Review of Observational Analysis Methods

Authored by: Observational Medical Outcomes Partnership
Corresponding author: Patrick Ryan, GlaxoSmithKline, patrick.b.ryan@gsk.com
Last revised: 13 February 2009

One of the goals of the Observational Medical Outcomes Partnership 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 scope of the Partnership's research. Each type of analysis may present different methodological challenges, require different algorithms, and utilize different data elements within the common data model.

The three analysis types are:

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. These analyses require an effective definition of the events of interest in the context of the available data (e.g. 'acute liver injury' may best be defined by a combination of medical diagnoses, procedure codes, and/or laboratory results). This is in contrast to identification of non-specified conditions, which may concurrently explore all outcomes for a given drug and will use a broad approach to define the set of potential outcomes. Where possible, outcomes definitions may be validated within the observational sources to provide parameters for interpreting monitoring analysis results.

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. This type of analysis is likely to be considered the initial step of a sequential review process, where all drug-outcome pairs are simultaneously explored and specific pairs identified for further attention. As such, identification may require relatively fewer fields from the common data model for the analysis, and may be executed with univariate and multivariate statistics and other data mining algorithms. It could be expected that a primary consideration for identification analyses is developing an efficient model to allow high-throughput computing across large sets of potential hypotheses about drug-outcome relationships.

Evaluation of a drug-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. Evaluation studies may require particular data elements specific to the study in question, and will commonly apply multivariate analyses to account for potential confounders.

Identification Methods Matrix

An important component of the OMOP effort is to develop a central repository of potential methods and their characteristics to facilitate the structure and development of protocol concepts. An initial list was developed by soliciting contributions and supporting publications from a workgroup, informal literature reviews, and informal reviews of presentations at relevant meetings. These findings are reported with the Methods Matrix.

The Methods Matrix

The matrix itself is intended to be a simplified display of the key classes (with examples) of methods, along with the working groups' current thinking regarding key features of the method as it may be applied or perform with observational data as well as some perceived strengths and limitations.

Please note that those methods that appear on the tab 'Other methods' in the spreadsheet have not been formally considered but may benefit from further consideration. There is an additional tab 'Key Patents' containing patents from a larger patent review that may also be of interest.

There are five classes of methods outlined. The reader is referred to summary documents for each of the methods in Appendix A.

1. Methods overview: Disproportionality analysis approaches from spontaneous adverse event reporting
2. Methods overview: Approaches based on selecting populations taking drug
3. Methods overview: Approaches based on selecting populations with condition
4. Methods overview: Surveillance approaches for general population
5. Methods overview: Other methods for consideration

The group of methods incorporated into OMOP will initially be feasibility tested and it is possible that many will fail feasibility testing because of the intensity of computing necessary. While this is purely a mechanical test of feasibility, it is possible that a method will fail at this stage and be eliminated from further consideration; some of the methods that are eliminated in this manner may actually have more desirable performance characteristics that will go untested because of the required computing feasibility test. It would be important for future research to re-consider our complete pool of methods to reduce the likelihood of excluding a method from consideration that may otherwise be appropriate.

1. Disproportionality Analysis Approaches from Spontaneous Adverse Event Reporting

Description:

One key goal of OMOP is to assess the feasibility and utility of observational data in identifying associations between drugs and conditions. As defined in the OMOP design, focus is placed on two primary types of conditions: 'health outcomes of interest' are those specifically-defined conditions that are the focus of ongoing surveillance for all medicines, and 'non-specified conditions' that may have other unanticipated relationships with drugs. For each type, OMOP intends to design and perform objective tests to evaluate the performance of alternative identification methods. Several alternative methods and approaches were identified through global introspection, literature review and patent review, but the working group acknowledges the list is not exhaustive of all potential approaches. Further work is recommended to perform a systematic methods review to ensure that best practices from all domains and disciplines are fully considered for their utility in supporting observational pharmacovigilance analyses.

This section outlines one recognized approach to identifying drug-condition associations with observational data. Within this section, we refer to 'disproportionality analysis' as a class of methods originally developed for use on spontaneous adverse event reporting systems to assess how much the observed frequency of a given drug-combination pair deviates from the expected frequency. The methods assess measures of reporting disproportionality rather than measure of different in incidence rates in the context where no true denominator is known. Disproportionality analysis algorithms (and software) for application to spontaneous reporting databases have been available for almost 10 years. A number of applications have been described in the literature. The recent survey paper by Almenoff et al discusses considerations appropriate for the use of these methods, and a number of issues that arise in their application. Spontaneous report databases consist of reports from physicians and others identifying adverse events reported by patients, along with drugs that the patient was taking concurrently or recently. The objective of the algorithms is to identify drug-event combinations that are reported more frequently than might be expected if there were no association between the mention of a drug and a mention of an adverse event in the reports, hence the term 'disproportionality'.

Some have suggested that these methods could be transported from the spontaneous world and directly applied to observational databases, while others have posited that 'denominator-less' methods are less appropriate in observational data where true denominators can be easily estimated. There have been few published systematic investigations of the value of applying disproportionality detection methods to large observational databases.

Rather than subjectively deciding the merits of this theoretical debate, it is an explicit goal of OMOP to perform empirical studies that can evaluate whether disproportionality analysis methods have utility relative to other recognized approaches in identifying drug-condition associations.

Constructing a drug-condition pair database from observational data

Implicit in the disproportionality analysis methods is the premise that the analysis be performed on a database of drug-condition pairs. In the spontaneous adverse event systems, this assumption is automatically satisfied: each adverse event case report represents a set of drugs and a set of adverse events which can be 'paired' and aggregated for analysis.

Most disproportionality analysis techniques developed for use on spontaneous adverse event reporting databases are effectively variants of frequency analysis on 2x2 contingency tables. Most approaches attempt to identify the ratio of observed to expected frequencies. They differ only in how they define 'expected' as the background inherent to the data. Frequentist approaches have an advantage over Bayesian methods of being much easier to compute, but there is no definitive method in terms of ability to identify drug-condition pairs requiring further evaluation. For the purposes of comparing different approaches, we will characterize our contingency tables as conforming to the following schema:

	Target Drug	All Other Drugs	
Target Condition	a	b	a+b
All Other Conditions	c	d	c+d
	a+c	b+d	n = a + b + c + d

In spontaneous data, 'a' would represent the number of adverse event cases where the target drug was co-reported with the target condition, 'b' is the number of pairs where the target condition was reported without the target drug, etc.

The concept can be applied to observational data, with the general intent being similar, though the level of disproportionality is in terms of explicit temporal relations within the data instead of inferred causal relations from AE reports. In observational data, however, some rules have to be established in order to extract drug-condition pairs. Since the data natively contains persons with drug utilization and condition incidence, it is necessary to establish practice for determining which drug uses are associations with which conditions for each person.

The method can be extended to observational data by constructing an analogous set of drug-outcome pairs from the claims or EHR. Specifically, drug-outcome pairs can be identified across all persons as those outcomes that are incident during drug exposure (DRUG_START_DATE < OUTCOME_START_DATE < DRUG_END_DATE). [Alternatively, a 'surveillance window' can be extended beyond the end of drug exposure to capture outcomes that are experienced near with end of drug use with some degree of tolerance (e.g. DRUG_END_DATE + 30d)].

All resulting drug-condition pairs are then extracted into a dataset, which can be further aggregated to calculate counts by drug and condition. Disproportionality statistics can be subsequently calculated. The drug-condition pairs can then be represented in a 2x2 contingency table for each specific drug-outcome pair, which can be used to calculate the particular metrics.

Frequentist approaches

Proportional reporting rate (PRR): PRR is the calculation of the proportions of specified reactions or groups of reactions for drugs of interest where the comparator is all other drugs in the database. If the drug and condition are independent, the expected value of PRR should be 1; PRR>1 indicate a greater than expected frequency of the report in the dataset. One drawback is that PRR may be undefined if no cases of a condition occur outside the target drug population (b=0). It may also provide unreasonably high values for rare events (where a, b, and c are small).

$$PRR = a/(a+c) / b/(b+d) = a*(b+d) / b*(a+c)$$

We calculate the confidence interval based on the standard epidemiology 2x2 table, as defined by Rothman et al.

$$PRR\ 95\%CI = e^{\ln(PRR) +/- 1.96*\sqrt{[1/a-1/(a+c)+1/b-1/(b+d)]}}$$

Reporting Odds Ratio (ROR): ROR is another metric for disproportionality, where the expected frequency of a given drug-condition pair is not dependent upon the observed frequency drug-condition pair. ROR has been proposed to mitigate the issue of non-selective under-reporting of specific drugs or events. Like PRR, ROR is an easily interpreted metric with an expected value of 1, such that ROR>1 may be indicative of signals.

One issue to consider with ROR is that events specifically related to the drug may not easily detected, because the metric can be infinite when there are zero cases in the comparator group. This situation may occur in particularly noteworthy instances of rare conditions that should require further evaluation. Rothman proposes further refinement by viewing the database as a case-control study, where one excludes from the control series those events that are suspected to be associated with the drug. Regardless of the use of ROR, the case-control paradigm could be helpful in examination of drug-condition pairs.

$$ROR = (a/c) / (b/d) = a*d / b*c$$

Chi-Square (χ^2): A measure of expectation of marginals may be calculated using a chi-squared test on one degree of freedom. χ^2 performs well if the expected cell count is not small[15]. If necessary as adjustment for small cell counts, Yates' correction can be used.

χ^2 with Yates' correction can be expressed as

$$\chi^2 = \sum [(|O-E|)^2 - ½ / E]$$

Where the summation is over all four cells of the contingency table, O is the observed frequency and E is the expected frequency of the reports.

Chi-square is one measure of statistical association commonly applied to 2x2 contingency tables. Here we use chi-square with 1 degree of freedom and Yates' correction, using the simplified formula provided by Sheskin et al.

$$\chi^2 = [n*(|a*d-b*c|-n/2)^2] / [(a+c)*(b+d)*(a+b)*(c+d)]$$

Bayesian approaches

Multi-item Gamma Poisson shrinker (MGPS): Multi-Item Gamma Poisson Shrinker is a data mining algorithm that computes the empiric Bayes geometric mean (EBGM) and corresponding 2-sided 90% confidence interval (EB05 < EB95) for each observed drug-condition pair in a reporting database.

EBGM values represent relative reporting rates (after Bayesian smoothing or "shrinkage") for drug-condition pairs in a given database. An EBGM of 5 means that a drug-condition pair has been reported 5 times as frequently as would be expected if reports involving the drug and reports of the condition were independent (i.e., no reporting association). A high relative reporting rate does not necessarily indicate a high incidence of the condition or suggest a causal relationship between the drug and the condition. The FDA has used EB05 ≥2 as a threshold for signal detection in spontaneous databases. This threshold ensures with a high degree of confidence that regardless of the number of reports, a particular drug-condition combination is being reported at least twice as often as would be expected if there were no association between the drug and the condition[10].

The "standard" stratification variables used in MGPS on spontaneous adverse event reporting data are age, sex, and year of report. Stratification is one approach to minimize the detection of apparent drug-condition associations that are actually due to independent relationships that may exist between a drug and a strata variable and a condition and the same strata variable. Analyses are stratified by year of report to reduce the chance of detecting signals that might arise due to "trendiness" in the reporting of specific drugs and/or events.

Bayesian confidence propagation neural network (BCPNN): Bayesian confidence propagation neural network (BCPNN) uses Bayesian statistics implemented in a neural network architecture to analyze all reported drug-condition combinations to identify unexpectedly strong relationships. This method is now in routine use for signal detection for WHO data, and is being explored within observational claims data.

The strength of the dependency between a drug and event is defined by a logarithm measure called the information component. The IC can be seen as the logarithm of the ratio of the observed rate of the drug-condition pair to the expected rate, under the null hypothesis of no association between the two[18]:

$IC = \log_2 p(x,y) / p(x)* p(y), = \log_2 p(y|x) / p(y),$

Where

a. p(x) is probability of a drug being listed on a case report [(a+c)/n]
b. p(y) is the probably of an event being listed in a case report [(a+b)/n]
c. p(x,y) is probability of a drug-event pair being listed [a/n]
d. p(y|x) is the conditional probability of y given x

Alternative approaches

Various literature references mention other possible metrics, including relative risk (RRR), Yule's Q, Kullbeck-Leibler disparity measure, and the empirical Bayes method by Gould.

Potential thresholds:

Thresholds have been commonly applied in spontaneous data to each of the metrics outlined above. However, the thresholds in use did not arise from empirical justification, but arbitrary demarcation lines used as a preliminary means of prioritization. As we perform the empirical studies in OMOP, it will be important to reference these thresholds to assess how well they perform in observational data, but also consider alternative thresholds for the same metrics. Example thresholds include:

Evans' criteria applied to spontaneous data: PRR>2, CHISQ>3, N>4
PRR>2 (or some other number > 1)
PRR_LB95>2 (or some other number > 1)
ROR>2 (or some other number > 1)
EBGM>2 (or some other number > 1)
EB05>2 (or some other number > 1)
IC>1

Expected Strengths:

Disproportionality analysis is fairly well understood within the pharmacovigilance community, and commonly applied to spontaneous adverse event reporting databases. Frequentist approaches are relatively efficient computationally, as all potential drug-condition pairs can be pre-processed at once and safety scientists can simply review the subset of pairs of interest.

Expected Limitations:

As defined here, disproportionality analysis methods do not make any adjustments for lengths of exposure or the reasons for exposure. The methods do not preclude such adjustments, but these modifications would need to be clearly defined prior to use. As a result without modification,

the method may perform differently for treatments requiring chronic use vs. acute use, or for outcomes that are incident immediately following exposure vs. after long-term use. As a 'denominatorless' approach, the comparator of 'all other drugs' may not provide the precision to identify unique relationships experienced within indicated populations at varying degrees amongst alternative treatment. These metrics does not adjust for other confounding factors related to exposure and outcome, so identified associations may be associated with the factors other than a causal drug-outcome relationship. Bayesian approaches may be computationally difficult to execute over extremely large data sources.

References:

Almenoff J, Tonning JM, Gould AL, et al., "Perspectives on the use of data mining in pharmacovigilance," Drug Safety, 2005: 28(11), 981-1007.

Gogolak VV. The effect of backgrounds in safety analysis: the impact of comparison cases on what you see. Pharmacoepidemiol Drug Saf 2003; 12: 249-52

Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001; 10: 483-6

van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002; 11: 3-10

Kaufman DW, Rosenberg L, Mitchell AA. Signal generation and clarification: use of case-control data. Pharmacoepidemiol Drug Saf 2001; 10: 197-203

DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999; 53 (3): 177-202

DuMouchel W, Pregibon D. Empirical Bayes screening for multi-item associations. Proc KDD 2001, San Diego, CA, ACM Press.

Bate A, Lindquist M, Edwards IR et al. A Data Mining Approach for Signal Detection and Analysis. Drug Saf 2002; 25 (6): 393-397

Lilienfeld, DE. A challenge to the data miners. Pharmacoepidemiol Drug Saf 2004; 13: 881-884

US FDA. Guidance for industry: good pharmacovigilance practices

and pharmacoepidemiologic assessment. US Food and Drug Administration Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, March 2005 [online]. Available from URL: http://www.fda.gov/cder/ guidance/6359OCC.htm#_Toc48124197 [Accessed 2005 Nov 29]

Kubota K, Koide D, Hirai T. Comparison of data mining methodologies using Japanese spontaneous reports. Pharmacoepidemiol Drug Saf 2004; 13: 387-94

Chan KA, Hauben M. Signal detection in pharmacovigilance: empirical evaluation of data mining tools. Pharmacoepidemiol Drug Saf 2005; 14: 597-99

Rothman KJ, Lanes S, Sacks ST. The reporting odds ratio and its advantages over the proportional reporting ratio. Pharmacoepidemiol Drug Saf 2004; 13: 519-23

Hauben M, Reich L. Potential Utility of Data-Mining Algorithms for Early Detection of Potentially Fatal/Disabling Adverse Drug Reactions: A Retrospective Evaluation. J Clin Pharmacol. 2005; 45 (4): 378-84.

Waller P, Heeley E, Moseley J. Impact analysis of signals detected from spontaneous adverse drug reaction Drug Saf 2005; 28 (10): 843-50

Almenoff JS, DuMouchel W, Kindman A, et al. Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting. Pharmacoepidemiol Drug Saf 2003; 12 (6): 517-21

Sheskin DJ. Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press, 2004.

Evans SJW, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse event reaction reports. Pharmacoepidemiol Drug Saf 2001; 10: 483-6.

Rothman K. Epidemiology: an introduction. New York, NY: Oxford University Press, 2002.

2. Approaches Based on Selecting Populations Taking Drugs

Description:

One key goal of OMOP is to assess the feasibility and utility of observational data in identifying associations between drugs and conditions. As defined in the OMOP design, focus is placed on two primary types of conditions: 'health outcomes of interest' are those specifically-defined conditions that are the focus of ongoing surveillance for all medicines, and 'non-specified conditions' that may have other unanticipated relationships with drugs. For each type, OMOP intends to design and perform objective tests to evaluate the performance of alternative identification methods. Several alternative methods and approaches were identified through global introspection, literature review and patent review, but the working group acknowledges the list is not exhaustive of all potential approaches. Further work is recommended to perform a systematic methods review to ensure that best practices from all domains and disciplines are fully considered for their utility in supporting observational pharmacovigilance analyses.

This section outlines one recognized approach to identifying drug-condition associations with observational data. The basic concept is to construct populations of persons taking drugs and assess the occurrence of conditions across those populations. Drug-condition associations can be identified when a condition occurs within one exposed population more often than some logical comparator (whether it is as compared to the overall population rate, the rate within another comparator population of interest, or the rate within the pre-exposed period of the target population). The concept is commonly referred to in the field of epidemiology as a 'cohort design', although there are many variants in the design that are discussed below. The reader is referred to the literature references provided for further illustration of how related approaches have been previously applied in the literature, and are encouraged to consider how such techniques could be modified for use in systematically exploring HOIs and non-specified conditions.

Starting point for consideration:

Assume the goal of the analysis is to identify any potential associations between any conditions and a particular target drug. As one basic approach, an analyst could select all persons in an observational data who have ever taken the target drug of interest. With this 'exposure cohort' constructed, various rates of occurrence of conditions could be calculated on the basis of when the condition occurred relative to the drug utilization. Among this prevalent cohort will be new users, which can also be used to define an inception cohort. For example, incidence rates can be estimated by calculating the number of occurrences of the condition, divided by the person-time at risk. One estimate of person-time at risk could be the person-time of exposure. Then, rates prior to exposure could be compared to rates of occurrence during exposure. Alternatively, rates during exposure could be compared to the known rate of occurrence in the overall population.

Additionally, the analyst could construct a 'comparison cohort' of all persons who have ever taken some comparator drug of interest. In this manner, rates of condition occurrence during target exposure can be compared to rates within the comparison cohort. Unadjusted metrics can be readily calculated for these comparisons using a common set of assumptions around the definitions of person-time and condition incidence.

Variants in design:

While the approach outlined above may serve as a starting point, several considerations must be made in order to establish a systematic solution that can be applied to all drugs and all outcomes.

Comparator selection:

Relative metrics that facilitate comparison require the selection of some comparator. The comparator selection remains an explicit decision by the analyst within this approach, though consideration can be made as to whether alternative comparators can be automatically selected without user intervention. Further research is needed to determine which comparisons are most appropriate for which circumstances. It is possible that multiple comparator groups may be preferable.

The most common comparator cohort design involves selecting some drug comparator group. In this regard, the population taking the target drug may be compared to some population of persons taking a comparator drug or set of drugs in the same therapeutic class, or a drug(s) for the same indicated condition. Incidence rates during exposure can then be compared to identify whether one cohort experiences a higher occurrence of the condition.

An argument in favor of a comparator drug is that the decision has been made to treat the patient, eliminating the potentially extreme version of confounding by indication that could result from using a comparison group that has been unexposed to any drug.

An alternative comparator group can be defined by a population of people with a specific condition of interest. For example, the target cohort may be compared to the population of persons with the indicated condition of the target drug. This comparator group may more closely resemble the idea of an 'unexposed' population, but consideration is required to determine if the two groups can be meaningfully compared against one another.

Within-cohort metrics do not require a distinct cohort comparator selection. Instead, persons within the target cohort act as self-controlled with periods of time prior to exposure compared to periods of exposure. Incidence rates for pre- and post-exposure can be estimated, and incidence rate ratio of post-/pre-exposure can be used to identify those outcomes where the exposed rate is highest relative to pre-exposure rates.

Cohort-overall metrics also do not require distinct cohort comparator selection. Instead, incidence rates post-exposure within the target cohort of interest can be compared to background overall incidence rate, estimated across the entire database. In this regard, an incidence rate ratio post-exposure/overall can be used to identify those outcomes that have higher incidence in cohort relative to overall expectation.

Cohort Matching:

An unmatched cohort design provides crude comparison of the target population and another population of interest (whether it is a drug or condition comparator, or the overall population). As a basic example, a cohort of all patients who ever took the target drug could be compared to a cohort of all patients who ever took a comparator drug. Rates of occurrence of specific conditions could then be calculated, representing the true observed rate for that population. However, consideration must be made when comparing rates across populations; other factors that distinguish the two populations may account for observed differences in the rates (referred to as confounding).

One potential alternative to address issues of confounding when comparing incidence rates across populations is matching. Comparisons are then made across the matched subset of the overall population of interest. One common approach applied in pharmacoepidemiology evaluation studies is to perform direct matching on a subset of covariates of interest. In particular, matching by age, sex, and time of exposure is commonly considered. It is unclear the extent to which a similar approach could be applied to identification within observational databases. One advantage of employing matching is that it is one approach to reducing associations (or lack of associations) due to confounding. Theoretical analysis has shown that matching in cohort studies completely controls for any potential confounding by the matching factors without requiring any special statistical methods. Note that you only control for confounding by the matching factors. On the other hand, one avoids having to make any assumptions about the shape of the relationship between the matching factor and the outcome. E.g., if matching on age, it doesn't matter whether the association between age and outcome is linear, quadratic, or more complicated. One disadvantage is that outcomes observed within the unmatched subset will not be considered in the analysis. Any match-based cohort approach necessarily discards patients taking the drug who did not satisfy the matching criteria from the analysis. So while all patients taking the drug may experience a condition of interest, match-based approaches only allow the analysis of some subset of that total exposed population. If only those subjects that matched are included in the analysis as a criterion of restriction, the subset not included in the analysis may be the population that demonstrates an association with the outcome, potentially resulting in false negative findings. There can be a loss of statistical power. Also, matching may be less computationally efficient than unmatched designs, and thus more difficult to execute within the systems infrastructure as a systematic solution to satisfy all drugs and all conditions. This presents a tradeoff of potentially increased comparability produced by matching with potential reduction in generalizability. Further consideration is required to determine whether one approach or both is required to assess safety.

Propensity score matched cohort design:

Propensity score matching is an approach being used increasingly for conducting evaluation studies within observational data. In addition, some, such as i3 Aperio, have considered applying this approach for identification. Note, this area is one of continuing controversy and research, with little consensus around best practice in design or interpretation.

The concept of propensity scoring is to construct one scalar value based on a set of covariates that can then be used to identify similar patients across cohorts. The propensity score is the conditional probability of selection of a particular treatment given a set of observed covariates. The score can then be used to select a comparator cohort with the same probability of receiving an intervention.

The theoretical goal of propensity scoring is to create cohorts that are balanced with respect to measured covariates. Some have expressed this as an attempt to approximate randomization amongst two groups. Just as with a randomized clinical trial where treatment can be randomly allocated at each site, the goal is to ensure in the observational study that every patient has a similar chance of being "allocated" to both treatment alternatives. In theory, with randomization, there would be two groups well balanced on all factors. Within this approach, the analyst must define the target cohort of interest (patients taking the target drug), a comparator cohort of interest (patients taking a comparator drug), and the set of covariates that require balancing between exposure groups. Some argue that in an ideal situation, you would build your propensity score on all covariates prior to exposure (all prior conditions and all prior drugs), but you may not be afforded that many variables. Then, the issue becomes which variables should be included. In addition to patient-level characteristics, provider-level effects should be introduced when possible. Others have argued that non-confounding correlates of exposure should be omitted, because their inclusion erodes precision and does not enhance validity.

An advantage of this approach is that it explicitly attempts to control for potential differences in exposed and unexposed by more clearly assuring that the members of the comparator group selected are more like the exposed patients with respect to reasons they were selected for treatment (confounding by indication). In an evaluation study, propensity score matching can also be used to reduce the number of covariates necessary in a multivariate statistical model.

The approach has several challenges in applying to identification within observational data. Propensity score matching can be computationally expensive and difficult to execute within a systems infrastructure. Currently, there is no agreed best practice to systematically select covariates for the propensity model. As with other approaches, because the current approach requires subjective decisions about which covariates to include in the model, it is possible that misspecification can introduce bias into the analysis if important confounders that are related both to the exposure and the outcome are not included in the model. When screening for non-specified conditions, the outcome is not defined (instead, you want to explore all possible outcomes), it may not be possible to construct one propensity score across the cohorts that can be used for all outcomes (since covariates related to outcome will differ for each non-specified condition). So, the approach either requires accepting the potential for introducing bias, or requires running a propensity score matched cohort design with a unique set of covariates for

each outcome. A unique set of covariates may be possible to identify for each of the Health Outcomes of Interest, but it is unclear how one could identify covariates for all other non-specified conditions. Others have argued that there if balanced is achieved across all prevalent covariates, no consideration of the outcome needs to be made. Further consideration is required to determine the feasibility of such an approach across all observational data sources.

Restriction:

As defined above, a basic approach is to compare the cohort of all patients who ever took the target drug with the cohort of all patients who ever took a comparator drug. Another potential issue is mixing: patients could have taken both the target drug and the comparator drug. Similarly, patients may not be comparable because they were ineligible to receive one drug (e.g. contraindications). One potential approach to address these issues is restriction: exclusion criteria can be imposed on the populations of interest. Restriction is commonly applied in randomized clinical trials and observational evaluation studies, though best practices for a systematic application of restriction is observational screening are unclear. In particular, there is no agreed systematic process for defining restriction criteria that can be imposed on all drugs to explore all conditions. As with matching, restriction presents a tradeoff with increased comparability at the expense of decreased generalizability. Further consideration is required to determine whether restriction is a necessary component to the identification process and to evaluate what incremental improvements are gained.

Censoring:

In the base case example illustrated above, all conditions that occurred during exposure may be captured as part of the rate estimation, with all time at risk used as a denominator. Some epidemiologic studies may alternatively invoke censoring to restrict the counts of conditions to only the 'first' occurrences. This is common for studies exploring chronic conditions (e.g. diabetes), or when prior onset of a condition may be confounded with future onset of condition (e.g. acute myocardial infarction). In these instances, person-time can be additionally censored to include only the period of exposure up until the point of first condition occurrence; subsequent exposure time is discarded for incidence rate calculations. Further consideration is required to determine whether censoring should be considered as part of an identification approach; while specific rules may be feasible for each Health Outcome of Interest, it is uncertain what constitutes best practice when systematically exploring all non-specified conditions.

Periods of observation:

A key challenge for observational screening is defining a common person-time at risk to be used across all conditions. Person-time at risk can be estimated as the variable period of exposure. Additionally, a 'surveillance window' can be added to end of exposure to incorporate uncertainty in actual drug cessation and half-life of medicine. For example, person-time at risk can be estimated as (drug use end – drug use start) + 30d. The length of the window might be based on PK considerations, e.g., 5 half-lives. While this approach may be amendable for many conditions, it may be insufficient for conditions with long latency (e.g. cancers).

Alternatively, a constant period following the start of exposure (e.g., 30 days from first drug use) can be applied. Short surveillance window periods (7 day) may be appropriate to capture relationships that may occur immediately following exposure (e.g., hypersensitivity reaction), while longer windows (e.g. 1 yr) following exposure may be used. Note that 'risk windows' that are related only to time of exposure start- and not the total length of exposure- may be a weaker sense of temporal association than those 'person-time exposure' periods tied directly to both exposure start and exposure duration. Additionally, consideration can be made for how risk can change over time after initiation of drug use, as discussed by Ray et al.

Metrics:

Within these 'cohort'-based approaches, a variety of alternative metrics can be considered. Below are illustrations of relative metrics, as commonly applied in epidemiologic studies. Note that, instead of relative metrics that use ratios, related metrics can be constructed that explore risk difference, attribute risk, or incremental risk.

Person-time metrics, like incidence rates and incidence rate ratios, or hazard ratios can be calculated if the person-time at risk can be reasonably estimated for each subject. Here, an incidence rate is the number of incidences of the outcome, divided by the total person-time at risk. An incidence rate ratio is one comparison of two incidence rates (e.g., IR target cohort / IR comparator cohort)

	Number of Condition occurrences	Total person-time exposed	Crude Incidence Rate
Target Drug	a	ta	a/ta
Comparator Drug	b	tb	c/tb

Incidence Rate Ratio = (a/ta) / (b/tb)

Person metrics, like relative risk, can be calculated based on a 2x2 contingency table, where each cell represents the number of persons with or without exposure and with or without outcome:

	Target Drug	Comparator drug	Outcome total
Has Condition	a	b	a + b
No condition	c	d	c + d
Drug total	a + c	b + d	n=a+b+c+d

RR = [a/(a+c)] / [b/(b+d)]

We calculate the confidence interval based on the standard epidemiology 2x2 table, as defined by Rothman et al.

RR 95%CI = $e^{\ln(RR) +/- 1.96*\sqrt{[1/a-1/(a+c)+1/b-1/(b+d)]}}$

Potential thresholds:

For each method, one could construct alternative observational screening thresholds to apply. As described in the OMOP design, screening for Health Outcomes of Interest requires a more conservative approach to ongoing surveillance where the screening metrics are primarily used for prioritization, while identification of non-specified outcomes may more regularly use thresholds to restrict the set of drug-condition candidates that warrant further review.

Alternative approaches could be based on confidence intervals (where drug-condition pairs are identified on the basis of the lower bound), or based on the point estimate of the metric (where the magnitude of the potential effect is considered). Examples are provided below:

Person-time metrics:
IRR_LB95> 1
IRR > 2 (or some other number > 1)

Person metrics:
RR_LB95>1 (or some other number > 1)
RR > 2 (or some other number > 1)

Adjustment:

In addition to the unadjusted metrics illustrated above, multivariate adjustments can be performed in attempt to address confounding. For example, a Cox proportional hazards model can be used to adjust for other covariates, like age and sex. Poisson regression can achieve the same adjustment. In the cited literature, the maxSPRT also allows for multivariate adjustment by age, sex and health plan. Other alternatives include approaches for stratification and direct standardization. Adjustment approaches offer the potential to reduce confounding and minimize false positive findings, but can also introduce bias by minimizing associations occurring in subpopulations of interest. Additionally, adjustment methods may be more computationally intensive than unadjusted methods. Further consideration is required to assess the feasibility of adjusted metrics, and to evaluate the tradeoff between unadjusted vs. adjusted approaches.

Multiplicity

When performing the analysis to identify associations across all non-specified conditions, some have argued that this approach amounts to performing multiple statistical tests (one test for each condition) and that adjustments for multiplicity should be considered. Approaches, such as Bonferroni adjustment and false discovery rates, have been offered. Others have argued that multiplicity adjustment is unnecessary in this exploratory phase provided an evaluation phase is anticipated, because it may increase the required information to observe a true association. Further consideration is required to determine whether adjustment for multiplicity is appropriate and necessary for inclusion in any identification methods.

Recurrent Analysis of Accumulating Data

It is reasonably anticipated that observational screening will be performed on a continuing basis during a product lifecycle. As such, analyses will be repeated over time as new data becomes available. Similar to the issue of multiplicity, some have agued that continuing reuse of data requires statistical adjustment to reduce false positive findings. One method, the maxSPRT, performs an explicit statistical adjustment to account for accumulating data. Other methods in other domains make no such adjustment, and some argue that the tradeoff between reducing false positive by potentially increasing false negatives in unwarranted. Further consideration is required to determine whether adjustment for recurrent analysis is appropriate and feasible for inclusion in any identification methods.

Expected Strengths:

Cohort designs are generally the strongest non-randomized design considered in most hierarchies of information. Approaches assessing populations taking drugs (i.e., cohort designs) are intuitive and can capture the full amount of exposure within the database. Metrics can be reasonably interpreted and compared. One cohort construction can facilitate exploration of a given drug across all outcomes (while a traditional case-control design can explore a given outcome against all drugs),

Expected Limitations:

In general, it is expected that person-time metrics should be more precise estimates that person metrics, because person-time at risk and length of exposure is a more granular denominator than number of persons. But since the data recorded may not reflect true exposure (but only a surrogate based on prescriptions written or filled, etc.), use of person-time may introduce some residual bias.

One consideration to selection of the cohorts is that new user cohorts may be small initially following product launch, thus requiring a prevalent cohort approach to maximize sample. Since the population of drug users evolves over time with physician experience, a cohort spanning the full time from market entry may not be wholly consistent.

One of the primary challenges with these approaches is the degree to which users have to make design decisions. Without best practices established, subjective decisions have to be made for all the variants explored above. There are no right answers to decisions around comparator, matching, censoring, so one set of decisions may be more appropriate for certain circumstances than others. Sensitivity analysis would be one way to assess the impact of these decisions.

3. Approaches Based on Selecting Populations with Condition

Description:
One goal of OMOP is to assess the feasibility and utility of observational data in identifying associations between drugs and conditions. As defined in the OMOP design, focus is placed on two primary types of conditions: 'health outcomes of interest' are those specifically-defined conditions that are the focus of ongoing surveillance for all medicines, and 'non-specified conditions' that may have other unanticipated relationships with drugs. For each type, OMOP intends to design and perform objective tests to evaluate the performance of alternative identification methods. Several alternative methods and approaches were identified through global introspection, literature review and patent review, but the working group acknowledges the list is not exhaustive of all potential approaches. Further work is recommended to perform a systematic methods review to ensure that best practices from all domains and disciplines are fully considered for their utility in supporting observational pharmacovigilance analyses.

This section outlines one recognized approach to identifying drug-condition associations with observational data. The basic concept is to construct populations of persons with a particular condition and assess the association of the condition across drug exposure. Drug-condition associations can be identified when a drug exposure occurs within the population with the condition of interest more often than some logical comparator (whether it is as compared to the overall population rate, the rate within another comparator population of interest, or the rate within the pre-condition period of the target population). The main concept reflected in this approach is that of a case-control study (where cases are compared to non-cases) or a self-controlled case series (where the time prior to becoming a case is used in lieu of a comparator). There are many variants in the design that are discussed below. The reader is referred to the literature references provided for further illustration of how related approaches have been previously applied in the literature, and are encouraged to consider how such techniques could be modified for use in systematically exploring HOIs and non-specified conditions.

Starting point for consideration:

In a case-control design, a case is defined as a person who experienced a condition of interest. For the purposes of OMOP, the condition of interest could be either an HOI with its associated definition, or a non-specified condition, defined by a single code (ICD-9 or medDRA) in the database. All exposures prior to condition onset are then assessed to identify associations between case status and exposure.

Assume the goal of the analysis is to identify any potential associations between a particular condition and any drug. As one basic approach, an analyst could select all persons in an observational database who has ever experienced the condition. With this group of cases

identified, various measure of association with drug exposures could be calculated on the basis of when the drug utilization occurred relative to the condition appearing. This will necessitate decisions about time windows for that exposure relative to the condition. For example, the prevalence of exposure to each drug among those with the condition can be estimated by calculating the number of people on each drug within the designated time window at or before the occurrence of the condition (can be expressed as a proportion) and compare the exposure proportion to that of the remainder of the population without the condition. Several options for selection of controls are possible, depending on the HOI-pair of interest. Some choices include comparing cases to an appropriate 'general population', to other comparator conditions of interest, or to a class of conditions treated by the same types of drugs. In a case-control sampling design, some options include self-controlled case series designs, sampling controls from source population taking all other drugs, or controls with HOIs that are not related to drug-exposure. Unadjusted metrics can be readily calculated for these comparisons using a common set of assumptions around the definitions of condition and exposure windows. An alternative to the standard case-control approach above would be to estimate an OR for all drugs/ drug classes for each condition in turn, vs. all other conditions using standard adjustment procedures.

Similarly, one could compare the proportion of use of the drug prior to and at the time of the event (case series designs- discussed below). Then, rates prior to exposure could be compared to rates of occurrence during exposure.

Variants in design:

Case-based approaches are useful when HOI is rare and/or acute and with varying exposure scenarios, including discontinuation period. These methods can be used when there is a long latency interval between exposure and HOI occurrence allowing consideration of risk during continued exposure and after discontinuation. A series of cases with the HOI is selected along with a representative sample from the base population from which cases were identified. Risk difference assessed by estimating the denominators of the proportions of cases among exposed and comparable unexposed subjects by means of a representative sample from the base study population. Analytic methods include relative risk (RR), hazard ratios, excess risk (ER), rate difference, incidence density function, hazard functions used to calculate ER and RR.

Self-controlled case series design is useful when there are multiple exposure risk periods and when subjects experience more than one HOI. This may be particularly useful when exposure effects are short term. Subjects are used as their own controls and implicitly controls for all fixed confounders. This requires that exposure probability is not affected by occurrence of an HOI and variability in time or age of HOI. This provides consistent estimates of relative incidence, but cannot estimate absolute incidence only relative incidence. It requires variability in the time of the event; if all events were to happen at exactly the same time then the method would fail (Whitaker et al). Analyses can include conditional regression to estimate the risk of the condition by comparing incidence during periods of exposure to particular drugs and compared with incidence during periods when people were not exposed to drug. The main advantage of this method is the elimination of inter-individual confounding, such as differences in comorbidity. A limitation of this approach, however, is that any change in disease severity

may affect attitudes to prescribing of drugs considered to increase the risk of event. When compared to case-control analyses (Tata et al), the case series analysis may minimize confounding by severity of the underlying disease or other unmeasured risk factors while residual confounding may inflate case–control results. A review of the variants of this design can be found in Whitaker et al. which begin with the origins in a paper by Maclure on the so-called "case-crossover" design.

Comparator selection:

Relative metrics that facilitate comparison require the selection of some comparator. The comparator selection remains an explicit decision by the analysis within this approach, though consideration can be made as to whether alternative comparators can be automatically executed without user intervention. Further research is needed to determine which comparisons are most appropriate for which circumstances.

The most common comparator involves selection of non-condition affected subjects. In this regard, the population with the condition may be compared to some population without the condition for the frequency of taking a comparator drug or set of drugs in the same therapeutic class, or a drug(s) for the same indicated condition. For case-series methods, the comparison is across time periods within the same person.

Matching

An unmatched design provides crude comparison of the population with a condition and another population of interest (whether it is a specific non-condition comparator or the overall population). As basic example, a comparator sample could be compared to a cohort of all patients who never experienced the condition of interest. Rates of drug exposure could then be calculated.

Matching is often used to control for a small number of potentially confounding factors. You may run into trouble if you try to match on too many factors – run out of matches. You should discourage matching on too many factors.

Comparisons are then made across the matched subset of the overall population of interest. One common approach applied in pharmacoepidemiology evaluation studies to perform direct matching on a subset of covariates of interest. In particular, matching by age, sex and plan, and time of exposure is commonly considered. It is unclear the extent to which a similar approach could be applied to identification within observational databases. In case-control studies, matching reduce associations due to confounding by may also introduce bias. Any match-based approach necessarily discards patients who did not satisfy the matching criteria from the analysis. Matching may be less computationally efficient than unmatched designs. This presents a tradeoff of potentially increased comparability produced by matching with potential reduction in generalizability. Further consideration is required to determine whether one approach or both is required to assess safety.

The use of **propensity scores** to adjust for measured confounding factors has become increasingly popular in cohort studies used for evaluation but relatively little has been explored

in case-control methods (and none for identification). The reader is referred to the summary publication by Mansson et al. to review the theory on the estimation and use of propensity scores in case-control studies and the results of simulation studies. The application of propensity scores is less obvious in case-control studies and Mansson's work revealed that there is an artifactual effect modification of the odds ratio by level of propensity score the magnitude of which decreases as the sample size increases and it is possible that the estimated propensity scores can fail to adjust fully for measured confounding factors as sample size increases.

Restriction:

Applying exclusion criteria can be imposed on the populations of interest to assure that the study is based on the selection of a particular type of event occurrence (e.g., initial occurrence, multiple occurrences over time in the same person). Restriction is commonly applied in randomized clinical trials and observational evaluation studies, though its role is observational screening is unclear. In particular, there is no agreed systematic process for defining restriction criteria that can be imposed on all potential conditions. As with matching, restriction presents a tradeoff with increased comparability at the expense of decreased generalizability. Further consideration is required to determine whether restriction is a necessary component to the identification process and to evaluate what incremental improvements are gained.

Censoring:

Some epidemiologic studies may invoke censoring to restrict the counts of conditions to only the 'first' occurrences. This is common for studies exploring chronic conditions (e.g. diabetes), or when prior onset of a condition may be confounded with future onset of condition (e.g. acute myocardial infarction). Further consideration is required to determine whether censoring should be considered as part of an identification approach; while specific rules may be feasible for each Health Outcome of Interest, it is uncertain what constitutes best practice when systematically exploring all non-specified conditions.

Periods of observation:

Exposure time windows, mentioned above, are the critical determinants of some of the case series designs. As risk is estimated at the variable period of exposure, a 'surveillance window' can be defined to incorporate uncertainty in actual drug cessation and half-life of medicine. For example, exposure time at risk can be estimated as (drug use end – drug use start) + 30d.

Alternatively, a constant period following the start of exposure (e.g. 30 days from first drug use) can be applied. Short surveillance window periods (7 day) may be appropriate to capture relationships that may occur immediately following exposure (e.g. hypersensitivity reaction), while longer windows (e.g. 1 yr) following exposure may be used. Note that 'risk windows' that are related only to time of exposure start- and not the total length of exposure- may be a weaker sense of temporal association than those 'person-time exposure' periods tied directly to both exposure start and exposure duration.

Metrics:

There are a number of different metrics that can be derived from these designs. While they use the same data (as in 2x2 contingency table), the metrics provide alternative ways of interpreting the results. Application of an odds ratio screening method for signal detection has been described in The Slone Epidemiology Center's case-control surveillance databases (Kaufman et al. 2001). The design provides an estimate of relative risk called the Odds Ratio (OR), which is calculated as follows:

Table. Simple 2x2 Table for Calculation of Odds Ratio

	No. of patients exposed to drug X within exposure window	No. of patients NOT exposed to drug X within exposure window	Total patients
Patients with incident ICD-9 diagnosis code	A	B	A+B
Control Patients	C	D	C+D
Odds Ratio (OR)			(A/D) / (B/C)

Both crude OR and adjusted estimates using the Mantel-Haenszel (M-H) procedure controlling for age, sex, health plan (other) will be calculated in this study.

Other measures of association

Excess Risk Measures

Methods using non-parametric linear regression models of disease incidence have been developed to estimate excess risk within a nested case control framework (Borgan et al). In this method, absolute risk estimates associated with a particular covariate history may be computed, accommodating continuous and time-varying covariates. This is useful for patient cohorts having chronic usage of a drug. In addition, excess and absolute risk estimates are additive over age intervals. The methods are also useful in nested case control studies with multiple controls per case, often used to reduce computational burden in large cohort studies, or conducted to gather covariate information on a sample of the cohort. The number of parameter functions is bounded by the number of subjects in the sampled risk set at each failure time, a limitation since most nested case-control studies only collect 1or 2 controls / case. Pooling of controls is an option but the variance estimator will underestimate the true variability.

Relative Excess Risk

Consider a case-control study designed to assess the risk of two drugs. Assume that in addition to the 2 drugs under evaluation, the study includes unexposed reference group of subjects who do not use either of the drugs. The table below shows the usual contents of frequency table from such a study.

Table. Typical display and notation for case-control study with 2 exposure groups and an unexposed group.

Exposure	Cases	Controls	Baseline Rate Ratio
Drug 1	a_1	b_1	$rr_1 = (a_1d) / (b_1c)$
Drug 2	a_2	b_2	$rr_2 = (a_2d) / (b_2c)$
No Exposure to either	c	d	1 (reference)

Since absolute rates are not estimable from case-control studies, the RER can be written as a function of the 2 baseline rate ratios:

RER – $(RR_1 - 1) / (RR_2 - 1)$, where RR_i estimated by the odds ratio $rr_i = (a_id / b_ic)$ where i = 1, 2. Assuming trinomial distributions for exposure among cases and controls, the estimator or RER is given by: rer – $(rr_1 - 1) / (rr_2 - 1)$ with 95% CI: rer $\pm$ $1.96 rer V^{1/2}$ where

$V = d((a_1(a_1 + b_1 + c + d)/D_1^2) + a_1(a_2 + b_2 + c + d)/D_2^2) + a_1 / (b_1 D_1) + a_2 D_2)$
$- (a_1 b_2 + 2 a_1 b_1 + 2 a_1 a_2) / D_1 D_2))$, where $D_1 = a_1d - b_1c$ and $D_2 = a_2d - b_2c$

Since the RER can have negative values further research is needed on the appropriateness of the log transform to estimate CIs.

Potential thresholds:

For each method, one could construct alternative observational screening thresholds to apply. As described in the OMOP design, screening for Health Outcomes of Interest requires a more conservative approach to ongoing surveillance where the screening metrics are primarily used for prioritization, while identification of non-specified outcomes may more regularly use thresholds to restrict the set of drug-condition candidates that warrant further review.

Alternative approaches could be based on confidence intervals (where drug-condition pairs are identified on the basis of the lower bound), or based on the point estimate of the metric (where the magnitude of the potential effect is considered).

Multiplicity:

When performing the analysis to identify associations across all non-specified conditions, some have argued that this approach amounts to performing multiple statistical tests (one test for each condition) and that adjustments for multiplicity should be considered. Approaches, such as Bonferroni adjustment, have been offered. Others have argued that multiplicity adjustment is unnecessary in this exploratory phase provided an evaluation phase is anticipated, because it may increase the required information to observe a true association. Further consideration is required to determine whether adjustment for multiplicity is appropriate and necessary for inclusion in any identification methods.

Recurrent Analysis of Accumulating Data

It is reasonably anticipated that observational screening will be performed on a continuing basis during a product lifecycle. As such, analyses will be repeated over time as new data becomes available. Similar to the issue of multiplicity, some have agued that continuing reuse of data requires statistical adjustment to reduce false positive findings. It is not clear the extent to which these methods make such adjustment, and some argue that the tradeoff between reducing false positive by potentially increasing false negatives in unwarranted. Further consideration is required to determine whether adjustment for recurrent analysis is appropriate and feasible for inclusion in any identification methods.

Expected Strengths:

The designs outlined above are generally the strongest non-randomized design considered in instances where the outcome of interest is rare; it falls below cohort studies among observational study design hierarchies for evidence value. Because OMOP would be potentially deploying these methods both for known conditions (HOIs) and unknown, its value as an identification tool (particularly among unknown conditions) requires further assessment. Unlike cohort approaches, assessing populations who have already experienced an event is not intuitive and decisions around time windows for exposure are critical to the implementation and interpretation of the findings. There is some appeal to identifying all conditions of interest (known HOIs) in a database.

Expected Limitations:

The main potential limitation for the application of these methods would be the computational effort and evaluation effort of applying it to non-specified conditions. These designs were developed to evaluate rare events so their application as an identification method itself may present limitations as well as a more general, non-specific approach as an identification tool. One of the primary challenges with these approaches is the degree to which users have to make design decisions. Without best practices establishes, subjective decisions have to be made for all the variants explored above. There are no right answers to decisions so one set of decisions may be more appropriate for certain circumstances than others.

References

Feldmann U. Epidemiologic assessment of risks of adverse reactions associated with intermittent exposure. Biometrics 1993; 49:419–428.

Feldmann U. Design and analysis of drug safety studies, with special reference to sporadic drug use and acute adverse reactions. Journal of Clinical Epidemiology 1993; 46:237–244.

Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. Statistics in Medicine 1996; 18:1–15.

Guess HA. Behavior of the exposure odds ratio in a case-control study when the hazard function is not constant over time. *J. Clin Epidemiol*. 1989; 42:1179-1184.

Kaufman DW, Rosenberg L, Mitchell AA. Signal generation and clarification: use of case-control data. *Pharmacoepidemiol Drug Saf*. 2001 May; 10(3): 197-203.

Maclure M, Mittleman MA. Should we use a case-cross-over design? Annual Review of Public Health 2000; 21:193–221.

Maclure M. The case-cross-over design: a method for studying transient effects on the risk of acute events. American Journal of Epidemiology 1991; 133(2): 144–153.

Mansson, R., M. M. Joffe, et al. (2007). On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol 166(3): 332-9.

Miettinen OS, Caro JJ. Principles of non-experimental assessment of excess risk, with special reference to adverse drug reactions. *J. Clin. Epidemiol*. 1989; 42: 325-331.

Miller E, Goldacre M, Pugh S, Colville A, Farrington CP, Flower A, Nash J, MacFarlane L, Tettmar R: Risk of aseptic meningitis after measles, mumps and rubella vaccine in U.K. children. Lancet 1993; 341: 979–982.

Nurminen M. Assessment of excess risk in case-base studies. *J. Clin. Epidemiol*. 1992; 45:1081-1092.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986; 73:1–11.

Suissa. The case-time-control design: further assumptions and conditions. *Epidemiology*1998. 9:441-445.

Tata, L. J., Fortun P.J., et al. . Does concurrent prescription of selective serotonin reuptake inhibitors and non-steroidal anti-inflammatory drugs substantially increase the risk of upper gastrointestinal bleeding? Aliment Pharmacol Ther 2005; 22(3): 175-81.

Vines SK, Farrington CP. Within-subject exposure dependency in case-cross-over studies. Statistics in Medicine 2001; 20:3039–3049.

Whitaker HJ, Farrington PC, Spiessens B, Musonda P. Tutorial in biostatistics: the self-controlled case series method. Statist. Med. 2006; 25:1768-1797.

4. Surveillance Approaches for General Population

Description:
One key goal of OMOP is to assess the feasibility and utility of design, focus is placed on two primary types of conditions: 'health outcomes of interest' are those specifically-defined conditions that are the focus of ongoing surveillance for all medicines, and 'non-specified conditions' that may have other unanticipated relationships with drugs. For each type, OMOP intends to design and perform objective tests to evaluate the performance of alternative identification methods. Several alternative methods and approaches were identified through global introspection, literature review and patent review, but the working group acknowledges the list is not exhaustive of all potential approaches. Further work is recommended to perform a systematic methods review to ensure that best practices from all domains and disciplines are fully considered for their utility in supporting observational pharmacovigilance analyses.

Public health surveillance encompasses a number of techniques that operate at a population level and most are intended to be applied as real-time or prospective monitoring. These are hypothesis-generation efforts or intended to enumerate a particular type of case as would be required to identify potential new 'outbreaks'. In general, the approach is to capture multiple points of observation before and after an intervention is introduced, an event is known to have occurred, or a particular point in time. In some approaches, a comparison group as similar as possible (without particular exposure or event of interest) is included or a reference population accessed to identify expected rates (this is the case where population data exist- for cancer, birth defects for example). The repeated measurements should be equally spaced in time and rarely are there considerations given for adjustments due to multiplicity. Interrupted time series designs improve on the non-random control group pre-test / post-test design by introducing serial measurements before and after the intervention. This minimizes the weaknesses of single measurements such as regression to the mean and, to some extent, history as a threat to internal validity. For the comparisons to be valid, there is an assumption that prescription patterns don't change over time, in terms of the population under consideration. The change in event rates is only interpretable if we believe the use of the drug is unconfounded by severity, e.g., through restriction to patients who would have gotten the prescription? The rate in the overall population would be affected by the introduction of the exposure, but looking at the whole population, especially if a small number of people are exposed, could attenuate any increase in risk.

Starting point for consideration:
Assume the goal of the analysis is to identify any potential associations between any conditions and a particular target drug. Because the value and distinction of these methods is their application in real-time/prospectively, in order to implement this method for pre-specified conditions (HOIs), we would need to select some index time or event (it may be the introduction of a new drug into the population), identify the drug exposure and comparator group (source for expected rates) of interest and initiate measurement prior to the anticipated introduction date and

at regular intervals following. This can be done retrospectively and would resemble a retrospective cohort approach. These methods are distinguished from a cohort in that the entire population is being monitored for general changes in established patterns, as one would find in an ecologic study. The metrics are usually are expressed as a rate, relative risk, attributable risk, standardized mortality ratio (SMR), and the measurement of interest could be defined as crossing a significance threshold of a relative measure (which can include) or if no comparator is used, crossing an absolute rate threshold.

The application of this method to non-specified outcomes is more difficult. As with all other methods, the difficulty is in the complexity and magnitude of the number of potential associations. However, the method could be used to identify new outcomes not observed in the pre-exposure period by applying the technique as described above prior to the launch of a new drug, and identifying new conditions in the post-exposure period across all of the conditions that are represented in the data that exceed the identification threshold. The method may be particularly effective at identifying shifts in resource use as reflected in health care utilization.

Variants in design:
There are several variations associated with this approach and include population health surveillance with comparative interrupted time series, population health surveillance time series analysis, and time series analysis with autoregressive integrated moving average (ARIMA). The basic approach is that of population surveillance which captures multiple points of observation before and after the introduction of a drug and includes a comparison group as similar as possible (reflecting the expectation in the population) to derive prevalence estimates of rate of events over time correlated with the rate of drug use over time (assuming that these techniques are applied to determine change in conditions with the introduction of a drug into the population). The repeated measurements should be equally spaced in time. Interrupted time series designs improve on the non-random control group pre-test / post-test design by introducing serial measurements before and after the intervention. This minimizes the weaknesses of single measurements such as regression to the mean and, to some extent, history as a threat to internal validity.

Hauben (2003) reviews several variations in population approaches to identifying changes in the expected frequency of events. Although most of the literature applies to spontaneous adverse event reports, there is a literature of the application of these techniques in classic public health surveillance including vaccine research. Several of these are briefly reviewed here.

CUSUM techniques are based on the premise that over time, positive and negative deviations around a mean negate each other and would be unlikely to exceed some designated threshold (control limit) unless a nonrandom process truly increases the underlying population mean. It detects sudden changes in the mean value of a quantity of interest and provides estimates of the timing and magnitude of change by calculating the cumulative sum of deviations from a set value in successive samples. If the CUSUM exceeds an *a priori*-defined threshold value (can include an acceptable 'false alarm' rate, and average length of monitoring time), an 'alert' is detected. The threshold is determined by the average time until the threshold is exceeded and an

alert is detected compared to a pre-specified expected background incidence. This technique is related to sequential Wald's sequential probability ratio test.

The Poisson method is similar and is a direct application of statistical theory where it is assumed that events follow a Poisson distribution, a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate, and are independent of the time since the last event. This requires that we identify conditions from the database occurring in a population as well as an estimate of background incidence of the condition and level of drug utilization against which these changes may be viewed. With estimated background incidence of an adverse event and the number of patients treated, rare coincidences of drug and event are modeled by a Poisson distribution with the probability of at least x coincidences of statistically independent drug and conditions per time period. The critical number of patients experiencing the condition for rejecting the null hypothesis of statistical independence between drug and event can be calculated. If the number of conditions is greater than the critical value, the hypothesis of independence between drug and condition is rejected. $\text{Prob}(X \geq x) = 1 - \sum_{e}^{-\mu} \mu^x / x!$ Hauben cites an example of Schoeder (1998) for the possible association of spinal and epidural hematoma (extremely rare adverse events after spinal and epidural anesthesia) after neuroaxial blockade and preoperative thrombo-prophylaxis with a low molecular- weight heparin. Pre-specified tables can be generated that portray background (expected) incidence, alpha level, and the maximum number of events at each exposure level that would constitute the critical threshold. Historically, many critical events detected after approval have low baseline rates, and some researchers (Clark et al., 1999) claim that no more than one to three spontaneous reports should be coincidental under a Poisson distribution and that for diseases with extremely low baseline incidence such as aplastic anemia, more than three reports is a strong signal. Translating this to data with actual denominators is more straightforward.

Brownstein et al. (2007) applied CUSUM and interrupted time series techniques to "…elucidate long-term temporal trends in rates of inpatient visits for MI to an integrated health system in Boston, Massachusetts and their correspondence with prescriptions...[and to] estimate the magnitude of this effect on macro-level trends in hospitalizations for MI." The condition of interest, serious coronary heart disease, was defined as acute MI requiring hospitalization. The expected mean and standard deviation of MI incidence for a baseline period of 1997–1998 was used against which the researchers tested for cumulating deviations above a target mean of 47.1 MI-related hospitalizations per 100,000 and a standard deviation of 2.8 per 100,000.

Time series analysis with autoregressive integrated moving average (ARIMA) has also been mentioned. The main steps required by ARIMA modeling are the selection of the time series, transformations of the series, model selection, parameter estimation, forecasting, and updating of the forecasts. Its usefulness resides mostly in providing an estimate of the variability to be expected among future observations. This knowledge is helpful in deciding whether or not an unusual situation, possibly an outbreak, is developing. Using ARIMA methods and historical data, robust models for expected rates of each outcome could be generated. As new data are

incorporated, these rates can be assessed against historical models of expected rates to determine the deviation from the expected controlling for multiple time and potential confounding factors.

Comparator selection:

The comparator for these techniques is an expected rate of the condition under surveillance. This can be derived from standard population references (as in birth defects or cancer surveillance) or from the population data itself. One may also consider the time window prior to the introduction of the new intervention as being a 'comparator' in which case it would be important to select a window and number of assessments prior to the intervention that will be meaningful.

Matching:

There is no matching in these analyses, as the comparison is to expected rates generated from another database or the population at large. If we were to directly match to develop a comparison population, the technique would essentially become a cohort study.

Restriction:

Restriction may be considered in this application as it applies to how definitions of the conditions of interest are defined or the population itself is defined. It is unclear if there is a simple systematic approach.

Censoring:

In the base case example above, all conditions that occurred during the pre- and post-exposure windows may be captured as part of the rate estimation. Given that the methods are seeking to identify trends and these data may represent counts of conditions based on clinical visits, censoring may be invoked to assure that the counts of conditions are restricted to a single event (the 'first' occurrence) in a given patient unless utilization frequency is the 'condition' of interest.

Periods of observation:

This is a critical defining characteristic of these methods and should be consistent with the time windows before and after the intervention of interest is introduced into the population.

Metrics:

For most of these methods, SMRs or similar observed to expected ratios are calculated and plotted over time. The example below, from Brownstein et al., (2007), illustrates the typical output for CUSUM.

Potential thresholds:
Threshold determination is specified within the metric itself but does leave some room for 'false alarm' rate to be tolerated.

Adjustment:
Generalized linear modeling has been used with Poisson and time series analyses to adjust for secular trends in the condition of interest and to correct for overdispersion of data.

Multiplicity
There are not inherent adjustments for multiplicity in these techniques as the time windows are treated as discrete, independent events.

Recurrent Analysis of Accumulating Data
These techniques rely on refreshed data at regular intervals however most treat each time window as discrete and independent.

Expected Strengths:
These techniques are based in public health surveillance and are best suited for specified HOIs for which there are credible expected rates available, and the conditions of interest are rare. These techniques are also a bit easier to communicate, as a wider audience is able to understand the rise in one exposure coincident with the rise in a particular condition.

Expected Limitations:
Ecologic fallacy occurs when one assumes that population trends that may be correlated (as in the frequency of drug exposure in the population is correlated with change in some condition frequency) as is possible in the Brownstein paper. These techniques would be complex to implement for non-specified conditions although if implemented, could provide additional hypotheses to consider. It is unclear how their performance characteristics (sensitivity, specificity, positive and negative predictive value) compare to more traditional cohort methods.

Before-after designs cannot discriminate between conditions and utilization leading up to an intervention and the impact of the drug on conditions after the intervention. For example, if there is a workup to exclude brain tumor or intracranial bleed prior to initiating a new therapy for migraine, it may appear that the new therapy reduces radiology and laboratory utilization when these were part of the 'work-up'. Similarly, there may be secular trends that occur concomitant to the introduction of the drug in the market. This underscores the importance of remembering that these are hypothesis-generating methods.

References

Brownstein JS, Sordo M, Kohane IS, Mandl KD (2007) The Tell-Tale Heart: Population-Based Surveillance Reveals an Association of Rofecoxib and Celecoxib with Myocardial Infarction. PLoS ONE 2(9): e840. doi:10.1371/journal.pone.0000840.

Clark JA, Berk RH, Klincewicz SL. Calculation of the probability of multiplicities in two cell-occupancy models: implications for spontaneous reporting systems. Drug Inf J 1999; 33: 1195-203.

Hauben M, Zhou X: Quantitative methods in pharmacovigilance. Drug Safety 2003; 26:159-186.

Schroeder DR. Detecting a rare adverse drug reaction using spontaneous reports [statistics]. Reg Anesth Pain Med 1998; 23 (6): 183-9

5. Other Methods for Consideration

During the initial peer review of the OMOP experimental design, the following commentary was offered by David Page (U Wisconsin) that requires further work during the methods development phase:

> As the proposal elegantly summarizes, one may use pre-defined HOIs or not, and one may seek to evaluate a proposed association or to discover (identify) previously unknown associations. I find the Identification Methods Matrix to be nicely organized and relatively complete regarding already-existing approaches. Nevertheless, I think a key opportunity is being overlooked for a novel use of powerful existing algorithms. This opportunity can be glimpsed when surveying the "Other methods" section of the matrix and considering what many of these methods have in common.
>
> "Other methods" contains an incomplete but wide variety of *data mining* and *supervised machine learning* algorithms, including classification trees and recursive partitioning, hierarchical models, neural networks, causal modeling, pattern discovery. These and many other related algorithms can all be applied *in essentially the same, natural way* within each of the four broad classes of methods in the Identification Methods Matrix, as I discuss in detail below (in reverse order). I believe doing so will significantly enhance the performance of a pharmacovigilance system beyond what is presently proposed.
>
> **Surveillance approaches for general population**: When a new drug is placed on the market, data mining and supervised machine learning algorithms can learn a model to predict who is on the drug, based on data about symptoms and diagnoses, other prescriptions, laboratory results, and other data of the types discussed in the proposal. Most likely, drug exposure trivially can be predicted with reasonable accuracy based on the symptoms or diagnoses for which the drug is prescribed; therefore, a patient's data must be censored to begin only at time of prescription of the new drug. If some combination of *subsequent* symptoms, diagnoses, drug prescriptions, etc. can be used to predict drug exposure with accuracy significantly better than chance (as assessed by cross-validation – see Final Comments in this review), then the model may have identified an adverse event. The adverse event may correspond to a single diagnosis code (a known HOI); or if not (no known HOI), the adverse event may be described indirectly in the model via a combination of existing diagnosis codes, drugs to treat the event, lab results indicative of the event, etc.
>
> Supervised learning and data mining algorithms suitable for such modeling include those mentioned in the "Other methods" page of the Identification Methods Matrix spreadsheet, but they also include a number of other approaches. Leading alternative approaches include support vector machines (SVMs) [e.g., Shawe-Taylor & Cristianini, 2000], logistic regression, Bayesian network learning algorithms (implicit in "causal modeling" already mentioned) [e.g., Heckerman, 1999], ensemble methods such as bagging

[Breiman, 1996] or boosting [Freund & Schapire, 1996] of classification trees or SVMs, or random forests [Breiman, 2001].

All of the preceding methods have in common that they require each data point – here, each patient – to be represented as a feature vector. But this feature vector requirement causes some loss of information from each patient's clinical history. For example, the feature vector can naturally encode that a patient has diagnosis D and prescription P, but not that P precedes D by at least two weeks. To operate with each patient's richer clinical record, rather than summarizing it as a feature vector, one may use relational machine learning and data mining techniques, such as inductive logic programming (ILP) [e.g., Dzeroski & Lavrac, 2001], graph mining [e.g., Cook & Holder, 2006] or statistical relational learning (SRL) [Getoor & Taskar, 2007], techniques not mentioned in the proposal. Despite the diversity of these tools, they all can be applied in the same manner to learn models as described above, only the models can be more expressive. ILP algorithms and some SRL algorithms have the added benefit that they can be used simply to find commonalities within *small (previously undefined) subsets* of the patients on a drug. For example, ILP learns rules that cover (explain) some minimal number of cases (patients on the drug) while not falsely covering controls (those not on the drug). An example rule might be:

A patient is likely on the new drug if he/she has since been prescribed with a beta-blocker and an acid-reducing medication within three months of each other.

Even if this rule is true of only 200 of the twenty thousand patients who started on the new drug six months ago, if it is true of only *five* patients out of another twenty thousand *not* on the new drug, then the rule may be capturing an adverse event.

All of the machine learning tools mentioned in this section can be applied to distinguish between patients on the drug and those not on the drug. They are relevant also when learning predictive models for the subsequent task types. While not all the methods necessarily need to be employed on every task type, it makes sense to at least test most of them to see which ones give the best performance for this pharmacovigilance application.

Approaches based on selecting populations with condition: Here the machine learning algorithm is tasked to learn a model to distinguish between people on the new drug and those not on the new drug, from among only people with the condition of interest. Again if an accurate model can be learned, it suggests a possible drug effect. For example, suppose among people suffering a myocardial infarction, those on the drug in question can be distinguished from those not on the drug because those on the drug tend to be younger and to have experienced a marked increase in blood pressure within the last year. This predictive model is then capturing a regularity in the data; the fact that a predictive regularity exists at all indicates the possibility that some of these MIs may have been drug-induced; the nature of the regularity adds weight to this possibility. Again, the same supervised learning and data mining algorithms are applicable here as in the preceding section.

Approaches based on selecting populations taking drug: If an HOI is known, supervised learning can be employed to distinguish between those patients who have the HOI and those who do not, among those on the drug. If an accurate model can be learned, it can be used to identify those patients at particularly high risk for the outcome. In contrast to the other two, preceding applications of machine learning, here one may want to limit the data to that *before* the initial prescription of the drug of interest. In this case, the predictive model is one that can be used before initiation of the drug to predict which patients are most at risk for the outcome if they take the drug. Alternatively, one may want to limit the data to that before the outcome occurs, to yield a model that may identify changes in patients on the drug as they progress toward the outcome.

Disproportionality analysis approaches from spontaneous adverse event reporting: Perhaps the best known of all data mining algorithms is the Apriori algorithm for association rule learning [Srikant & Agrawal, 1994]. Association rules identify two or more attributes (features with specific values) in the data points (patients, in this case) that appear together more frequently than one would expect assuming independence of the attributes. Algorithms for learning association rules rank the associations according to one of several possible scoring functions that take into account correlation and frequency of the attributes. Association rule mining can be applied here specifically for associations where at least one of the attributes in question is a drug. An association between a drug and a pair of diagnosis may be indicative of an adverse event, as may be an association between a drug and another drug. Here one probably will want to use data on each patient only after initiation of the drug. Otherwise the method will yield trivial associations, such as between the drug and the diagnosis for which the drug is indicated.

Final Comments

Because of the temporally sequential nature of clinical histories, one might also consider algorithms for learning from sequence data, such as hidden Markov models and dynamic Bayesian networks. Nevertheless, given the wide variability in the durations and number of observations for each patient, it probably is more appropriate to use relational learning methods such as ILP or SRL, rather than sequence models, to take advantage of the longitudinal aspects of the data.

With such rich data sets as those described in the proposal, the multiple comparisons problem arises, or equivalently, the problem of overfitting the training data (developing a model that fits the training data well but doesn't extend to unseen cases). The best approach to avoid this problem is to test any predictive model, or purported association between drug and (possibly complex) event, on unseen data. This can be accomplished by a held-aside test set that is only examined after a model or association has been discovered. When cases are rare, data is limited, so it is usually preferable to perform cross-validation. For example, in the applications of supervised learning I've described above, one would want to estimate the accuracy of a learned model by ten-fold cross-validation. If cross-validation is infeasible (very limited data, or many different

associations are proposed), then to avoid overfitting or the multiple comparisons problem, it is necessary to use much more stringent conditions for saying an association is of interest. False discovery rate, as mentioned in "Other methods," provides one mechanism for doing so.

Supervised machine learning and data mining provide a suite of powerful algorithms that are a natural fit for this application and have not been given sufficient consideration in formal testing. Types of algorithms that should be considered include association rule mining algorithms such as Apriori, support vector machines (SVMs), Bayesian network learning algorithms such as TAN [Friedman, Geiger & Goldszmidt, 1997], naïve Bayes or K2 [Heckerman, 1999], logistic regression, ensemble methods such as random forests, and relational learning algorithms – particularly ILP algorithms such as Aleph, and SRL algorithms such as SAYU [Davis *et al.,* 2007] and Markov Logic Networks [Domingos & Richardson, 2006]. Association rules, classification trees (and recursive partitioning), ILP and SRL approaches are particularly attractive because the models are easily inspected visually and easily comprehended.

References

Rakesh Agrawal & Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Databases* (VLDB-94), pp. 487-499, Morgan Kaufmann, 1994.

Leo Breiman. Bagging predictors. *Machine Learning* 24 (2): 123-140, 1996.

Breiman, Leo. Random Forests. *Machine Learning* 45 (1), 5-32, 2001.

Diane Cook & Lawrence Holder. *Mining Graph Data*. New York: Wiley, 2006.

Jesse Davis, Elizabeth Burnside, Ines Dutra, David Page, Raghu Ramakrishnan, Vitor Santos Costa and Jude Shavlik. Learning a New View of a Database: With an Application to Mammography. In L. Getoor and B. Taskar, Eds., *Introduction to Statistical Relational Learning*, Chapter 17. Cambridge, MA: MIT Press, 2007.

Pedro Domingos & Matt Richardson. Markov Logic Networks. *Machine Learning*, 62, 107-136, 2006

Saso Dzeroski and Nada Lavrac, Eds. *Relational Data Mining*. Berlin: Springer, 2001. Chapters 3-7, 15.

Yoav Freund and Robert E. Schapire A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119-139, 1997. http://www.cse.ucsd.edu/~yfreund/papers/adaboost.pdf.

Nir Friedman, Dan Geiger & Moises Goldszmidt. Bayesian network classifiers. *Machine Learning,* 29, 131-163, 1997.

Lise Getoor & Ben Taskar. *Introduction to Statistical Relational Learning*. Cambridge, MA: MIT Press, 2007.

David Heckerman. A Tutorial on Learning with Bayesian Networks. In *Learning in Graphical Models*, M. Jordan, Ed. Cambridge, MA: MIT Press, 1999. Also appears as Technical Report MSR-TR-95-06, Microsoft Research, March, 1995.

John Shawe-Taylor & Nello Cristianini. *Support Vector Machines and other kernel-based learning methods.* Cambridge, U.K.: Cambridge University Press, 2000

