Empirical Bayesian Data Mining for Discovering Patterns in Post-Marketing Drug Safety

David M. Fram
Lincoln Technologies, Inc.
40 Washington Street, Suite 220
Wellesley Hills, MA 02481
781-237-8722
david.fram@lincolntechnologies.com

June S. Almenoff, MD, PhD
GlaxoSmithKline
Five Moore Drive
Research Triangle Park, NC 27709
888-825-5249
june.s.almenoff@gsk.com

William DuMouchel PhD
AT&T Shannon Laboratory
180 Park Avenue, Rm C283
Florham Park, NJ 07932
520-299-8624
dumouchel@research.att.com

ABSTRACT

Because of practical limits in characterizing the safety profiles of therapeutic products prior to marketing, manufacturers and regulatory agencies perform post-marketing surveillance based on the collection of adverse reaction reports ("pharmacovigilance").

The resulting databases, while rich in real-world information, are notoriously difficult to analyze using traditional techniques. Each report may involve multiple medicines, symptoms, and demographic factors, and there is no easily linked information on drug exposure in the reporting population. KDD techniques, such as association finding, are well-matched to the problem, but are difficult for medical staff to apply and interpret.

To deploy KDD effectively for pharmacovigilance, Lincoln Technologies and GlaxoSmithKline collaborated to create a web-based safety data mining web environment. The analytical core is a high-performance implementation of the MGPS (Multi-Item Gamma Poisson Shrinker) algorithm described previously by DuMouchel and Pregibon, with several significant extensions and enhancements. The environment offers an interface for specifying data mining runs, a batch execution facility, tabular and graphical methods for exploring associations, and drilldown to case details. Substantial work was involved in preparing the raw adverse event data for mining, including harmonization of drug names and removal of duplicate reports.

The environment can be used to explore both drug-event and multi-way associations (interactions, syndromes). It has been used to study age/gender effects, to predict the safety profiles of proposed combination drugs, and to separate contributions of individual drugs to safety problems in polytherapy situations.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining, scientific databases.*

Keywords

Data mining, empirical Bayes methods, association rules, post-marketing surveillance, pharmacovigilance.

1. INTRODUCTION

It is widely recognized that there are practical limits on the degree to which safety profiles of therapeutic products (drugs, vaccines, medical devices) can be fully characterized before these products are approved for marketing: pre-marketing studies are inherently too short, with study populations that are too small and too homogeneous, to be able to detect important but relatively rare adverse events. The opportunity for a new drug's true side effect profile to reveal itself is often realized after the drug is approved and then used in conjunction with other therapies. To provide an objective basis for monitoring and assessing the safety of marketed products, pharmaceutical companies and regulatory agencies have implemented post-marketing surveillance activities ("pharmacovigilance") based in large measure on the collection of spontaneously generated adverse reaction reports. Report initiation (by health professionals and consumers) is generally voluntary; by contrast, the pharmaceutical companies are generally under legal obligation to follow up on reports that they receive and to pass them along to various regulatory authorities.

As a result of these pharmacovigilance efforts, a number of large databases of spontaneous adverse event reports have come into existence: Each major pharmaceutical company has a proprietary database of reports focused on cases in which one of the company's products was considered to be "suspect" (~500,000 reports for the larger pharmaceutical companies). In addition, there are several combined databases maintained by government regulatory agencies and health authorities that are available in varying degrees for public use; some of these combined databases contain as many as ~3,000,000 reports. Although the various private and public databases differ in detail, the core content is fairly consistent: For each adverse event case, there is a demographic record (age, gender, date of event, seriousness of event), one or more therapeutic product records (generic or trade name, suspect or concomitant designation, route of administration, dosage), and one or more records documenting a sign, symptom or diagnosis (typically represented as a coded "event term" based on a standardized medical coding dictionary). Individual databases may also contain narratives (from which event terms were coded), outcomes (e.g., hospitalization, death), and report source (consumer or health professional, domestic or foreign).

These databases of spontaneous reports represent the largest existing sources of information relating specifically to the side-effect profiles of marketed therapeutic products. Systematic analysis of this data has proved difficult both for conceptual and practical reasons: concerns about the effects of underreporting, the lack of easily linkable measures of exposure (the "denominator problem"), inconsistencies and evolution over time in naming/coding practices, and technical issues relating to database access and computational tractability. In the absence of systematic analysis methods, the emphasis historically has been on largely non-analysis-based "monitoring" through such techniques as case-by-case examination of newly reported cases, tabulations of counts of events for specific drugs, and detailed review of all the data fields (including, specifically, the free-text medical narratives) of cases associated with a possible safety concern. These traditional approaches tend to be highly dependent on the knowledgeability and alertness of individual safety reviewers. They also suffer from an absence of contextual information: in isolation, it can be very difficult to tell whether 10 cases of a specific drug-event combination is disproportionately frequent such that is "interesting" and merits further investigation.

There has been a growing interest, originating at government health authorities, in the potential use of statistical data mining ("association finding", "disproportionality analysis") as a means of extracting knowledge from pharmacovigilance databases. Such "safety data mining" holds the promise of contributing to a more systematic approach to the monitoring of drug safety and to the earlier detection of potential problem areas. In 1997, Fram and DuMouchel began a long-standing research partnership with Dr. Ana Szarfman, Medical Officer at the Food and Drug Administration Center for Drug Evaluation and Research, to experiment with the application of DuMouchel's Gamma Poisson Shrinker ("GPS") and Multi-Item Gamma Poisson Shrinker ("MGPS") techniques to FDA's combined SRS and AERS databases and to validate results retrospectively against known drug-induced adverse reactions [1,2,3,4]. Independent, parallel efforts have proceeded at the UK Medicines Control Agency (MCA) where Dr. Stephen Evans has explored the use of proportional reporting ratios (PRR's)[5] and at the Uppsala Monitoring Centre where Edwards, Lindquist, Bate and others have pursued the use of techniques based on Bayesian neural networks[6].

GlaxoSmithKline ("GSK") and Lincoln Technologies ("Lincoln") began a collaboration in 2001 to apply safety data mining techniques both to provide direct leverage in addressing GSK's core business problems of pharmacovigilance and risk management and, more broadly, to create decision support tools to assist with a variety of complex safety-related business issues. This work began in a "service bureau" mode, where Lincoln staff performed data mining runs using the MGPS software and the publicly available FDA adverse event data ("FOI AERS") and delivered analysis results in tabular and graphical form to GSK pharmacovigilance staff. At this stage of the collaboration, data mining projects required manual integration of results from several different software tools utilized for the distinct steps of data extraction/transformation, data mining, and output visualization. Positive early scientific results demonstrated the practicality and utility of the overall approach, and also suggested the desirability of providing GSK pharmacovigilance staff with a means for direct access to the technology to support even wider application of the approach across a number of GSK's safety surveillance sites.

The focus of this paper is to report on the joint effort by Lincoln and GSK to define, design, implement, test, and deploy a web-based visual data mining environment ("WebVDME") that packages sophisticated KDD techniques for pharmacovigilance in a format that is accessible to the intended medical end-user community.

2. REQUIREMENTS & ARCHITECTURE

The WebVDME project was undertaken on the understanding that the initial focus would be on creating a custom solution to meet GSK's specific needs, but that the software would eventually become a commercial product generalized and enhanced for deployment at other pharmaceutical industry and government clients. The application would start out being hosted at Lincoln, but could later be moved to reside on a system or systems at GSK. Lincoln and GSK also realized that the project would need to follow a documented System Development Life Cycle (SDLC) development process appropriate to a major application in a regulated industry and would also benefit from close interaction between the developers and end users to try out and refine system features.

The formal process began with a list of GSK's primary business requirements for the system, which included:

- Implementation as a "thin client" web-based application (no software, controls, applets, etc., to be installed on client computer), compatible with access through GSK's existing firewalls and compliant with GSK's major IT standards (MS Windows 2000 servers, Oracle database).
- Data mining based on MGPS, including identification of signals related to two-way (drug-event) and multi-way (drug interaction, multi-event syndrome) associations, with end-user control over choices regarding stratification and subsetting.
- Access to the major public U.S. drug and vaccine databases, with ability to select dictionary level and combining strategy for drug and event categories (e.g., use of "brand" versus "generic" names for drugs, lumping of closely synonymous adverse event symptom terms) .
- A user interface suitable for direct use by medical staff.
- Output of data mining results in graphical and tabular form by means of the web interface, including screening and subsetting of results, and also the ability to download results for use with Excel and with other third-party graphics and statistical packages.

Based on these results, a technical architecture for the application was designed as shown below in Figure 1, below.

The various computer system components depicted in the architecture perform the following roles:

- *Client computers* (standard PC workstations) – support the operation of Internet Explorer to provide browser access to the data mining environment. Also support optional software packages (e.g., Microsoft Excel) for use with data downloaded from the data mining environment.

- *Application server* (Intel/Windows 2000 Server/Advanced Server or Sun Sparc/Solaris 2.8) – runs web server and Java Server Page (JSP) server to support the WebVDME application, which is implemented as JSP pages and supporting Java classes.

- *Oracle server(s)* (Oracle 8i or 9i) – runs the Oracle database that contains the safety database(s), data mining results, and administrative/control information for WebVDME.

- *Data mining server* (Intel/Windows 2000 Server/Advanced Server or Sun Sparc/Solaris 2.8) – runs MGPS data mining algorithm plus Java support classes.

The present configuration hosted by Lincoln for GSK includes a dedicated 2-processor Windows server (Dell PowerEdge 2650), with 4 GB of main memory and 100 GB of disk space.

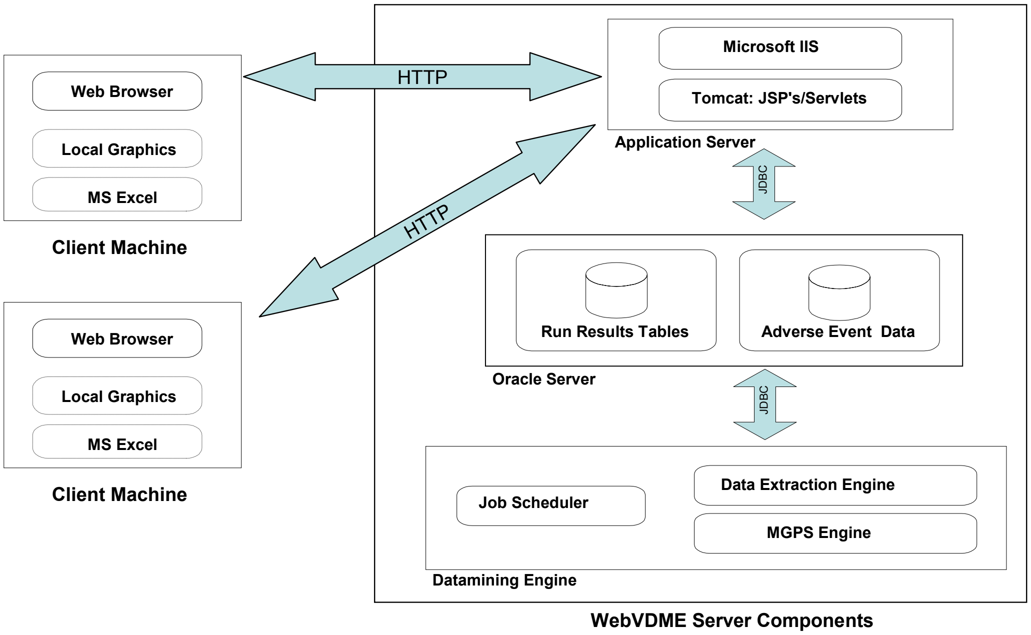


Figure 1. WebVDME Architecture

3. DATA SOURCES & PREPARATION

As described above, there exist a variety of public and company-specific pharmacovigilance databases which share a common conceptual organization (case reports that contain information on demographics, drugs, and events) but differ in details of field and table naming, presence or absence of specific attributes, and use of specific medical coding dictionaries. To insulate both the application and the end users from superficial variation in the detailed formatting of the target safety database, we decided on a design that supports multiple database "configurations" each of which defines a mapping between user-visible "variables" and specific database tables and columns, including specification of plausible roles for the variables (e.g., a drug name, an event name, an attribute suitable for use in stratification or subsetting, etc.) These configurations can be set up, using a set of configuration management WebVDME web pages, by expert users who are familiar both with the target database schema and the pharmacovigilance application; medical end-users work with the application-specific variables so defined (end-users do not need to be aware of the target database schema). Multiple configurations can be used to support user access to several different databases (e.g., to do a data mining run first on the public data and then on in-house data), and also to several different versions of the same database (e.g., different chronological snapshots, or versions that include or exclude so-called "concomitant" medications).

The specific database that was the initial focus of the GSK collaboration is the public-release version of FDA's AERS database (FOI AERS), which contains approximately 3,000,000 adverse event reports collected from 1968 to the present (available as series of quarterly updates from NTIS). Substantial pre-processing of this data is required to make it appropriate for MGPS safety signaling (this pre-processing must be carried out with each new quarterly release of the public data).

There are several distinct reasons for pre-processing. First, because the data was collected over more than three decades, using several different database organizations, mapping and recoding is required to present a uniform view of the data.

Second, drug names are collected as free text, with substantial variation among submitting organizations and individuals in how trade, generic, or mixed names are entered; whether packaging, route-of-administration and dosing information is provided; and in punctuation and spelling; etc. Also, the desired granularity of drug information for use in data mining can vary depending on the intended analysis goal (while it is often most useful to consider drugs based on the same molecular entity as equivalent, trade names may be useful in distinguishing between drugs produced by different manufacturers or between uses of a common substance for different disease indications). Lincoln uses an extensive set of parsing and lookup tables to make consistent generic and trade names available.

Third, there is a need to remove duplicates. Adverse event databases typically contain multiple versions of the same case because of regulatory requirements to submit an initial expedited report plus a series of follow-up reports as addition information about a case because available. For purposes of statistical data mining these multiple versions need to be collapsed to a single "best representative version" of the case. Beyond this versioning problem, there are several other significant sources of case duplication: multiple reports of the same medical event by different manufacturers and reports arriving through different pathways. In the context of an association-finding technique, report duplication represents a common source of false positives (especially for otherwise rare higher-order associations). Lincoln has implemented sophisticated "fuzzy-equality" case-matching algorithms to identify likely duplicate reports.

Based on the success of the initial project accessing FOI AERS, GSK has decided to extend the use of the data mining system to operate on its internal adverse event ("OCEANS") database. This extension to working with the OCEANS data does not require software modification because of WebVDME's configuration layer, described above, that takes care of the mapping between user-visible data mining variable names and the specific logical and physical structures used in the pharmacovigilance database.

4. MGPS & EXTENSIONS

The analytical core of the WebVDME application is a high-performance implementation of the MGPS (Multi-Item Gamma Poisson Shrinker) algorithm described previously by DuMouchel and Pregibon [2]. MGPS is based on the metaphor of the "market basket problem", in which a database of "transactions" (adverse event reports) is mined for the occurrence of interesting (unexpectedly frequent) itemsets (e.g., simple drug-event pairings or more complex combinations of drugs and events representing interactions and/or syndromes). Interestingness is related to the factor by which the observed frequency of an itemset differs from a nominal baseline frequency. The baseline frequency is usually taken to be the frequency that would be expected under the full independence model, in which the likelihood of a given item showing up in a report is independent of what other items appear in the report. (Other choices for the baseline frequency are possible; see discussion of "comparative analysis" below.)

For each itemset in the database, a relative reporting ratio RR is defined as the observed count N for that itemset divided by the expected count E. When using the independence model as the basis for computing the expected count, MGPS allows for the possibility that the database may contain heterogeneous strata with significantly different item frequency distributions occurring in the various strata. To avoid concluding that an itemset is unusually frequent just because the items involved individually all tend to occur more frequently in a particular stratum ("Simpson's paradox"), MGPS uses the Mantel-Haenszel approach of computing strata-specific expected counts and then summing over the strata to obtain a database-wide value for the expected count.

To improve upon the estimation of "true value" for each RR (especially for small counts), the empirical Bayesian approach of MGPS assumes that the many observed values of RR are related in that they can be treated as having arisen from a common "super population" of unknown, true RR-values. The method assumes that the set of unknown RR is distributed according to a mixture of two parameterized Gamma Poisson density functions, and the parameters are estimated from a maximum likelihood fit to the data. This process provides a "prior distribution" for all the RR's, and then the Bayes rule can be used to compute a posterior distribution for each RR. Since this method improves over the simple use of each N/E as the estimate of the corresponding RR, it can be said that the values of N/E borrow strength from each other to improve the reliability of every estimate.

The improved estimates of RR—referred to as EBGM (Empirical Bayes Geometric Mean) values—are actually derived from the expectation value of the logarithm of RR under the posterior probability distributions for each true RR. EBGM is defined as EBGM = exponential of expectation value of log(RR). EBGM has the property that it is nearly identical to N/E when the counts are moderately large, but is "shrunk" towards the average value of N/E (typically ~1.0) when N/E is unreliable because of stability issues with small counts. The posterior probability distribution also supports the calculation of lower and upper 95% confidence limits (EB05, EB95) for the relative reporting ratio. A technical summary of MGPS is included at the end of this paper.

Two new extensions to the MGPS data mining algorithm were developed in response to pharmacovigilance data mining needs:

1. *Ability to shrink toward the all-2-factor model when looking at higher-order effects*. In the study of multi-item associations in pharmacovigilance (drug interactions, syndromes), it is important to be able to distinguish effects that result from the synergistic combination of multiple items from effects that are only the consequence of the pairwise associations. An extension to MGPS supports computing the EBGM scores for higher-order effects based directly on expected values that can be estimated from already-computed two-factor relationships. With this enhancement, EBGM's for higher-order combinations are significantly high only when the observed count differs significantly from what would be expected from the component two-factor relationships.

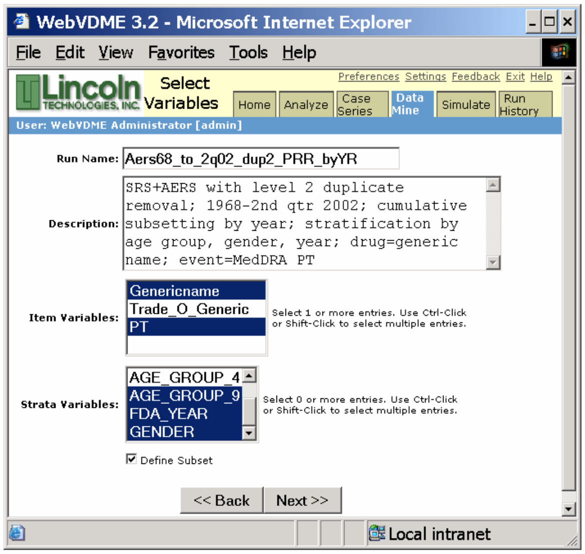
2. *Highlighting period-to-period change*. MGPS can make use of baseline values ("expected counts") that have been derived in some fashion other than from the standard full independence model, and is also able to perform iterative data mining runs on subsets that contain all database records belonging to a series of time windows. In this extension, MGPS generates the expected counts for the second and subsequent iterations from the EBGM estimates computed up through the previous iteration. The resulting signal scores are a measure of change from a prior time period, which is useful for temporal trend analysis (e.g., detecting an altered safety profile due to a change in pharmaceutical manufacturing or in physician prescribing patterns).

The WebVDME project benefited substantially from the availability of an initial C++ implementation of MGPS developed over a period of years in collaboration with DuMouchel, and also from a recent NIH-sponsored activity to develop a high-performance implementation of the method. The association-counting phase of the algorithm (which accounts for much of the time and space required for execution) uses a modification of the "a priori" method to prune counting of higher-order associations when the counts for their component lower-order associations imply that the higher-order count cannot meet a minimum count threshold [7]. Counting is implemented using a hash table to store the count data, with an ability to write out and merge partial results when working with very large problems that would exceed available physical memory or address space. In order to speed up the maximum likelihood estimation of the parameters of the prior distribution in the empirical Bayes model, a summary of the table of counts and baseline values, computed using the data squashing methods of DuMouchel et al [8] is substituted for the much larger baseline values file.

5. USER INTERFACE & CAPABILITIES

The WebVDME user interface is organized around a set of "tabs" and links that lead to major components of the system. The principal tabs are: Data Mining (specifying and initiating data mining runs), Analyze (exploring data mining results), and Case Series (reviewing the details of specific cases identified through data mining). There are also a set of administrative functions related to creating new users and to granting privileges for using the different components of the system in a production environment.

MGPS data mining runs are defined through a "wizard" (a multi-step series of web page dialogs) that guides the user in selecting the variables to be used as the source of data mining items, in setting up stratification and subsetting, and in making various other technical choices such as the maximum number of items in an itemset and the minimum count required to consider an itemset. The first page in the wizard is shown in Figure 2 below. When a run definition is completed, it can be submitted for execution (either immediately or at a scheduled time). Execution proceeds in the background; the user can continue working to define or analyze other runs, or can log out. Email notification of the completion of submitted runs can be requested.


WebVDME 3.2 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Lincoln TECHNOLOGIES, INC.
Select Variables
Preferences Settings Feedback Exit Help
Home
Analyze
Case Series
Data Mine
Simulate
Run History
User: WebVDME Administrator [admin]
<< Back
Next >>
Local intranet

Run Name: Aers68_to_2q02_dup2_PRR_byYR

Description: SRS+AERS with level 2 duplicate removal; 1968-2nd qtr 2002; cumulative subsetting by year; stratification by age group, gender, year; drug=generic name; event=MedDRA PT

Item Variables: Genericname, Trade_O_Generic, PT — Select 1 or more entries. Use Ctrl-Click or Shift-Click to select multiple entries.

Strata Variables: AGE_GROUP_4, AGE_GROUP_9, FDA_YEAR, GENDER — Select 0 or more entries. Use Ctrl-Click or Shift-Click to select multiple entries.

☑ Define Subset

Figure 2. Selecting variables for use in data mining.

The results of a run are stored in an Oracle-based output table that can be accessed by clicking the Analyze tab or the Analyze link in the Run History table. The Run Results table can be quite large (e.g., 500,000 rows), and a set of filtering capabilities is provided to help focus on associations in the table of particular interest. The Filter dialog is shown in Figure 3.

Through filtering, the user can select a set of specific drugs and events to show in the table (using either the primary terms for drugs and adverse events or, if supported by the specific safety database, by higher-level term categories in a dictionary). It is also possible to focus on a specific dimension (e.g., 2-way associations in a run that involves both 2-way and 3-way associations), or on a specific pattern of item types (e.g., only associations involving two drugs and one event). Additional selection criteria can be provided through a SQL WHERE clause (e.g., specifying EB05 > 2 to focus on associations that reliably occur at least twice as frequently as would be expected from independence).

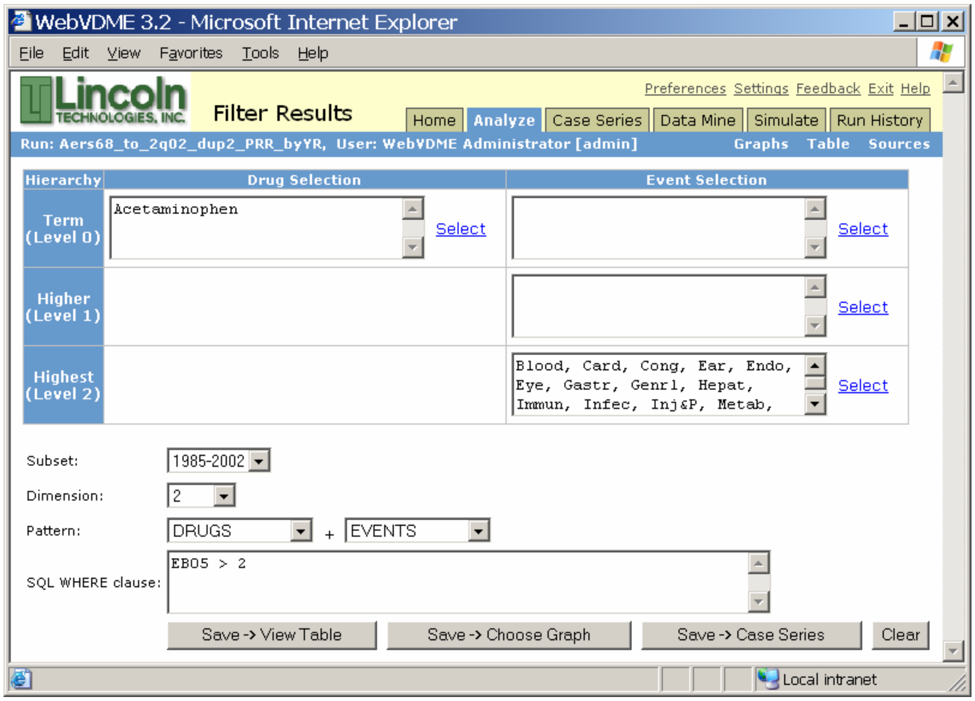
Filtered results can be displayed as a table as shown in Figure 4. By default, the columns shown include the items in the association, the observed (N) and expected (E) counts, RR, EBGM, and the 5th and 95th percentile confidence limits (EB05, EB95). Additional results columns are optionally available. Filtered tabular results can also be downloaded as a spreadsheet.

By clicking on the leftmost column (with the magnifying glass), the user can drill down to a list of the specific adverse event reports behind the association. This list can be used to drill down, further, to full details for individual case reports, or can be saved into a "case series" that can be used to facilitate evaluation of a potential safety signal via a careful case-by-case review.

WebVDME also supports a set of application-specific graphs. Figure 5 shows one graph type, used to track the emergence of safety signals over time when performing a MGPS run involving the iterative analysis of cumulative subsets.

Graphs are implemented as GIF output; display of the graph takes place wholly on the server and can be viewed on any browser. This simple approach to graphics simplifies deployment relative to schemes involving downloading controls or applets to the client, which can run into client computer configuration and firewall issues. Some interactivity is provided through use of "mouseover" to show the statistics behind elements of the graph, and by providing drilldown from graphical elements to the supporting case reports. In cases where more sophisticated graphics are required, WebVDME provides for download of data mining output to graphics packages popular in the pharmaceutical industry.

Additional system capabilities include site administration, including control over how batch jobs are assigned to processors, and a variety of information displays ("audit trails" documenting run and analysis choices that lie behind a display, descriptions of runs and case series, on-line "help" screens describing system use). WebVDME has also been integrated with a data simulation capability (developed with support from CDC), to generate artificial background databases and signals for use in evaluating MGPS signal detection operating characteristics through Monte Carlo techniques.


WebVDME 3.2 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Lincoln TECHNOLOGIES, INC.
Preferences Settings Feedback Exit Help
Home Analyze Case Series Data Mine Simulate Run History
Graphs Table Sources
Select
Select
Select
Select
Save -> View Table
Save -> Choose Graph
Save -> Case Series
Clear
Local intranet

Filter Results

Run: Aers68_to_2q02_dup2_PRR_byYR, User: WebVDME Administrator [admin]

Hierarchy	Drug Selection	Event Selection
Term (Level 0)	Acetaminophen	
Higher (Level 1)		
Highest (Level 2)		Blood, Card, Cong, Ear, Endo, Eye, Gastr, Genrl, Hepat, Immun, Infec, Inj&P, Metab,

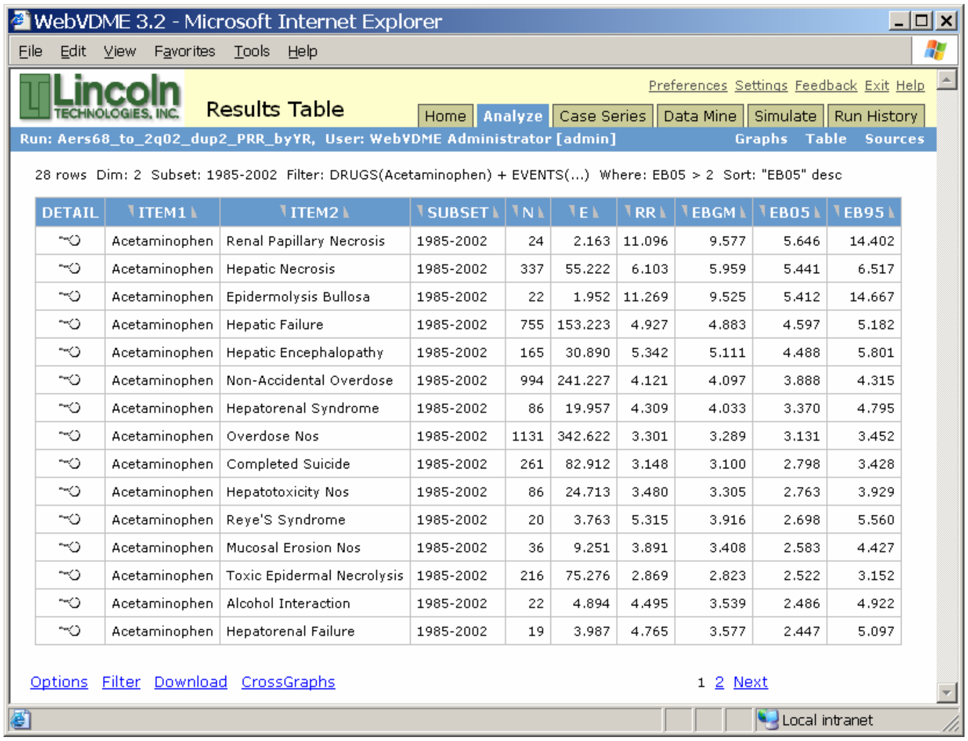
Subset: 1985-2002

Dimension: 2

Pattern: DRUGS + EVENTS

SQL WHERE clause: EB05 > 2

Figure 3. Filter results dialog. Attention is focused on drug-event associations involving the drug Acetaminophen and any event term belonging to a list of specified system organ classes (Blood, Card, Cong, etc.). The example also includes further restrictions on the temporal subset (1985-2002) and the association signal score (EB05 > 2).


WebVDME 3.2 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Lincoln TECHNOLOGIES, INC.
Preferences Settings Feedback Exit Help
Home Analyze Case Series Data Mine Simulate Run History
Graphs Table Sources
Options Filter Download CrossGraphs
1 2 Next
Local intranet

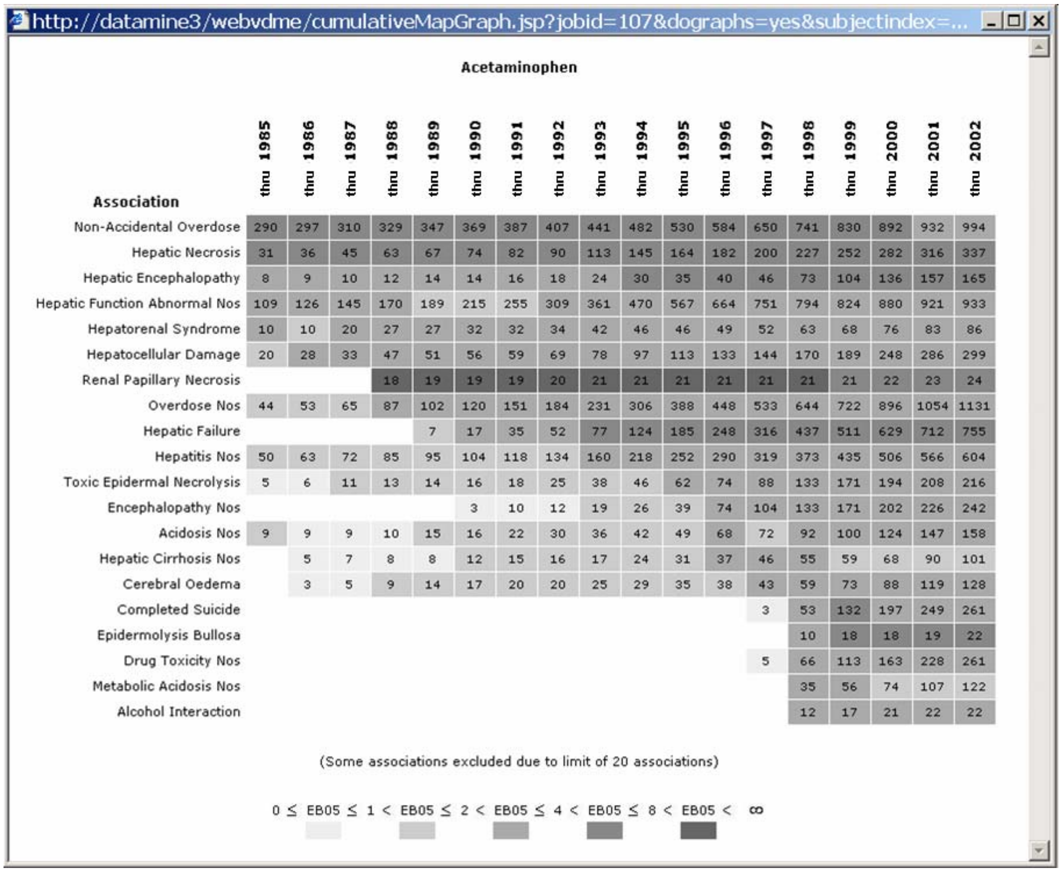
Results Table

Run: Aers68_to_2q02_dup2_PRR_byYR, User: WebVDME Administrator [admin]

28 rows Dim: 2 Subset: 1985-2002 Filter: DRUGS(Acetaminophen) + EVENTS(...) Where: EB05 > 2 Sort: "EB05" desc

DETAIL	ITEM1	ITEM2	SUBSET	N	E	RR	EBGM	EB05	EB95
⊸	Acetaminophen	Renal Papillary Necrosis	1985-2002	24	2.163	11.096	9.577	5.646	14.402
⊸	Acetaminophen	Hepatic Necrosis	1985-2002	337	55.222	6.103	5.959	5.441	6.517
⊸	Acetaminophen	Epidermolysis Bullosa	1985-2002	22	1.952	11.269	9.525	5.412	14.667
⊸	Acetaminophen	Hepatic Failure	1985-2002	755	153.223	4.927	4.883	4.597	5.182
⊸	Acetaminophen	Hepatic Encephalopathy	1985-2002	165	30.890	5.342	5.111	4.488	5.801
⊸	Acetaminophen	Non-Accidental Overdose	1985-2002	994	241.227	4.121	4.097	3.888	4.315
⊸	Acetaminophen	Hepatorenal Syndrome	1985-2002	86	19.957	4.309	4.033	3.370	4.795
⊸	Acetaminophen	Overdose Nos	1985-2002	1131	342.622	3.301	3.289	3.131	3.452
⊸	Acetaminophen	Completed Suicide	1985-2002	261	82.912	3.148	3.100	2.798	3.428
⊸	Acetaminophen	Hepatotoxicity Nos	1985-2002	86	24.713	3.480	3.305	2.763	3.929
⊸	Acetaminophen	Reye'S Syndrome	1985-2002	20	3.763	5.315	3.916	2.698	5.560
⊸	Acetaminophen	Mucosal Erosion Nos	1985-2002	36	9.251	3.891	3.408	2.583	4.427
⊸	Acetaminophen	Toxic Epidermal Necrolysis	1985-2002	216	75.276	2.869	2.823	2.522	3.152
⊸	Acetaminophen	Alcohol Interaction	1985-2002	22	4.894	4.495	3.539	2.486	4.922
⊸	Acetaminophen	Hepatorenal Failure	1985-2002	19	3.987	4.765	3.577	2.447	5.097

Figure 4. Tabular display of data mining results corresponding to filter in Figure 3. A total of 28 drug-event associations match the filter criteria. Downward- and upward-pointing triangles next to each column heading provide control over sort order; in this case the results are presented in order of descending EB05.



http://datamine3/webvdme/cumulativeMapGraph.jsp?jobid=107&dographs=yes&subjectindex=...

Acetaminophen

Association	thru 1985	thru 1986	thru 1987	thru 1988	thru 1989	thru 1990	thru 1991	thru 1992	thru 1993	thru 1994	thru 1995	thru 1996	thru 1997	thru 1998	thru 1999	thru 2000	thru 2001	thru 2002
Non-Accidental Overdose	290	297	310	329	347	369	387	407	441	482	530	584	650	741	830	892	932	994
Hepatic Necrosis	31	36	45	63	67	74	82	90	113	145	164	182	200	227	252	282	316	337
Hepatic Encephalopathy	8	9	10	12	14	14	16	18	24	30	35	40	46	73	104	136	157	165
Hepatic Function Abnormal Nos	109	126	145	170	189	215	255	309	361	470	567	664	751	794	824	880	921	933
Hepatorenal Syndrome	10	10	20	27	27	32	32	34	42	46	46	49	52	63	68	76	83	86
Hepatocellular Damage	20	28	33	47	51	56	59	69	78	97	113	133	144	170	189	248	286	299
Renal Papillary Necrosis				18	19	19	19	20	21	21	21	21	21	21	21	22	23	24
Overdose Nos	44	53	65	87	102	120	151	184	231	306	388	448	533	644	722	896	1054	1131
Hepatic Failure					7	17	35	52	77	124	185	248	316	437	511	629	712	755
Hepatitis Nos	50	63	72	85	95	104	118	134	160	218	252	290	319	373	435	506	566	604
Toxic Epidermal Necrolysis	5	6	11	13	14	16	18	25	38	46	62	74	88	133	171	194	208	216
Encephalopathy Nos						3	10	12	19	26	39	74	104	133	171	202	226	242
Acidosis Nos	9	9	9	10	15	16	22	30	36	42	49	68	72	92	100	124	147	158
Hepatic Cirrhosis Nos		5	7	8	8	12	15	16	17	24	31	37	46	55	59	68	90	101
Cerebral Oedema		3	5	9	14	17	20	20	25	29	35	38	43	59	73	88	119	128
Completed Suicide													3	53	132	197	249	261
Epidermolysis Bullosa														10	18	18	19	22
Drug Toxicity Nos													5	66	113	163	228	261
Metabolic Acidosis Nos														35	56	74	107	122
Alcohol Interaction														12	17	21	22	22

(Some associations excluded due to limit of 20 associations)

0 ≤ EB05 ≤ 1 < EB05 ≤ 2 < EB05 ≤ 4 < EB05 ≤ 8 < EB05 < ∞

Figure 5. Specialized graph for displaying evolution of signal scores over time (for Acetaminophen). Columns represent cumulative yearly results. Rows represent different adverse event terms, ordered from top to bottom according to the first occurrence chronologically of EB05 > 2. Gray-scale indicates signal-score strength; number in each cell gives count of reports mentioning the drug and the event as of the corresponding point in time.

6. DEVELOPMENT PROCESS

Over the 8-month WebVDME development period, the principal Lincoln team members included a project manager/application specialist, 3 senior software developers, and a senior technical writer/software quality specialist. GSK project participants included pharmacovigilance managers and staff and a software quality team. The month-by-month sequence of development activities in the project was as described in Table 1.

Table 1. Development Project History

Month	Activities
pre-Feb 2002	Custom data mining analyses carried out by Lincoln for GSK, early discussion of the web site concept.
Feb 2002	Contract draft completed, including project plan and major requirements
Mar 2002	Formal contract signing, early prototype development, requirements identification teleconferences with GSK project team.
Apr 2002	First live demo of data mining and results viewing (using 2 prototypes focused on controlling data mining runs and on tabular/graphical results display.
May 2002	Separate prototypes integrated, test data acquired and cleaned, initial "help" documentation, requirements and design documentation completed.
Jun 2002	First integrated end-to-end demonstration, user interface tuning in response to GSK feedback, formal test plan and test suite documentation completed.
Jul 2002	Completion of first formal testing cycle, WebVDME test release available to GSK for hands-on testing, identification of key areas for improvement (MGPS memory usage on large runs, ability to filter data mining results using higher-level terms and patterns, graph performance on large datasets)
Aug 2002	Development and testing of improvements, re-execution of test suite, migration to new server, demonstration and signoff on improvements
Sep 2002	User manual completed, end-user and administrator training developed and provided, user acceptance testing performed and documented by GSK, formal release to production.

We attribute the combined team's ability to keep to this ambitious schedule to several factors:

- Ability to make substantial use of prior design and implementation experience by the team members in supporting pharmacovigilance use of data mining with predecessor tools, in developing and deploying several JSP-based applications and in creating, tuning, and validating our high-performance C++ implementation of MGPS.

- Intense and enthusiastic participation by GSK staff throughout the development process, especially in the hands-on testing of interim releases and the production release of the software.

- Conservative choice of operating environment (using a well-established production environment: Oracle, TomCat, Internet Explorer, and solid development tools: JBuilder Enterprise, Visual SourceSafe). Testing requirements were simplified by standardizing on a single SQL database supplier (Oracle) and a single browser environment (IE 5.5 and later).

Important architectural choices in the design of the system were:

- Strict use of server-centric development technologies (HTML, GIF, JSP pages) that facilitate widespread deployment without concern for the details of client computer configurations, network firewalls, corporate security policies, etc.

- Maintenance of all important data resources (source databases, data mining results tables, configurations, etc.) in Oracle to provide for stable storage, security, and rapid results retrieval.

- Integrated batch execution of data mining runs, with end-user capabilities for submitting and monitoring runs from within the system, so compute-intensive data mining runs can take place in the background without interfering with interactive use.

A specific technique used by the project team to support rapid design, implementation, and evaluation while still generating the written design artifacts necessary for validation was to create and maintain a comprehensive set of context-specific "help" screens starting very early in the project. These "help" screens were used simultaneously as design documentation for the user interface, as the primary source for generation of specifications, test scripts and end-user documentation, and as a means of technical communication among the developers, testers, and GSK users.

Software construction was performed by a small group of expert programmers, under the direction of Lincoln's software architect and chief technical officer. Code was constructed using standard development tools (JBuilder) within Lincoln's JSP/Java architecture, following a common coding style. All sources were maintained in SourceSafe (accessed through the web-based SourceOffSite front end) and were checked in frequently. Programmers performed coding and unit testing in a laptop computer environment that was capable of running the application and database. Builds and installs on shared servers took place regularly (once or twice a week early in the cycle, daily later in the cycle). Problems were entered in and tracked using the ProblemTracker software from the time that initial coding and unit testing were completed on a module. Throughout the project, performance problems were the most important cause for rework of software modules, and a continued effort was necessary to achieve reasonably good performance on large-scale data mining problems – these were generally due to technical issues such as database loading or indexing, Java garbage collection, etc., rather than to the data mining algorithm itself.

A major focus of the project was testing. All developers were responsible for performing unit testing of software modules that they develop or modify. Automated regression testing was used primarily to ensure that changes to the MGPS statistical algorithm did not have unintended effects on the computation of signal scores. An extensive suite of manual test scripts was developed to support formal testing requirements. While both the regression testing and the formal manual testing were effective in detecting situations where changes broke previously-working software, we found that aggressive testing of the system on large, realistic problems was still necessary to reach a high level of reliability.

GSK staff made crucial contributions to the testing process through design and execution of formal user acceptance tests (based on application of WebVDME to realistic problems of interest). GSK was also involved in informal early testing of new functionality, so that poor interface choices could be caught and corrected quickly. User participation in testing was facilitated by the release of a first version of the software several months before the production version. This release contained most of the functionality and permitted end-to-end experimentation while several of the more challenging features were being completed.

Lincoln has ongoing maintenance responsibility for WebVDME, which has evolved from a custom application to a pharmacovigilance software product that can be installed at a user organization or provided as an external service. Several other pharmaceutical companies have begun pilot testing of the application (including one that began testing during the latter stages of the implementation project described here), and the WebVDME software has been delivered to FDA and to CDC as well. Recently, Lincoln entered into a formal Cooperative Research and Development Agreement (CRADA) with FDA, under which WebVDME is being enhanced to serve as an internal data mining resource for medical officers and safety evaluators associated with the monitoring of marketed drugs and vaccines.

7. APPLICATIONS AT GSK

At GSK, WebVDME data mining has been used to assist with safety screening as part of day-to-day pharmacovigilance and risk management practice, and to carry out special projects for hypothesis generation and refinement as an adjunct to formalized methods (clinical trials, pharmacoepidemiological studies).

Project areas have included:

- Comparisons of safety profiles of drugs within a drug class

- Exploration of possible drug-drug interactions

- Analysis of adverse events that could be attributed to one of several drugs in multi-drug therapy ("innocent bystander"/ "guilty bystander" problems)

- Peri-approval risk management planning for new drug applications

- Study of adverse events in special populations

- Analysis of signal evolution over time to understand trends

Example scenarios of use include:

Effects in Special Populations: As one input to deciding whether it was appropriate to conduct clinical trials in a special population of a highly effective drug for a serious condition, a study can be conducted to examine whether there is evidence from data mining that a known adverse event might occur more frequently in that special population. The AERS database can be essentially divided into two data sets based on membership in the special population, and signal scores and their 90% confidence interval values calculated for the drug-event pair in each group and compared.

Analysis of Possible Drug Combination: In the evaluation of a potential new therapy based on the combination of two marketed drugs (which had in some cases been co-prescribed by physicians), data mining can be used as one tool for ascertaining whether toxicity associated with the primary drug might be exacerbated by the presence of the secondary drug. The analytical approach can be based on recoding cases in the database to distinguish between cases where only the primary drug is reported and cases where both primary and secondary drugs are reported.

Analysis of Adverse Events in Polytherapy: When several products are co-prescribed, safety signals may emerge where it is difficult to discern which products are properly associated with the event of interest. Data mining can be used to conduct analyses of specific subgroups where each drug of interest is used in combination and also in the absence of others ("monotherapy"); this is illustrated for hypothetical data in Figure 6 below. While these analyses must be interpreted with extreme caution and are intended only as a first step towards hypothesis generation, comparison of signals seen in these groups may help to clarify associations and potentially help to direct the focus of future clinical studies.

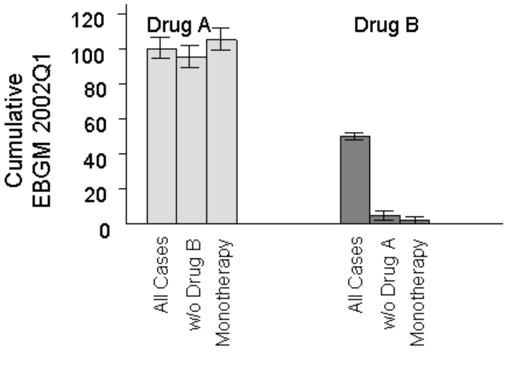


Figure 6. Discriminating drug contributions in polytherapy

8. CONCLUDING REMARKS

We were fortunate in collecting key user input early in the design of this application, and many aspects of the development were accomplished more easily than anticipated. Several areas were more difficult and yielded only to sustained and repeated effort, including data preparation and data cleaning, performance measurement and optimization, and testing on large problems. These would be good areas for improved tools or techniques.

We believe that the development, deployment, and use of WebVDME demonstrates that sophisticated KDD tools can be "packaged" in a form where they can be used effectively by end-users to explore complex problems in a mission-critical application. These applications can be factored conveniently into functionality for staff with different roles (administrator, data mining specialist, data mining results reviewer). Currently available mainstream web-site development technologies (web and application servers, relational databases) are capable of supporting the creation, within reasonable time and budget constraints, of server-based KDD applications that can be readily deployed and maintained. Further, many of the components of the application (user administration, batch queue operation, output filtering, table and graph display) are relatively general purpose and will be reused by Lincoln in other data-centric applications.

9. TECHNICAL SUMMARY OF MGPS

For an arbitrary itemset, it is desired to estimate the expectation $\lambda = E[N/E]$, where N is the observed frequency of the itemset, and E is a baseline (null hypothesis) count; e.g., a count predicted from the assumption that items are independent. An itemset is defined by its members $i, j, k,...$, which occur as subscripts to N, E, and other variables, so that, for example, N_{ij} is the number of reports involving both items i and j, E_{ijk} is the baseline prediction for the number of reports including the itemset triple (i, j, k), etc.

A common model for computing baseline counts is the assumption of within-stratum independence; when E is computed under this assumption we shall often denote it by $E0$. Assume that all reports are assigned to strata denoted by $s = 1, 2, ..., S$. Let:

P_i^s = proportion of reports in stratum s that contain item i
n_s = total number of reports in stratum s

Baseline frequencies for pairs and triples are defined under independence as:

$$E0_{ij} = \Sigma_s\, n_s\, P_i^s\, P_j^s \qquad E0_{ijk} = \Sigma_s\, n_s\, P_i^s\, P_j^s\, P_k^s$$

For itemsets of size 3 or more, an "all-2-factor" loglinear model can be defined as the frequencies $E2$ for the itemsets that match all the estimated pairwise two-way marginal frequencies but contain no higher-order dependencies. For triples, $E2_{ijk}$ agree with the estimates for the three pairs:

$$\lambda_{ij}E0_{ij} \qquad \lambda_{ik}E0_{ik} \qquad \lambda_{jk}E0_{jk}$$

For 4-tuples, $E2_{ijkl}$ agrees with 6 such pairs, etc.

Then for itemsets of size 3 or more we compare the estimated frequency to the all-2-factor prediction by simple subtraction. For example, in case of triples:

$$Excess2_{ijk} = \lambda_{ijk}E0_{ijk} - E2_{ijk}$$

The parameters λ above are estimated by their geometric means, denoted *EBGM,* of their empirical Bayes posterior distributions. For simplicity, the formulas below use just two subscripts, for itemsets of size 2, such as the occurrence of drug i and symptom j in a medical report. Estimates for other itemset sizes are computed analogously. Let:

N_{ij} = the observed counts

E_{ij} = the expected (baseline) counts

$RR_{ij} = N_{ij} / E_{ij}$ = ratio of observed to baseline

We wish to estimate $\lambda_{ij} = \mu_{ij} / E_{ij}$, where $N_{ij} \sim \text{Poisson}(\mu_{ij})$. Assume a superpopulation model for λ_{ij} (prior distribution) based on a mixture of two gamma distributions (a convenient 5-parameter family of distributions that can fit almost any empirical distribution):

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P) = P\ g(\lambda; \alpha_1, \beta_1) + (1 - P)\ g(\lambda; \alpha_2, \beta_2)$$

$$g(\lambda; \alpha, \beta) = \beta^{\alpha}\ \lambda^{\alpha-1}\ e^{-\beta\lambda} / \Gamma(\alpha)$$

Estimate the prior distribution from all the (N_{ij}, E_{ij}) pairs. Estimate the 5 hyperparameters:

$$\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, P)$$

by maximizing the likelihood function L(θ) in 5 dimensions:

$$L(\theta) = \Pi_{i,j}\{P\ f(N_{ij}\ ; \alpha_1, \beta_1, E_{ij})\ +\ (1 - P)\ f(N_{ij}\ ; \alpha_2, \beta_2, E_{ij})\}$$

$$f(n; \alpha, \beta, E)\ =\ (1 + \beta/E)^{-n}(1 + E/\beta)^{-\alpha}\ \Gamma(\alpha + n) / \Gamma(\alpha)\ n!$$

If a threshold (minimum count) for the observed counts is used, these formulas are modified to condition on $N_{ij} \geq n^*$ (where n^* = the threshold count).

Given θ, the posterior distributions of each λ_{ij} are also a mixture of gamma distributions used to create "shrinkage" estimates. Assuming that θ and E are known, then the distribution of N is:

$$\text{Prob}(N = n)\ =\ P\ f(n; \alpha_1, \beta_1, E) + (1 - P)\ f(n; \alpha_2, \beta_2, E)$$

Let Q_n be the posterior probability that λ came from the first component of the mixture, given $N = n$. From Bayes rule, the formula for Q_n is:

$$Q_n = P\ f(n; \alpha_1, \beta_1, E)/[P\ f(n; \alpha_1. \beta_1,E)+(1 - P)\ f(n; \alpha_2, \beta_2, E)]$$

Then, the posterior distribution of λ, after observing $N = n$ can be represented as:

$$\lambda|N = n\ \ \sim\ \ \pi(\lambda; \alpha_1 + n, \beta_1 + E, \alpha_2 + n, \beta_2 + E, Q_n)$$

where (as above):

$$\pi(\lambda; \alpha_1, \beta_1, \alpha_2, \beta_2, P)\ =\ P\ g(\lambda; \alpha_1, \beta_1) + (1 - P)\ g(\lambda; \alpha_2, \beta_2)$$

We will use the expectation value:

$$E[\log(\lambda_{ij})\ |\ N_{ij}, \theta]$$

as a means of estimating the "true" value of λ_{ij}

To obtain a quantity on the same scale as RR, we define the Empirical Bayes Geometric Mean:

$$EBGM_{ij} = e^{E[\log(\lambda_{ij})|\ N_{ij},\ \theta]},\ \ \text{where}$$

$$E[\lambda\ |\ N = n,\theta] = Q_n\ (\alpha_1 + n)/(\beta_1 + E) + (1 - Q_n)\ (\alpha_2 + n)/(\beta_2 + E)$$

$$\begin{aligned} E[\log(\lambda)\ |\ N = n,\theta] = \quad & Q_n \quad [\psi(\alpha_1 + n) - \log(\beta_1 + E)]\ + \\ & (1 - Q_n) \quad [\psi(\alpha_2 + n) - \log(\beta_2 + E)] \end{aligned}$$

where $\psi(x) = d(\log \Gamma(x))/dx$. In the same way, the cumulative gamma distribution function can be used to obtain percentiles of the posterior distribution of λ. The 5th percentile of λ is denoted:

$EB05_{ij}$ = Solution to: $\text{Prob}(\lambda < EB05\ |\ N_{ij}, \theta) = 0.05$
and is interpreted as a lower 1-sided 95% confidence limit.

10. ACKNOWLEDGMENTS

Work on the MGPS data mining technology and its embodiment in WebVDME has been supported in part by a Small Business Innovation Research Grant to Lincoln Technologies from the NIH National Center for Research Resources (1-R44-RR17478).

We acknowledge Ana Szarfman of FDA for her passionate interest in drug safety and for her inspiration and enthusiasm through many years of collaboration.

We thank GSK colleagues Mark Mickunas for serving as the information technology project manager and for assisting with the preparation and transfer of the OCEANS data; and Eric Smith for being a true "super user" and for contributing importantly to the definition, evaluation and testing of the WebVDME software. We would also like to thank the entire GSK user group for their input into the design of the system.

We also express our appreciation to the following Lincoln Technologies colleagues: Leslie Hepburn for producing accurate and complete system documentation, help pages, and test scripts, all under the pressure of very tight timetables; Rick Ferris for filling in on short notice as a versatile application developer, end-user trainer, and system tester; Renel Fredriksen for continually enhancing both the functionality and performance of the MGPS C++ code; Mike Prindle for getting the architecture right and for personally building the lion's share of the web site software; Mike Yang for his cleverness, patience, and persistence in carrying out the data pre-processing necessary to turn the messy "as-delivered" FOI AERS data into a valuable safety signaling resource; and Chan Russell for his multiple technical and managerial contributions and for his invaluable assistance in the editing and production of this paper.

11. REFERENCES

[1] DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA Spontaneous Reporting System (with discussion). *Am Stat*. 1999; 53(3): 177-190.

[2] DuMouchel W and Pregibon D. Empirical Bayes screening for multi-item associations. *Proc KDD 2001*, ACM, NY, 67-76.

[3] Szarfman A, Machado SG, and O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety*. 2002; 25(6): 381-392.

[4] O'Neill RT and Szarfman A. Some FDA perspectives on data mining for pediatric safety assessment. *Curr Ther Res Clin Exp*. 2001; 62(9): 650-663.

[5] Evans SJW, Waller PC and Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepi and Drug Safety*. 2001; 10: 483-486.

[6] Bate A, Lindquist M *et. al.* A data mining approach for signal detection and analysis. *Drug Safety*. 2002; 25(6): 393-397.

[7] Agrawal R and Srikant R. Fast algorithms for mining association rules. *Proc 20th VLDB Conf.* Santiago Chile, 1994.

[8] DuMouchel W, Volinsky C *et. al.* Squashing flat files flatter. *Proc KDD 1999*. ACM Press, San Diego CA, 6-15.