiratory musculature; weakness may lead to the need for mechanical ventilation support and tracheotomy. No references were found, however, for this association using PubMed®, the Web of Science®, or PsycINFO® databases (Table 7).

3.8. Pattern discovery data trend characterization

The rules generated by the PD program were examined for "unknown", "less well known", and "established" biomedical knowledge, as described. One hundred rules were randomly examined, and of those 75 were removed from consideration because they included negative information (i.e. low glucose). The remaining 25 consisted of 6 (24%) well-known, 8 (32%) less well known, and 11 (44%) unknown associations in the biomedical literature (Tables 5C, 8).

3.9. Pattern discovery: medically-known correlations

Three medically known associations were generated by PD, and verified as previously described, are summarized in Table 8. They include: valvular disease+warfarin+cardiac arrhythmias [58,59]; cardiac arrhythmias+valvular disease+echocardiogram+congestive heart failure [60,61], and congestive heart failure+valvular disease+hypertension [62,63].

Table 8
Selected pattern discovery (PD) results vs. search engine literature publications (–= no information)

PD rule	Search terms	PubMed results	Web of science results	Psych INFO results
% Cluster 34	(Heart valve diseases) (Warfarin)	53	—	—
Valvular_disease = Positive AND Rx:	(Cardiac arrhythmias)			
Warfarin = Filled IMPLIES	(TS = heart valve diseases) AND (TS = warfarin)	—	0	—
Cardiac_arrhythmias = Positive	AND (TS = cardiac arrhythmias)			
Fraction of consequent given antecedent: 0.74206;	(heart valve diseases.mp. or exp	—	—	0
Num. Var.: 3; Type Score: 9	Heart Valves/) AND (warfarin.mp.) AND			
	(exp "Arrhythmias (Heart)"/ or cardiac arrhythmias.mp.)			
% Cluster 14	(cardiac arrhythmias) (heart valve diseases)	98	—	—
Cardiac_arrhythmias = Positive AND	(echocardiogram) (congestive heart failure)			
Valvular_disease = Positive AND Echocardiogram =	(TS = cardiac arrhythmias) AND (TS = heart	—	0	—
Performed IMPLIES Congestive_heart	valve diseases) AND (TS = echocardiogram)			
_failure = Positive	AND (TS = congestive heart failure)	—	—	0
Fraction of consequent given antecedent: 0.75439;	(heart valve diseases.mp. or exp Heart Valves/)			
Num. Var.:4; Type Score: 12	AND (exp "Arrhythmias (Heart)"/ or			
	cardiac arrhythmias.mp.) AND			
	(congestive heart failure.mp.) AND			
	(echocardiogram.mp.)			
% Cluster 14	(Congestive heart failure) (Heart valve diseases)	550	—	—
Congestive_heart_failure = Positive AND	(Hypertension)			
	(TS = congestive heart failure) AND (TS = heart valve diseases)	—	1	—
Valvular_disease = Positive IMPLIES	AND (TS = hypertension)			
Hypertension = Positive	(congestive heart failure.mp.)			
Fraction of consequent given antecedent: 0.76739;	AND (exp HYPERTENSION/ or exp ESSENTIAL			
Num. Var.:3; Type Score: 9	HYPERTENSION/ or hypertension.mp.)			
	AND (heart valve diseases.mp. or exp Heart	—	—	0
	Valves/)			

% Cluster 319 Diabetes_uncomplicated=Positive AND Physical_therapy_exercises_manipulation_etc= Performed AND Computerized_axial_tomography_(CT)_scan_head= Performed IMPLIES Fluid_and_electrolyte_disorders=Positive Fraction of consequent given antecedent: 0.71429; Num. Var.:4; Type Score: 12	(Acid-Base Imbalance OR Water-Electrolyte Imbalance) AND (diabetes) (physical therapy) (head CT scan) (TS = Acid base imbalance OR TS = water eletcrolyte imbalance) AND (TS = diabetes) AND (TS = physical therapy) AND (TS = head ct scan) (exp DEHYDRATION/ OR acidosis.mp or alkalosis.mp. OR exp POTASSIUM/ OR hypokalemia.mp OR exp ELECTROLYTES/ OR hypokalemia.mp.) AND (exp DIABETES MELLITUS/) AND (physical therapy.mp.) AND (head ct scan.mp.)	0 — — —	— 0 — —	— — — 0
% Cluster 1289 Deficiency_anemias = Positive AND Rx: Omeprazole = Filled IMPLIES Hypertension = Positive Fraction of consequent given antecedent: 0.72; Num. Var.:3; Type Score: 9	(Deficiency anemias) (Omeprazole) (Hypertension) (TS = iron deficiency anemias) AND (TS = omeprazole) AND (TS = hypertension) (omeprazole.mp.) AND (exp HYPERTENSION/ or exp ESSENTIAL HYPERTENSION/ or hypertension.mp.) AND (iron deficiency anemia.mp.)	0 — — —	— 0 — —	— — — 0

3.10. Pattern discovery: medically-unknown correlations: 2 examples

Diabetes/physical therapy/head CT scan/fluid and electrolyte disorder

A strong correlation (fraction of consequent given antecedent: 0.71429, Type Score: 12) exists between diabetes uncomplicated, physical therapy, head CT scans, and fluid and electrolyte disorder (Table 8); however, no citations were found in the literature to support these associations.

Deficiency anemias/omeprazole/hypertension

The correlation between deficiency anemias, omeprazole, and hypertension was strong (fraction of consequent given antecedent: 0.72, type score: 9) (Table 8), but zero references were found in the biomedical literature to support this combination of terms. A strong association has, however, been reported between iron deficiency anemia and long-term ingestion of omeprazole [64]. Additionally, experimental animal models have demonstrated that maternal iron restriction during pregnancy causes hypertension in adult offspring due to a deficit in nephron number [65].

4. Discussion

The use of large repositories of patient-specific biological, clinical, and associated administrative data generated during the routine delivery of medical care has historically been limited to utilization management, quality assurance, and more recently, disease management. Selected portions of these data have also been incorporated into research protocols and studies, usually within disease or procedure-specific retrospective or prospective studies. In general, however, the data generated through routine care procedures have not been considered of sufficient quality and integrity to use as the sole and primary source of data for clinical research, especially research examining new approaches to diagnosis or treatment, including new pharmaceutical agents or devices.

With increasing reliance on primary electronic capture of a wide variety of clinical data, and increasingly biological data, the quality and integrity of the resulting clinical repositories has improved. Large-scale associations among a wider "population-based" repository of clinical and biological data that have no a priori assumptions can facilitate in the generation of new hypotheses that may subsequently stimulate confirmatory experimentation. This approach is attractive because it has the potential to generate new insights into basic biological and applied clinical applications at a very low cost.

In this preliminary study, we examined a large clinical dataset using three distinct data mining approaches: CliniMiner®, Predictive Analysis, and Pattern Discovery. We found many correlations or "rules" abstracted by the data that appear to be reflections of well-established medical associations, such as the relationship between drug and alcohol abuse and AIDS. In the future, filtering tools could eliminate these well-established associations. We then isolated an additional subset of associations that were confirmable by references in the published scientific/clinical literature. The remainder of the reported associations can be classified in a variety of ways; some appear novel and plausible, meaning that they have validity and may be worthy of further investigation. For example, the novel reporting of psychoses and peptic ulcer disease with paralysis may simply represent the association of three relatively severe conditions with one prompting the subsequent development of the others. More interestingly, this association might point to an underlying common inflammatory, autoimmune, or even infectious etiology.

We conclude from these results that unsupervised data mining of large clinical repositories is feasible. The records used in this project were minimally processed and the categories chosen for inclusion were

very limited subsets of more comprehensive data that are available. This greatly constrained the number and complexity of the potential associations. None-the-less, these preliminary associations appear to have potential utility. These results may also represent a first step toward the use of large quantities of biological and clinical data as the basis for new approaches to scientific discovery and hypothesis generation. We would emphasize, however, that much more work needs to be completed before such efforts are widely implemented. In addition, medical reference databases may find it useful to require that all authors explicity codify the clinical components of their work using standards such as the International Classification of Diseases. This would greatly speed the automation of identifying potentially novel associations between searches, like the ones presented here, with the medical literature.

Our team is currently expanding the size of the database used in this study and plans to extend its components to include acute diagnoses, detailed pathology reports, and patient outcomes. These categories should substantially expand the potential associations. We will also merge data from the VCU data warehouse (for which we have a joint services agreement) to expand our existing patient cohort, and plan to use new representation modifier capabilities, such as the Medical Language Extraction and Encoding System (MedLEE). MedLEE is an application to extract, structure, and encode clinical information in textual patient reports so that the data can be used by subsequent automated processes [66]. The application of this technology will permit us to use the textual data contained within UVA's CDR to be represented in HL-7 and/or XML for further processing.

For filtering the data mining results through comparison with the existing biomedical literature, we will employ tools, such as Collexis, that provide new capabilities to represent the relationships from full text articles in a semantic network that then can be more directly compared to the data mining output. Given a full set of documents, Collexis constructs a concept fingerprint of each document, which is then stored in a catalog. The software reads the collection of fingerprints, and creates the associative concept space (ACS); this is then stored in a database. The API browser visualizes the ACS models and is used to: (i) input a seed term then output/find all related concepts, (ii) input concepts, output a path between them (hypothesis testing), and (iii) retrieve references that support the found relationships [67]. Finally, we hope to develop new methods to combine these two outputs, the associations from large data repositories and new representations of biomedical knowledge, in ways that would more directly and efficiently lead to the generation of new ideas.

5. Summary

This report provided the initial results from an unsupervised data mining search of 667,000 clinical records that were compiled from an academic medical center data repository using a new data mining approach, HealthMiner®. These data contained comprehensive demographic, socio-economic, clinical, and in selected cases, biological and outcomes information. Our principal goal was to investigate the potential value of searching these databases, without bias, for novel biomedical insights.

HealthMiner® consists of three clinical data mining tools: CliniMiner® (also referred to as FANO in earlier publications), Predictive Analysis, and Pattern Discovery. These methods are related in that they are unsupervised rule discovery techniques. The majority of rules generated for CliniMiner® and Predictive Analysis represented well-established medical knowledge that could be directly confirmed with reference to the biomedical literature. A minority of the associations reported were unknown to the

published literature, however, and, upon further examination, may represent useful knowledge for hypothesis generation and experimentation. For example, CliniMiner® identified a strong relationship between the co-occurrence of paralysis+peptic ulcer+rheumatoid arthritis. Input of these combined terms into three large, national reference databases yielded zero information regarding their relatedness, signifying that this triplet association was a candidate for further academic consideration.

We conclude that it is feasible to combine and apply large-scale data mining search tools to complex clinical datasets. Although much work remains to be accomplished to make this approach widely applicable, it holds promise as a potentially valuable alternative to traditional hypothesis-driven scientific discovery. This effort may represent a first step in the development of a non-hypothesis driven approach to scientific discovery based on information obtained from a large clinical data repository. We are currently collaborating with Virginia Commonwealth University to expand the scope and information of our electronic patient records for continued knowledge discovery.

Acknowledgements

This work was supported in part by the University of Virginia School of Medicine Grant DR00907 (W.A. Knaus) and the Virginia Tobacco Settlement Foundation Grant 8520003 (W.A. Knaus). The faculty of the University of Virginia and Virginia Commonwealth University declare that they have no financial interests in the research or algorithms described in this manuscript. The authors would like to thank J. L. Preston for her technical assistance in preparation of this manuscript.

References

[1] K.W. Scully, R.D. Pates, G.S. Desper, A.F. Connors, F.E. Harrell, K.S. Pieper, R.L. Hannan, R.E. Reynolds, Development of an enterprise-wide clinical data repository: Merging multiple legacy databases, J. Am. Med. Inform. Assoc. (Suppl.) S (1997) 32–36.

[2] J.S. Einbinder, K. Scully, Using a clinical data repository to estimate the frequency and costs of adverse drug events, J. Am. Med. Inform. Assoc. (Suppl.) S (2002) S34–S38.

[3] J.H. Holmes, D.R. Durbin, F.K. Winston, Discovery of predictive models in an injury surveillance database: An application of data mining in clinical research, Proc. AMIA Symp. (2000) 359–363.

[4] S.M. Downs, M.Y. Wallace, Mining Association rules from a pediatric primary care decision support system, Proc. AMIA Symp. (2000) 200–204.

[5] S.E. Brossette, A.P. Sprague, J.M. Hardin, K.B. Waites, W.T. Jones, S.A. Moser, Association rules and data mining in hospital infection control and public health surveillance, J. Am. Med. Inform. Assoc. 5 (1998) 373–381.

[6] J.C. Prather, D.F. Lobach, L.K. Goodwin, J.W. Hales, M.L. Hage, W.E. Hammond, Medical data mining: Knowledge discovery in a clinical data warehouse, Proc. AMIA Symp. (1997) 101–105.

[7] D. Haughton, J. Deichmann, A. Eshghi, S. Sayek, N. Teebagy, H. Topi, A review of software packages for data mining, Am. Stat. 57 (2003) 290–309.

[8] H.A. Morrow-Jones, E.G. Irwin, B. Roe, Consumer preferences for neotraditional neighborhood characteristics, Housing Policy Debate 15 (2004) 171–202.

[9] J.R. Brence, D.E. Brown, Data mining corrosion from eddy current non-destructive tests, Comp. Ind. Eng. 43 (2002) 821–840.

[10] A.G. Dean, R.F. Fagan, B.J. Panter-Connah, Computerizing Public Health Surveillance Systems, Principles and Practice of Public Health Surveillance, Oxford University Press, New York, 1994, pp. 200–217.

[11] V. Maojo, C.A. Kulikowski, Bioinformatics and medical informatics: collaborations on the road to genomic medicine?, J. Am. Med. Inform. Assoc. 10 (2003) 515–522.

[12] http://www.bisti.nih.gov/bisti_recommendations.cfm.
[13] H.-C. Lin, S. Xirasagar, Institutional factors in cesarean delivery rates: policy and research implications, Obstet. Gynecol. 103 (2004) 128–136.
[14] E.D. Peterson, L.P. Coombs, E.R. DeLong, C.K. Haan, T.B. Ferguson, Procedural volume as a marker of quality for CABG surgery, J. Am. Med. Assoc. 291 (2004) 195–201.
[15] B.J. Bock, C.T. Dolan, G.C. Miller, W.F. Fitter, B.D. Hartsell, A.N. Crowson, W.W. Sheehan, J.D. Williams, The data warehouse as a foundation for population-based reference intervals, Am. J. Clin. Pathol. 120 (2003) 662–670.
[16] H.T. Sorensen, S. Friis, B. Norgard, L. Mellemkjaer, W.J. Blot, J.K. McLaughlin, A. Ekbom, J.A. Baron, Risk of cancer in a large cohort of nonaspirin NSAID users: a population-based study, Br. J Cancer 88 (2003) 1687–1692.
[17] T.B. Ferguson Jr., L.P. Coombs, E.D. Peterson, Preoperative beta-blocker use and mortality and morbidity following CABG surgery in North America, J. Am. Med. Assoc. 287 (2002) 2221–2227.
[18] H.N. Reynolds, M. McCunn, U. Borg, N. Habashi, C. Cottingham, Y. Bar-Lavi, Acute respiratory distress syndrome: estimated incidence and mortality rate in a 5 million-person population base, Crit. Care (London) 2 (1998) 29–34.
[19] W.A. Knaus, D.P. Wagner, J. Lynn, Short-term mortality predictions for critically ill hospitalized adults: science and ethics, Science 18 (1991) 389–394.
[20] J.-R. LeGall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, J. Am. Med. Assoc. 270 (1993) 2957–2963.
[21] S. Lemeshow, D. Teres, J. Klar, J.S. Avrunin, S.H. Gehlbach, J. Rapoport, Mortality probability models based on an International cohort of intensive care unit patients, J. Am. Med. Assoc. 270 (1993) 2478–2486.
[22] M.M. Pollack, K.M. Patel, U.E. Ruttimann, PRISM III: an updated pediatric risk of mortality score, Crit. Care Med. 24 (1996) 743–752.
[23] C.E. Kennedy, N. Aoki, Generating a mortality model from a pediatric ICU (PICU) database utilizing knowledge discovery, Proc. AMIA Symp. (2002) 375–519.
[24] A.J. Butte, I.S. Kohane, Unsupervised knowledge discovery in medical databases using relevance networks, Proc. AMIA Symp. (1999) 711–715.
[25] S.-M. Lee, P.A. Abbott, Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers, J. Biomed. Inf. 36 (2003) 389–399.
[26] R. Korrapati, S. Mukherjee, K.V. Chalam, A Bayesian framework to determine patient compliance in glaucoma cases, Proc. AMIA Symp. (2000) 1050.
[27] E. Burnside, D. Rubin, R. Shachter, A Bayesian network for mammography, Proc. AMIA Symp. (2000) 106–110.
[28] D. Aronsky, P.J. Huag, Automatic identification of patients eligible for a pneumonia guideline, Proc. AMIA Symp. (2000) 12–16.
[29] B. Robson, Clinical and pharmacologic data mining; 1. Generalized theory of expected information and application to the development of tools, J. Proteom. Res. 2 (2003) 283–302.
[30] B. Robson, R. Mushlin, Clinical and pharmacologic data mining 2. A simple method for the combination of information from associations and multivariances to facilitate analysis decision and design in clinical research and practice, J. Proteom. Res. 3 (2004) 697–711.
[31] S. Weiss, N. Indurkhya, Lightweight rule induction, Proc. Int. Con. Mach. Learn. 23 (2000) 1135–1142.
[32] http://www.os.dhhs.gov/ocr/hipaa/.
[33] S. Weiss, R. Galen, R. Tadepalli, Maximizing the predictive value of production rules, J. Art. Intell. 45 (1990) 47–71.
[34] J.A. Lyman, K. Scully, S. Tropello, J. Boyd, J. Dalton, S. Pelletier, C. Egyhazy, Mapping from a clinical data warehouse to the HL7 reference information model, Proc. AMIA Symp. 23 (2003) 920.
[35] A. Elixhauser, C. Steiner, D.R. Harris, R.M. Coffey, Comorbidity measures for use with administrative data, Med. Care 36 (1998) 8–27.
[36] G. Lake-Bakaar, R. Grimson, Alcohol abuse and stage of HIV disease in intraveneous drug abusers, J. R. Soc. Med. 89 (1996) 389–392.
[37] W.M. Compton III, L.B. Cottler, A.B. Abdallah, D.L. Phelps, E.L. Spitznagel, J.C. Horton, Substance dependence and other psychiatric disorders among drug dependent subjects: Race and gender correlates, Am. J. Addict. 9 (2000) 113–125.
[38] S.L. Eames, J. Westermeyer, R.D. Crosby, Substance use and abuse among patients with comorbid dysthymia and substance disorder, Am. J. Drug Alcohol Abuse 24 (1998) 541–550.
[39] A. Tolaymat, F. Al-Mousily, J. Sleasman, S. Paryani, R. Neiberger, Hyponatremia in pediatric patients with HIV-1 Infection, S. Med. J. 88 (1995) 1039–1042.

[40] W. Van Paesschen, C. Bodian, H. Maker, Metabolic abnormalities and new-onset seizures in human immunodeficiency virus-seropositive patients, Epilepsia. 36 (1995) 146–150.
[41] J.S. Coselli, S.A. LeMaire, L.D. Conklin, C. Koksoy, Z.C. Schmittling, Morbidity and mortality after extent II thoracoabdominal aortic aneurysm repair, Ann. Thorac. Surg. 73 (2002) 1107–1116.
[42] F.W. Lafferty, C.A. Hubay, Primary hyperparathyroidism, Arch. Intern. Med. 149 (1989) 789–796.
[43] A.P. McLaughlin, J.E. Altwein, W.O. Kessler, R.F. Gittes, Hazards of gallamine administration in patients with renal failure, J. Urol. 108 (1972) 515–517.
[44] http://www.mayoclinic.com. Kidney failure, signs and symptoms, 2004.
[45] http://www.niams.nih.gov/hi/topics/arthritis/rahandout.htm, Handout on Health: Rheumatoid arthritis, Medications, NIH Publication No. 04-4179, 2004.
[46] B.A. Rawlins, F.P. Girardi, O. Boachie-Adjei, Rheumatoid arthritis of the cervical spine, Rheum. Dis. Clin. North Am. 24 (1998) 55–65.
[47] N.E. Neff, G. Kuo, Acute manic psychosis induced by triple therapy for H. pylori, J. ABFP 15 (2002) 66–68.
[48] E. Gomez-Gil, F. Garcia, L. Pintor, J.A. Martinez, M. Mensa, J. de Pablo, Clarithromycin-induced acute psychoses in peptic ulcer disease, Eur. J. Clin. Microbiol. Infect. Dis. 18 (1999) 70–71.
[49] L.J. Cass, L. Alexander, M. Enders, Complications of corticotropin therapy in multiple sclerosis, J. Am. Med. Assoc. 197 (1966) 105–110.
[50] United States Renal Data System, R.N. Foley, C.A. Herzog, A.J. Collins, Blood pressure and long-term mortality in the United States hemodialysis patients: USRDS Waves 3 and 4 study, Kidney Int. 62 (2002) 1784–1790.
[51] M.F. Lucas, C. Quereda, J.L. Teruel, L. Orte, R. Marcen, J. Ortuno, Effect of hypertension before beginning dialysis on survival of hemodialysis patients, Am. J. Kidney Dis. 41 (2003) 814–821.
[52] F. Imbert-Bismut, V. Ratziu, L. Pieroni, F. Charlotte, Y. Benhamou, T. Poynard, Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study, The Lancet 357 (2001) 1069–1075.
[53] N. Callewaert, H. Van Vlierberghe, A. Van Hecke, W. Laroy, J. Delanghe, R. Contrerras, Noninvasive diagnosis of liver cirrhosis using DNA sequencer-based total serum protein glycomics, Nat. Med. 10 (2004) 429–434.
[54] H. Jarbin, A.-L. Von Knorring, Suicide and suicide attempts in adolescent-onset psychotic disorders, Nord. J. Psychiatry. 58 (2004) 115–123.
[55] C.-H. Kim, K. Jayathilake, H.Y. Meltzer, Hopelessness neurocognitive function and insight in schizophrenia: relationship to suicidal behavior, Schiz. Res. 60 (2003) 71–80.
[56] N.I. Cherny, The management of cancer pain, CA: a Cancer J Clin. 50 (2000) 70–116.
[57] www.astrazeneca.com.
[58] M. Fukuchi, K. Kumagai, M. Sakuma, Y. Kagaya, J. Watanabe, K. Tabayashi, K. Shirato, Warfarin-intractable, intraatrial thrombogenesis in a 52-year old woman with mitral stensosis and chronic atrial fibrillation, Tohoku J. Exp. Med. 203 (2004) 59–63.
[59] V. Pengo, F. Babero, A. Biasiolo, C. Pegoraro, F. Noventa, S. Iliceno, Prevention of thromboembolism in patients with mitral stenosis and associated atrial fibrillation: effectiveness of low intensity (INR target 2) oral anticoagulant treatment, Thromb Haemost. 29 (2003) 760–764.
[60] M. Vaturi, Y. Shapira, H. Vaknin-Assa, A. Oron, R. Matesko, A. Sagie, Echocardiographic markers of severe tricuspid regurgitation associated with right-sided congestive heart failure, J. Heart Valve Dis. 12 (2003) 197–201.
[61] T.C. Martin, Echocardiographic findings in a comtemporary Afro-Caribbean population referred for evaluation of atrial fibrillation or flutter, West Ind. Med. J. 50 (2001) 294–296.
[62] W.B. Kannel, R.B. D'Agostino, H. Silbershatz, A.J. Belanger, P.W.F. Wilson, D. Levy, Profile for estimating risk of heart failure, Arch. Intern. Med. 159 (1999) 1197–1204.
[63] J. He, L.G. Ogden, L.A. Bazzano, S. Vupputuri, C. Loria, P.K. Whelton, Risk factors for congestive heart failure in US men and women, Arch. Intern. Med. 161 (2001) 996–1002.
[64] M.A. Khatib, O. Rahim, R. Kania, P. Molloy, Iron deficiency anemia induced by long-term ingestion of Omeprazole, Dig. Dis. Sci. 47 (2002) 2596–2597.
[65] S.J. Lisle, R.M. Lewis, C.J. Petry, S.E. Ozanne, C.N. Hales, A.J. Forhead, Effect of maternal iron restriction during pregnancy on renal morphology in the adult rat offspring, Brit. J. Nutr. 90 (2003) 33–39.
[66] http://lucid.cpmc.columbia.edu/medlee/demo/.
[67] E.M. Van Mulligen, C. Van Der Eijk, J.A. Kors, B.J. Schijvenaars, B. Mons, Research for research: tools for knowledge discovery and visualization, Proc. AMIA Symp. 23 (2002) 835–839.

Chid Apte, Ph.D. is the manager of the Data Analytics Research group within the Mathematical Sciences Department of IBM's Research Division, and the Research Relationship Manager for Business Intelligence Solutions. He has over 20 years of experience in conducting and leading research and advanced development in the areas of data mining based business intelligence and knowledge-based systems. Dr. Apte has worked in diverse areas of applications, including manufacturing quality control, portfolio management, insurance and financial risk management, targeted marketing, automated help desks, lifetime value modeling, clinical and healthcare data mining, and market intelligence. He is a senior member of the IEEE, a member of the AAAI and ACM SIGKDD, has published extensively in his areas of expertise, and is actively involved in organizational aspects of leading data mining conferences. He received his Ph.D. in Computer Science from Rutgers University and B. Tech. in Electrical Engineering from the Indian Institute of Technology, Bombay. His current research interests are focused on leveraging machine learning and computational statistics for analytics applications to business and science.

Simona Cohen, M.Sc. has been a research staff member in IBM Haifa Labs since 1993. She holds a M.Sc. in Computer Science (1989) and a B.Sc. in Computer Science (1986) both from the Technion, Israel Institute of Technology. Prior to joining IBM, she was a research assistant in the Technion and worked in LanOptics in Israel and in Graphnet in New Jersey, USA. Her interest areas include information integration and knowledge management systems especially in the biomedical domain. Mrs. Cohen is the Haifa project leader of the IBM Clinical Genomics solution, which enables research institutions and biopharmaceutical companies across the world to integrate, store, analyze and better understand genotypic and phenotypic data for medical research and patient care.

Carleton T. Garrett M.D., Ph.D. is professor of Pathology and Director of the Division of Molecular Diagnostics in the Department of Pathology of Virginia Commonwealth University. He is also medical director of the CLIA'88 certified molecular diagnostics laboratory in the Molecular Diagnostics Division. Dr. Garrett received his MD from The Johns Hopkins School of Medicine and his Ph.D. in Oncology from the University of Wisconsin. He performed his residency training in anatomic pathology at The Johns Hopkins Hospital and the University of Wisconsin General Hospitals in Madison and is board certified in anatomical pathology. In addition to his clinical responsibilities, Dr. Garrett manages a human cancer specimen acquisition service at VCU for cancer researchers and performs cancer research using gene expression microarrays. Previously, he was principle investigator of a project "Acquisition of Human Cancer Residual Tissue Samples and Microarray Gene Expression Analysis" which was part of a multi institutional three million dollar grant funded by the Virginia Commonwealth Technology Research Fund entitled "*Cancer Genomics and Development of Diagnostic Tools and Therapies*". He also served as the Program Director for the latter grant.

William A. Knaus, M.D. is the Evelyn Troup Hobson Professor and Chair of the Department of Public Health Sciences at the University of Virginia Health System. Dr. Knaus received his medical degree from West Virginia University School of Medicine in 1972 and served as the Director of the ICU Research Unit at George Washington University from 1978–1995. There, he created a clinical research unit focused on developing a severity of illness and prognostic scoring system for critically ill hospitalized patients, APACHE (Acute Physiology, Age, Chronic Health Evaluation). The ICU Research Unit was further supported and expanded with public and private grant funds from an initial database of 500 to over 1,000,000 cases worldwide. Dr. Knaus also designed and successfully managed one of the largest and most well-supported ($30 million) clinical trials of physician decision-making, The SUPPORT (Study to Understand Prognoses, Preferences, and Outcomes from Treatment) Trial. In his capacity as Chair of the Department of Public Health Sciences at the University of Virginia Health System, Dr. Knaus has designed and developed a new clinical department within the School of Medicine. He developed an integrated clinical and administrative data repository (CDR) to support research and management efforts throughout school of medicine and health system. In 2000, Dr. Knaus was elected to The Institute of Medicine National Academy of Sciences. He is currently leading several university-wide bioinformatics integration efforts.

Jason Lyman, M.D., M.S., is currently an Assistant Professor of Clinical Informatics in the Department of Public Health Sciences at the University of Virginia School of Medicine. In addition, he is Clinical Director of the Clinical Data Repository (CDR), an enterprise-wide data warehouse supporting clinical research at UVA. His research interests include clinical decision support, data warehousing, patient safety, and physician order entry. Dr. Lyman has active teaching responsibilities in the undergraduate medical school curriculum as well as in his departmental master's degree program. Dr. Lyman has prior clinical experience in pediatrics and has completed an NLM-funded fellowship and master's degree in Clinical Informatics at Oregon Health Science University.

Greg Miller, Ph.D. is a Professor in the Pathology Department at Virginia Commonwealth University. He serves as Director of Pathology Information Systems and Director of Clinical Chemistry. He received a Ph.D. in Biochemistry from the University of Arizona in 1973; did post-doctoral training in Clinical Chemistry at the Ohio State University; and became a Diplomat of the American Board of Clinical Chemistry in 1976. His current professional activities include Chair of the CLSI Area Committee on Clinical Chemistry and Toxicology, Consultant to the College of American Pathologists Chemistry Resource Committee, chair of the NIH/National Kidney Disease Education Program Laboratory Working Group, and member of the American Diabetes Association Laboratory Working Group for Standardization of Insulin Assays.

Rudy Muller, B.S. Computer Science, is a Computer Systems Engineer with Virginia Commonwealth University. His specializations at VCU include system architecture design, programming, and network management.

Irene M. Mullins, M.S. is an Instructor in the Department of Public Health Sciences at the University of Virginia Health System. She received a B.A. *cum laude* with High Honors in Biology from Mount Holyoke College, in 1997 and a Master's degree in population genetics at Virginia Polytechnic Institute and State University, in 2000. She has since collaborated on several molecular technique-based projects at the University of Virginia Health System. Her current role as a research collaborator for the Department of Public Health Sciences translational research initiative has resulted in three independent experimental projects involving the genetics of immune control of melanoma metastasis and data mining of patient records for hypothesis-generation. She is currently pursuing several clinical research projects and applying to medical school.

Daniel Platt, Ph.D., received a Ph.D. in condensed matter physics from Emory in 1992. He has been worked at the IBM Computational Biology Center since its founding, working in the Bioinformatics and Pattern Discovery group. His current interests have expanded to encompass redescription mining and the derivation of inference rules from mined patterns in application to medical records. He is also interested in and involved with population genetics studies.

Isidore Rigoutsos, Ph.D. is the manager of the Bioinformatics and Pattern Discovery group at the Computational Biology Center of IBM's Thomas J. Watson Research Center in Yorktown Heights, NY where he has been since 1992. Dr. Rigoutsos received his B.S. degree in Physics from the National University of Athens and the Ph.D. degree in Computer Science from New York University's Courant Institute of Mathematical Sciences. Since January of 2000, he has been a Visiting Lecturer at the Department of Chemical Engineering at the Massachusetts Institute of Technology where he teaches a Spring Semester and a Summer Professional course, both in Bioinformatics. Dr. Rigoutsos is a Fulbright Scholar, a senior member of the Institute of Electrical and Electronics Engineers (IEEE), a member of the International Society for Computational Biology (ISCB), the American Society for Microbiology, and the American Association for the Advancement of Science (AAAS). In 2003, Dr. Rigoutsos was elected a Fellow of the American Institute for Medical and Biological Engineering (AIMBE). He is the author/co-author of numerous peer-reviewed publications, and holds 13 U.S. and 2 European patents. He is an Associate Editor for the journal "Genomics," and on the Editorial Board of "Bioinformatics," "Human Genomics," "International Journal of Bioinformatics Research and Applications," and "Gene Therapy and Molecular Biology." He is also a Founding Member of the Hellenic Society for Computational Biology. Additionally, he serves on the Advisory Board of the Master's program in Bioinformatics of Oxford University in the United Kingdom.

Barry Robson, B.Sc.(Hons), Ph.D., D.Sc. (IBM Distinguished Engineer), was the Strategic Advisor at IBM's T. J. Watson Research Center, at Yorktown Heights, NY, where he played a key role in proposals leading to IBM's DiscoveryLink, Blue Gene protein science and Secure Health and Medical Access Network (S.H.A.M.A.N.) projects. He is active in regard to studies in innovation and technical vitality at corporate and national level; he served on the Innovation Frontiers and the National Innovation Initiative and contributed to the important report "Innovate America. National Innovation Initiative Report" (Council on Competitiveness, December 2004). He is also the Program Director Computational Medicine, and a Council Member of the Deep Computing Institute. He was recently Professional Interest Communities Chair in computational biology and medicine and will continue to participate through the contemporary Chair. His scientific and medical expertise and interests are in regard to biomolecular medicine, healthcare and the digital patient record with pharmacogenomic and other data, information technology support of bio-ethics, and high dimensional clinical data mining for diagnosis, prognosis, and research.

Kenneth W. Scully, M.S. received his B.S in physics in 1971 from Wheaton College and a M.S. in Computer Science from the University of Colorado Boulder in 1983. Since 1996, he has been the Database Administrator and Technical Lead for the

Clinical Data Repository (CDR) project, an integrated data warehouse containing clinical and financial information from the UVA Health System that is accessible from a Web browser to UVA researchers, clinicians, and staff at the University of Virginia Health System.

Mir S. Siadaty, M.D., M.S. is an Assistant Professor of Clinical Informatics and Biostatistics in the Department of Public Health Sciences at the University of Virginia Health System. He received his M.D. from Tehran University of Medical Sciences in 1988, and his M.S. in biostatistics from the University of Minnesota in 2002. In addition to his formal training in both medicine and statistics, Dr. Siadaty has computer science expertise. He has published on the synthesis of biomedical knowledge by more explicit statistical methods for meta-analysis. Currently, Dr. Siadaty's research is focused on pooling two huge bodies of information, the biomedical knowledge (an instance of which is PubMed of National Library of medicine, with 15 million published papers indexed) and patient data (such as UVa Clinical Data Repository with over one million patients digitized data), with the goal to discover novel regularities, and generate new hypotheses worthy of focused research. The ultimate goal would be to provide a tool that could lead to new basic and applied discoveries that would advance research, clinical care, and improve human health.

Sholom Weiss is a research staff member at the IBM T. J. Watson Labs and a professor (emeritus) of computer science at Rutgers University. He is an author and coauthor of many papers on artificial intelligence and machine learning, including a book entitled "Text Mining: Predictive Methods for Analyzing Unstructured Information" (Springer, 2005). His current research interests emphasize innovative methods of data mining. He is a fellow of the American Association for Artificial Intelligence.