

OBSERVATIONAL MEDICAL OUTCOMES PARTNERSHIP

Partnership Update

Year One Accomplishments

At the end of the first full year of operations for the Observational Medical Outcomes Partnership (OMOP), we should recognize that a great deal has been accomplished. The OMOP Research Laboratory is up and running and six distributed partners are now part of the OMOP research core, along with thirteen-plus methods development partners. Contributions to the OMOP Methods Library are growing, and the OMOP Principal Investigators have completed the initial set of definitions for our Health Outcomes of Interest (HOI). We also delivered a first-of-its-kind Observational Medical Data Set Simulator, or OSIM, to support methods development and testing. We launched a contest to attract methods developers to contribute innovative statistical analyses and data-mining algorithms to our program.

The OMOP research core has grown to include over 100 researchers and technologists, each contributing to our ambitious research agenda. Our plans for 2010 are just as ambitious: our distributed partners and research lab staff will put the Methods Library to work across the wide array of data environments and will begin to analyze the results of these experiments.

In this issue of the OMOP newsletter, you will find a brief overview of OMOP's 2009 Symposium, an update on the OMOP Research Laboratory, a primer on the OMOP Methods Library, and an interview with OMOP Scientific Advisory Board member David Page. We are fortunate to have the ongoing support and engagement from all of our stakeholders, advisory boards, and executive board, and we look forward to another successful year.

In this Issue

Partnership Update

2009 OMOP Symposium

Methods

Advisory Board Q & A

The Research Lab

Follow OMOP

2009 OMOP Symposium

In Review

The Observational Medical Outcomes Partnership (OMOP) held its first annual Symposium on November 12, 2009, in Bethesda, Maryland. The primary purpose of the Symposium was to discuss the progress of Phase 1 of the OMOP Research Program and to engage attendees in a dialogue regarding a number of OMOP's work streams. Approximately 200 people attended the event, and over 100 individuals followed the day's activities via webcast. The wide spectrum of stakeholders that comprise the OMOP public/private partnership was well represented at the event, as were all of the organizations within OMOP's network of research partners.

Throughout the day, the audience heard and engaged in discussion with the OMOP Principal Investigators, members of the OMOP Advisory Boards, and members of the Executive Board. A variety of topics pertaining to the use of observational healthcare data, technical requirements for active surveillance, data models, statistical and analytical methods development, and Health Outcomes of Interest (HOI) definitions dominated the marathon agenda.

Thomas P. Scarnecchia, OMOP Executive Director, kicked off the event by welcoming the participants and attendees to the inaugural OMOP Symposium and then introduced Dr. Janet Woodcock, Director, Center for Drug

Evaluation and Research, Food and Drug Administration (FDA), and Chair of the OMOP Executive Board, who set the stage for the day by stressing the importance of OMOP's research agenda.

The keynote address was delivered by Dr. Clement McDonald, Director of the Lister Hill National Center for Biomedical Communications at the National Library of Medicine, and member of the OMOP Healthcare Informatics Advisory Board. Dr. McDonald engaged the audience in an animated discussion about the challenges of using observational data and the realities of the data that is available today to support observational studies.

A general overview of the OMOP project and its progress was presented by Mark Overhage, MD, PhD, of the Regenstrief Institute, Inc., followed by Bram Hartzema, PharmD, MSPH, PhD, FISPE, University of Florida, who discussed summary results of the OMOP data provider survey.

The heart of the meeting consisted of panel discussions facilitated by members of the OMOP research team and the Advisory Boards. The first panel was moderated by Paul Stang, PhD, Johnson & Johnson Pharmaceutical Research and Development, and was comprised of the lead investigators from the OMOP Distributed Research Partners, who described their data environments and their experiences working with the



From left to right:

Emily Welebob, Christian Reich, Program Managers; Patrick Ryan, Co-Investigator; Thomas Scarnecchia, Executive Director; Marc Overhage, Paul Stang, Judith Racoosin, Bram Hartzema, Principal Investigators

OMOP Research Plan within those environments. This was followed by four panel discussions held as concurrent breakout sessions that covered “Technical Requirements for Active Surveillance”, “Defining Health Outcomes of Interest”, “OMOP Common Data Model and Vocabulary”, and “Methods Development and Evaluation”. These panels were led by members of the OMOP research team and were reported back to the general session by members of the OMOP Advisory Boards.

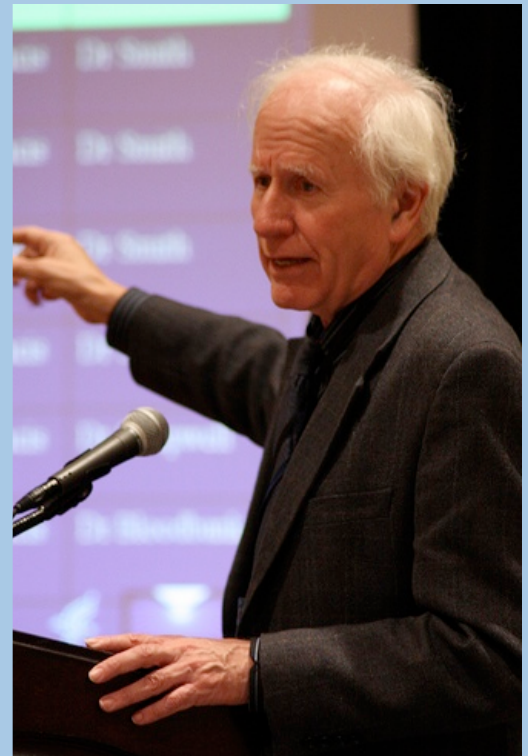
OMOP was also very fortunate to have representatives from several major international institutions present international perspectives on drug safety initiatives, including the EU-ADR, the Pharmaceuticals and Medical Devices Agency of Japan (PMDA) and the Medicines and Healthcare Products Regulatory Agency (MHRA).

To conclude, Gerald Dal Pan, MD, MHS Director, Office of Surveillance and Epidemiology, FDA, Rebecca Burkholder, Vice President Health Policy, The National Consumers League, Ronald L. Krall, MD, former Senior Vice President and Chief Medical Officer, GlaxoSmithKline, and Stephen Spielberg, MD, PhD, Marion Merrell Dow Chair in Pediatric Pharmacogenomics, Children’s Mercy Hospital, and Dean Emeritus, Dartmouth Medical School, gave their perspectives on drug safety systems and the importance of research. This panel shared their thoughts regarding the quality of available data to support observational studies, and the importance of conducting research to understand both the potential and the limitations of using healthcare data today for such studies.

The Symposium was closed by Amy Porter, Executive Director of the FNIH, with an open invitation to the broader research community to participate in the evaluation of other sources of data, in the creation of standards around a common data model and vocabulary mappings, in the development of standardized software tools and analytical methods, and in the expansion of the Health Outcomes of Interest Library.

Overall, the Symposium proved to be successful in achieving its goal of stimulating discussion within our stakeholder community and research network regarding the OMOP Research Program. We anticipate one or more symposiums during 2010 to continue this important dialogue and build upon the community that is emerging in observational methods development.

The presentation materials from the Symposium are available at <http://omop.fnih.org/OMOP2009Symposium>.



Dr. Clement McDonald,
OMOP Advisory Board Member



Rebecca Burkholder,
OMOP Executive Board Member



Dr. Stephen Spielberg,
OMOP Executive Board Member

Observational Analysis Methods

The Development and Evaluation

One of the goals of the Observational Medical Outcomes Partnership (OMOP) is to define methods that can assess the feasibility and utility of using observational data to identify and evaluate associations between drugs and health-related conditions. There are three distinct types of analysis within the scope of the Partnership's research (figure below). Each type of analysis may present different methodological challenges, require different algorithms, and utilize different data elements within the Common Data Model.

OMOP is building a library of methods developed for the OMOP Common Data Model. Initially all methods will be developed and tested within the OMOP Research Lab (OMOP RL) against the central data sets. The OMOP RL provides the core IT infrastructure needed to support research conducted using OMOP licensed data. The OMOP RL provides OMOP researchers with access to data, statistical analysis tools, and a methods library.

"OMOP provides a unique environment to rigorously and empirically assess the performance characteristics of alternative analysis approaches," stated Dr. David Madigan, Professor of Statistics at Columbia University, and OMOP Methods Lead. Methods will be developed to execute against the OMOP Common Data Model, and will be made publicly available to the broader research community.

In order to develop, test, and implement various methods (new and existing), OMOP initiated a call for participation to identify potential resources, tools, and

skills to develop and execute methods within the OMOP RL. The methods are being developed in various statistical programs – SAS, R, Perl, SQL, BBR, BXR, and PROLOG. This call resulted in fourteen methods collaborators. The collaborators and their methods can be found in Table 1 following this article.

In addition to these collaborators, OMOP is actively encouraging a broader collaboration with the research community. One vehicle for broader collaboration is the **OMOP Cup** that is in progress and has two related challenges. Challenge 1 explores how well the method works when provided an entire longitudinal data set. The goal is to accurately classify which drugs are associated with which outcomes. Challenge 2 evaluates the timeliness of detection of drug-event associations by having methods run against data sequentially as it accumulates over time.

OMOP is providing a large simulated data set (<http://omop.fnih.org/osim>) that resembles observational data that can be extracted from insurance claims or electronic medical records. OMOP seeks methods that will identify relationships in the data between drugs and medical outcomes. The goal is to develop methods that correctly identify true drug-event associations while minimizing false positive findings. Methods will be evaluated by how closely they predict the known relationships that exist in the data.

Detecting drug events in observational healthcare data can be a difficult problem to solve. To maintain the momentum, in early December 2009, OMOP offered

Three Analysis Types

Monitoring of Health Outcomes of Interest: The goal of this surveillance analysis is to monitor the relationship between any drug and a specific outcome of interest.

Identification of Non-Specified Conditions: This exploratory analysis aims to generate hypotheses from observational data by identifying associations between drugs and conditions that were previously unknown.

Evaluation of a Drug and Condition Association: This hypothesis-strengthening analysis is consistent with traditional pharmacoepidemiology practice where a drug-outcome association has been identified and more formal investigation is requested.

Progress Prizes to two OMOP Cup competitors who had the best method performance in each challenge. In addition, all competitors in the Top 10 will be recognized as OMOP Cup Award Winners and will be invited to submit their methods for publication on the OMOP website.

After two months of entries, the OMOP Cup has reached its first milestone with the announcement of the Progress Prize. Four awards totaling \$5,000 were given out to participants who came up with the highest-performing methods that improve the state of the art in identifying adverse drug reactions in medical records. The competition brought together competitors from epidemiology, drug safety, statistics, and machine learning in a cross-disciplinary challenge. The Progress Prizes are only the first part of the OMOP Cup, which culminates with \$15,000 in additional prizes in March 2010.

The OMOP Cup is still open to new entrants; go to <http://omopcup.orwik.com> for additional information.

Progress Prize Winners

Challenge 1:

1st Place: David Vogel and Eric Gottschalk of Data Mining Solutions won \$2,500

2nd Place: Robin Sabhnani of Carnegie Mellon University won \$1,000

Challenge 2:

1st Place: Robin Sabhnani of Carnegie Mellon University won \$1,000

2nd Place: Lisa Friedland of University of Massachusetts-Amherst won \$500

Table 1. Collaborators and Methods in Detail.

Program / Collaborator	Description
Disproportionality Analysis / Ivan Zorych	Disproportionality analysis methods for drug safety surveillance represent the primary class of analytic methods for analyzing data from spontaneous report systems (SRSs). SRSs receive reports that are comprised of one or more drugs, one or more adverse events (AEs), and possibly some basic demographic information (in addition to narrative and text data). Disproportionality analysis methods include the multi-item gamma-Poisson shrinker (MGPS), proportional reporting ratios (PRR), reporting odds ratios (ROR), and Bayesian confidence propagation neural network (BCPNN).
Multi-Set Case-Control Estimation / Ivan Zorych	The multi-set case-control estimation program leverages the basic design of a case-control study to enable estimates of drug and condition associations across a large set of drugs and conditions. The algorithm extracts the information necessary to yield an odds ratio, but can be applied simultaneously to multiple conditions (each acting as distinct case definitions for case-control substudies), and allows for all exposures to be evaluated for each outcome. In this manner, multi-set case-control estimation can be used to study specific drug and condition relationships, but is also scalable to be applied within an active surveillance context to both monitoring of health outcomes of interest, and identification of non-specified conditions. Multi-set case-control estimation can be seen as supplemental to traditional case-control surveillance programs, which typically execute on one drug and condition pair at a time and may offer additional study design customization for each analysis.
Case-Crossover / Brian Sauer	Case-crossover designs have been proposed as alternatives to case-controlled studies when assessing the relationship between a transient drug exposure and acute outcomes. The case-crossover design uses within subject comparisons of drug exposures over time to estimate the rate ratio of the outcome associated with the drug under study.

Observational Analysis Methods

The Development and Evaluation

Program / Collaborator	Description
CSSP / Lingling Li	The conditional sequential sampling procedure (CSSP) is a practical group sequential method with a finite number of interim tests to test if the drug of interest leads to an elevated risk compared to a comparator drug. CSSP is designed for settings in which information for both the drug of interest and the comparator drug accumulates over time. The original CSSP can only adjust for a few categorical variables. We are developing two enhanced variants, the propensity score (PS) -stratified CSSP and the PS-weighted CSSP, to allow the flexibility of adjusting for multiple categorical and continuous baseline covariates. The PS is defined as the conditional probability of receiving the drug of interest given measured covariates.
Local Control / Robert Obenchain	The Local Control (LC) approach to analysis of observational studies is a robust alternative to traditional Covariate Adjustment methods using multivariable statistical models. LC focuses on making fair head-to-head comparisons between two treatments for the same condition. The key LC strategy is to make treatment comparisons only within clusters of relatively well-matched patients. In other words, LC is a "hypothesis-strengthening" method with a unique way of adjusting for imbalance, selection bias, and confounding among treatment cohorts. Finally, LC quantifies the full distribution of observed Local Treatment Differences (LTDs), thereby characterizing all aspects of patient differential response (PDR) to treatment.
Cohort Methods / Siu Hui	We are implementing a series of methods with increasing sophistication in the domain of hypothesis generation for identifying non-specified conditions. This exploratory analysis aims to generate hypotheses from observational data by identifying associations between drugs and conditions for which the relationships were previously unknown. This data mining process serves as an initial step in signal detection to prioritize review of drug and outcome pairs to ensure patients safety.
Case-Control Surveillance / Karen Benoit	The case-control matching surveillance method consists of a SAS program with an outer condition loop, a case-control-matching routine, and an inner drug loop. The user must supply basic study parameters, such as enrollment criteria, desired number of controls per case, matching options, and persistence and exposure windows. If the user does not supply lists of conditions of interest or drugs of interest, the program will run against all conditions and all drugs within the database. The condition loop characterizes all patients who meet the enrollment criteria as cases or possible controls by whether they had the condition of interest. Cases are matched to controls by year of birth and gender and, if specified by the user, location and/or race. The temporal relationship between drug exposure and condition is evaluated, and cases and controls are summarized into a classic 2-by-2 table from which an Odds Ratio is calculated.
Statistical Relational Learning / David Page	Many standard statistical methods work with data in a single flat file or table. Statistical relational learning (SRL), by contrast, works directly with relational data distributed across many tables. In the OMOP context, for example, the data reside in person tables, drug era tables, condition era tables, etc. The Page group is at the forefront of current SRL research. The SAYU algorithm extracts rules from relational data. Current work is adapting SAYU to the OMOP context, focusing on both the HOI task as well as non-specified outcomes.
MaxSPRT / Lingling Li	The maximized sequential probability ratio test (MaxSPRT) is a sequential analysis method designed for continuous or frequent (e.g., weekly) monitoring of a potential elevated adverse event risk following an introduction of a drug or vaccine of interest. It consists of two variants, the Poisson MaxSPRT for historical controls and the Binomial MaxSPRT for matched concurrent controls. We propose to implement the Poisson MaxSPRT and further enhance its capability of adjusting for confounding by fitting a Poisson regression model to the historical control data to estimate the expected baseline rates.

Observational Analysis Methods

The Development and Evaluation

Program / Collaborator	Description
High Dimensional Propensity Scoring Cohort / Alan Brookhart and Eric Brinsfield	High-dimensional propensity scoring is a multi-step algorithm to implement high-dimensional proxy adjustment in observational data. Steps include (1) identifying data dimensions, e.g., diagnoses, procedures, and medications; (2) empirically identifying candidate covariates; (3) assessing recurrence of codes; (4) prioritizing covariates; (5) selecting covariates for adjustment; (6) estimating the exposure propensity score; and (7) estimating an outcome model. Used in conjunction with a new user cohort design, high-dimensional propensity scoring offers a novel approach to minimizing confounding when assessing the relative association among patients exposed to alternative medicines and the occurrence of a health outcome of interest.
Bayesian Logistic Regression / Ivan Zorych	Bayesian logistic regression is a high-dimensional statistical method that allows for millions of covariates to predict occurrence of ADEs. The Bayesian approach to logistic regression has several advantages, including: the avoidance of over-fitting, efficiency during model-prediction time, and scalability to large numbers of covariates.
Cohort Screening / Ivan Zorych and Prosanos Corporation	Cohort screening is an extension of a traditional cohort epidemiology design where the rate of ADE occurrence can be compared across groups of patients exposed to different medicines. Cohort screening can allow for comparisons within a cohort population, between treatments, and relative to the overall population at large.
Univariate and Multivariate Self-Controlled Case Series / Shawn Simpson	The self-controlled case series (SCCS) method is a study method for investigating the association between a transient exposure and an ADE. The method uses only cases; no separate controls are required as each case acts as its own controls. OMOP is developing two approaches for SCCS: 1) a univariate approach that can estimate the association of one drug with a given condition, and 2) a multivariate version that estimates the association between many drugs and a given outcome.
Temporal Pattern Discovery / Niklas Noren	Temporal pattern discovery is a novel methodology for event history data focusing explicitly on the detailed temporal relationship between pairs of events. The proposed measure contrasts the observed-to-expected ratio in a time period of interest to that in a predefined control period. The method applies statistical shrinkage towards the null hypothesis of no association. This provides protection against spurious associations and is an extension of the statistical shrinkage successfully applied to large-scale screening for associations among events in cross-sectional data, such as large collections of adverse drug reaction reports.

Advisory Board Q & A

Dr. David Page, OMOP Scientific Advisory Board Member

An important component of the Observational Medical Outcomes Partnership (OMOP) effort is to develop a central repository of potential methods and their characteristics in order to facilitate the structure and development of protocol concepts. Currently, there are several methods contributors to OMOP who are developing and evaluating analysis methods. The group of methods being developed is undergoing feasibility testing within the OMOP Research Lab (OMOP RL) and on simulated data prior to being released to the OMOP Distributed Partners and subsequently placed into the public OMOP Methods Library (<http://omop.fnih.org/MethodsLibrary>).

In December 2009, Dr. David Page, Professor, Departments of Biostatistics and Medical Informatics and Computer Science, University of Wisconsin-Madison, and member of the OMOP Scientific Advisory Board (SAB) shared with us his thoughts about using analysis methods on observational healthcare data sets for research and the opportunities that lie ahead when sharing methods broadly.



David Page received his Ph.D. in computer science from the University of Illinois at Urbana-Champaign in 1993. He was a research scientist in the Oxford University Computing Laboratory from 1993 to 1997, where he also served as a visiting member of the Faculty of Mathematics from 1995-1997.

Dr. Page is now a professor at the University of Wisconsin-Madison, in the Dept. of Biostatistics and Medical Informatics (School of Medicine and Public Health) and Dept. of Computer Sciences. He is also a member of the University of Wisconsin Comprehensive Cancer Center and the Genome Center of Wisconsin, and he is a member of the scientific advisory boards for the Wisconsin Genomics Initiative and the Observational Medical Outcomes Partnership.

Dr. Page's primary research interests are in machine learning analysis of clinical and genetic data and in learning statistical models from multi-relational data.

Dr. Page has substantial experience applying machine learning and data mining to clinical and genetic data to construct predictive models, and is collaborating with OMOP as a methods collaborator specifically with statistical relational learning (SRL) to address the need for algorithms to analyze data sets with relational (multi-table and/or temporal) data.

Q: What is the best way to get a computer-implemented method of analyzing a data set of pharmacovigilance data? What does OMOP need to do to make sure that the Distributed Partners can execute the methods?

DP: There are at least two aspects when developing or enhancing analysis methods. The first critical aspect is the need to get a lot of people involved in order to develop novel methods and test existing methods to see what works best on what type of data. Today, researchers implement various statistical software packages in a lot of development languages but during development we do not want early on to force people to have "commercial grade" methods in one platform with bells and whistles. If we do, the outcome may be that we stop thinking of new ideas and develop software only for one platform. We do not want to unify the platform too early--then we risk no exploration and the development of novel new methods. The second issue is the exact opposite--to make it easier for the research community to implement one platform [that] would be efficient; hence, we have competing goals. At this point, with all the disparate healthcare data sets and current state of post-market drug surveillance, we still need to explore and test in various environments as we do not know what should be the recommended practice for the different data sets and various

drug safety questions.

It is important for OMOP to emphasize the entire research program for methods development, testing, and evaluation with Distributed Partners. I would like to see the Distributed Partners encouraged to have in their computational environments SAS, R, and other statistical tools so that they can implement the majority of the OMOP methods. Once the Distributed Partners can execute the various methods, then it may be a future activity for OMOP to assist in taking analysis methods code developed in the different languages and consolidate [it] into a platform. One of the strengths of OMOP's methods development strategy is the availability of the OMOP RL. Not all researchers have access to electronic healthcare data to develop and test upon. The other strength is [that] the common data model and the simulated data [are] in the same format. For our group, it has been a tremendous benefit to test on the simulated data (Observational Medical Dataset Simulator - <http://omop.fnih.org/osim>) [...] and then use the licensed healthcare data sets in the OMOP RL.

Q: Discuss the benefits of the OMOP Methods Library.

DP: The Methods Library is a huge step forward because it allows us to compare our algorithms with other leading algorithms, thus enhancing our development. It is important to have the data sets to evaluate the methods upon as this helps us to have rigor and formulate critical comparisons. Within the machine learning and data mining community, there are sites [...] that contain a collection of algorithms for data analysis and predictive modeling. These sites, such as the OMOP Methods Library, give access to different industries and collaborators--academic and research, industry, entrepreneurs, and government.

Q: A number of methods for assessing signals in pharmacovigilance exist and are currently

implemented. Is there a "best" method for or [can you] discuss the pros and cons of the existing methods for post-market surveillance?

DP: OMOP is trying to answer these questions--we cannot answer until we have all the methods (there is no single method to implement) and analysis. OMOP is doing this by bringing everybody to the table and providing standard development guidelines in the Methods Library. We have to be attentive to the data sets that the methods are executed upon. The size and characteristics of the data set you are applying the method on are important to know and be aware of. Learning curves show how the methods vary with the size of the data set. In a learning curve you typically see the accuracy ramp up quickly with more data, and then at some point it levels off, so that adding more data beyond that point does not help much. But if you are looking for a rare adverse drug event (ADE) then you may need a lot of data before you reach this leveling point; that is, before you get all the incidences you need in order to identify the ADE.

Q: When developing methods, how do you balance the trade off between false positives and false negatives?

DP: This same trade off has been in the news a lot lately; for example, with the great debate sparked when the U.S. Preventive Services Task Force announced new guidelines last month for mammograms regarding screening. There are two types of error that can occur with such things as laboratory tests, trials, and predictions. A false positive is when there is no disease but the results come back as positive. A false negative is when there actually is a disease but the results come back as negative. It is a trade off and we need to know what the trade off looks like. ROC and Precision-Recall (PR) curves are great ways to work on this tradeoff and will allow us to get more accurate methods. To weigh the false positives and false negatives will require much more discussion than this [...] newsletter. While we

obviously don't want false negatives – ADEs that we miss – we also cannot tolerate too many false positives if every positive prediction takes great expense and effort to track down. We will have to estimate the cost/danger of such followup on false positives and also estimate the cost of each false negative, recognizing that the cost of a false negative depends in part on the nature of the ADE (is it a myocardial infarction or a rash?) and how long it may take to accumulate enough data to make a better prediction of that ADE (might we pick it up more accurately next month?). Once we have a good estimate of these costs, then we can adjust our tradeoff using our ROC and PR curves.

Q: OMOP is studying two analysis problems: identification of non-specified conditions and monitoring of health outcomes of interest. The notable difference is how the outcome is defined. Describe how methods may perform differently for each of these problems.

DP: The monitoring of health outcomes of interest is complex, but the easier problem of the two since you know what outcome/ADE you are trying to find. You simply look for drugs associated with that outcome. When you do not know the ADE it is much tougher. For example, the ADE may not correspond to an ICD9 code or may not have been characterized previously. ADEs are not always characterized, in which case methods for monitoring cannot be directly applied to identification of non-specified conditions.

Q: What is the value of bringing together different disciplines to foster innovative methods development?

DP: Coming from different viewpoints and bringing different disciplines together will yield new algorithms. Computer science, statistics, epidemiology, and many other fields have been growing closer together in what they are doing.

You see people from the various disciplines, and having them talk to one another in the context of pharmacovigilance can only benefit all of us. I see a lot of interest in collaborating across these disciplines. Many exciting new approaches can come from combinations of existing approaches from these different fields, and in some cases we will see interesting similarities and other relationships between approaches from different fields. Sometimes these relationships are obscured because we have different words for the same things -- because the fields use different vocabularies.

Q: What role do you see for simulated data in evaluating the performance of methods? What about applying the method to real sources across the OMOP data community?

DP: There are times that one should be suspicious of simulated data, but it can be very helpful in getting new methods up and running. One must be cognizant of what important aspects of the "real data" it is missing; however, there are two major benefits for using simulated data in developing and evaluating the performance of methods. It lets us develop the algorithms and get the code up and running before it goes to the OMOP RL. Also, for cases like the OMOP Cup you cannot give the real data to people--this gives OMOP a way to compare methods when you could not otherwise compare them (no access to real data). Again, you can get useful feedback from these comparisons. If the methods do not work well on the simulated data ("truth"), then it may not work well (computational feasibility) on the real data and require further development.

Research Lab

Multiple Technology Challenges and Novel Approaches

When the idea of establishing an informatics environment to support the research activities of the Observational Medical Outcomes Partnership (OMOP) was originally proposed to the initial organizers and later the Executive Board, the concept was simply to acquire technology to manage data and support the analysis activities of the partnership. What ultimately emerged over the course of the past twelve months goes far beyond providing computing resources to the OMOP research team. The evolution of the OMOP Research Lab, or OMOP RL, from a purely technical computing platform to a multi-disciplined, virtual learning laboratory for observational methods research is a reflection of the breadth of the challenge the OMOP research team faces in developing a large portfolio of analytical methods to run across a distributed set of disparate data sources. While the OMOP RL does fulfill its core mission of supporting the evaluation of databases and the development and evaluation of analytical methods, the OMOP RL also brings a variety of disciplines and subject matter experts together that are defining the “information architecture” of distributed analysis as well as functioning as a center of coordination across our network of research partners. These partners support OMOP as contributors of analytical methods or by independently running the method libraries on their own data within their own environment.

As a consequence, OMOP’s RL has evolved to be centralized development center as well as a coordination center for our research partners. It is also evolving to respond to a host of technology challenges that our research agenda is uncovering.

Centralized Development Center

The development center provides the computational and data management facilities needed to support the OMOP research program. It houses the five de-identified observational data sets OMOP purchased, covering Medicaid, Medicare, commercial claims, and EHR data sets, each between 200 gigabytes and 1 terabyte in size. These data sets were transformed from their native raw formats into the OMOP Common Data Model (CDM),

standardizing their structure (database schema) and content (standard vocabulary). This enables the development of a library of analytical tools, but it also allows researchers to compare and evaluate these methods against each other and across a variety of different data sources.

Both the OMOP research team and our method developers are accessing the OMOP RL securely from the public internet. The datasets currently being utilized remain within the OMOP RL and no data leaves the lab for analysis.

The development center is also home to the efforts needed to assemble and map a wide variety of healthcare terminologies into a logical set to support analysis activities across the wide variety of data environments we are encountering within our partner network. The mapping process and resulting libraries are housed within the OMOP RL.

Coordination Center

The OMOP RL is also used to manage the distributed activities of the OMOP Research Core. The OMOP RL functions as a software distribution and technical support center for our distributed partners. This requires the distribution of well-documented software code that had been tested and benchmarked within the OMOP RL. The OMOP RL also functions as the collection center for analysis results from each distributed partner. We anticipate that several terabytes of analysis results will be returned from our six distributed partners. In effect, the OMOP RL is creating a small scientific community by providing support for secure communications and controlled information exchange amongst our partners.

Unique Technical Needs

Despite the commonalities with clinical research, where patients are exposed to drugs and efficacy and adverse events data are recorded and statistically analyzed, the paradigm of observational outcomes research follows a different pattern (Table 2), resulting in a technology environment that is much more computationally demanding and dynamic. These

differences influenced the technology and practices deployed within the OMOP RL and are shaping its future direction.

Observational analysis calls for database and analysis servers that can support very large tables and data transformation activities that can deal with the relative “dirtiness” of source data. Methods are also being contributed and developed in a wide variety of software environments such as SAS, R, C++, Perl, Python, and Prolog, while utilizing iterative prototyping programming techniques resulting in many variants of the same analysis method. This rules out the typical software development and validation methodologies deployed in a clinical environment and requires many of the development practices utilized in bioinformatics research environments. As a result, we are also exploring new approaches to validate our processes and software within the OMOP RL.

The Challenge of Scale

Although the OMOP RL is successfully fulfilling its mission, it still faces a number of challenges. The traditional approach to developing observational methods is to build a routine for the surveillance of one drug or HOI. While this approach is useful to address a specific drug safety question, it is not practical to support OMOP’s goals. OMOP is extending observational methods to run across many drugs and conditions as defined in our HOI-Drug Pairs. In theory, a method could be run against 40 million combinations of drugs and conditions. This will result in the need to carefully manage the workload on our analysis servers and may mean that some resource intensive analytical methods will not be practical in our infrastructure or that of our distributed partners.

As a solution, OMOP is exploring how to make the computing environment within the OMOP RL far more elastic to address peak demands of

Table 2. Characteristics of Clinical and Observational Outcomes Research.

Element	Clinical Research	Observational Outcomes Research
Quality	High, collected for the purpose of the study. Data are queried for completeness.	Lower, collected for the purpose of insurance claim processing or medical record keeping; collected ad hoc with unknown completeness.
Level of Control	High, all data have a full chain of custody from the source records to the data set for statistical analysis and FDA submission.	Light, some data are pooled, sold commercially and de-identified for HIPAA compliance with limited possibility to access the source records.
Number of Subjects	Typically a dozen to the low thousands.	High, between 1.5 and 160 million lives.
Depth of Data Per Person	High, typically hundreds of data items per patient, with data elements collected according to a predefined study protocol.	Varying, but for purposes of outcome research only a few data elements relevant (drugs, outcomes, demographics, procedures, visits).
Analysis Methods	Relatively simple and well established, relying on well-defined datasets.	Novel and complex, with the attempt to control for the relatively “dirty” data.
Analysis Environments	SAS	SAS, R, C++, Perl, Python, Prolog

computational intensive analytical methods. The OMOP Statistics Team has routinely utilized the Amazon EC2 Cloud Computing environment to generate and analyze simulated data. We are now experimenting with SAS on the Amazon EC2 environment as a means to support methods development and testing. Research into the implications on the design of analytical methods and data management practices is underway. Several of our methods developers are redesigning their algorithms and software to exploit massively parallel computing architectures. Since Cloud computing provides an economically attractive on-demand computing infrastructure, we hope to learn about the feasibility of utilizing these services for observational analysis applications.

Static Data

For the current series of experiments, OMOP is utilizing a snapshot of data provided by GE and Thompson Reuters from the spring of 2009. The rationale behind the snapshot was to focus our resources on methods and not on creating a real-time data environment. In order to accommodate updates to the OMOP datasets, the transformation process from the native data formats to the OMOP Common Data Model will need to be adjusted to process incremental updates. Vocabularies will also require processes and knowledgeable staff to periodically update them to reflect the dynamic nature of medical terminologies across our disparate data sources.

While creating a real-time data environment is easily achievable within the current lab, the vocabulary effort will most likely require a broad healthcare community effort, as well as the development of standards to which the community can conform.

It is also important that these processes be replicated across our network of distributed data partners. This will result in increased demand for staff and processing resources at each distributed partner. While we are not attempting to create a real-time data environment within the OMOP RL, we are studying the question to understand the impact and inform those efforts that require a real-

Technology Behind the OMOP Research Lab

- **Oracle Database Server:** Sun M5000 server with 16 processors (8x dual-core CPUs), 64 GB memory, redundant power, network and storage, dual-port FibreChannel (FC) host bus adapters (HBA) for Storage Area Network (SAN) connectivity
- **Statistical Application Server (SAS, R):** 2 Sun M5000 servers with 12 processors each (6x dual-core CPUs, 32 GB memory, redundant power, network, and storage, dual-port FibreChannel (FC) host bus adapter (HBA) for Storage Area Network (SAN) connectivity
- **Statistical Analysis Cloud (SAS, R):** up to 250 processors
- **SAN:** 20 TB of total usable storage
- **Virtual Citrix Secure Gateway/Web Interface and Virtual Citrix XenApp/Presentation remote access servers**

time environment.

Early Lessons

The OMOP RL has evolved rapidly since the effort began to build it last spring. While it is successfully supporting the OMOP research program, it is also providing a unique learning laboratory for the distributed aspects of the partnership as well as technological challenges that large scale observational analyses bring. That said, for many of our researchers, the real value of the OMOP RL is the observational data resources that OMOP has made available and that would otherwise be out of their reach.

Follow OMOP

Upcoming Events

2nd Annual Sentinel Initiative Public Workshop:

January 11, 2010
Washington, DC

DIA/FDA CDER/CBER Computational Science

Annual Meeting:

March 22 – 23, 2010
Bethesda, MD

International Society for Pharmacoepidemiology

(ISPE) Mid-Year Meeting:

April 10 -12, 2010
Raleigh, North Carolina

Midwest Biopharmaceutical Statistics Workshop:

May 18-20, 2010
Muncie, IN

16th World Congress of Basic and Clinical Pharmacology :

July 19, 2010
Copenhagen, Denmark

JSM 2010 - Joint Statistical Meetings:

July 31 – Aug 5, 2010
Vancouver, BC

OMOP Newsletter

©2009 Foundation for the National Institutes of Health.
All Rights Reserved.

Observational Medical Outcomes Partnership

Foundation for the National Institutes of Health
9650 Rockville Pike
Bethesda, MD 20814-3999
Phone: 301-402-5311
Website: <http://omop.fnih.org>



FOUNDATION
FOR THE
National Institutes of Health