Background

The Observational Medical Outcomes Partnership (OMOP, http://omop.fnih.org) is a public-private partnership managed by the Foundation for the National Institutes of Health (FNIH), chaired by the Food and Drug Administration (FDA), supported by a consortium of pharmaceutical research organizations, with active participation from academia, private industry, providers and other stakeholders throughout the healthcare system. OMOP was formed to conduct methodological research to inform the appropriate use of observational healthcare data for studying the effects of medical products. As part of its research, OMOP developed tools and capabilities for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum, and established a shared resource to enable collaborative research to advance the science. In keeping with its mission, all of the research and work products from OMOP are placed into the public domain for use and dissemination across the broader research community.

OMOP has conducted a series of experiments to generate empirical evidence about the performance of observational analysis methods in their ability to identify true risks of medical products and discriminate from false findings. These experiments were designed to inform the development of a risk identification and analysis system, as envisioned by various pharmaceutical research companies and now mandated for the FDA by Congress through the FDA Amendment Act of 2007. We define a ‘risk identification and analysis system’ as a systematic and reproducible process to generate evidence efficiently to support the characterization of the potential effects of medical products from across a network of disparate observational healthcare data sources.

Recent Findings

In June 2012, the OMOP research team presented results from its latest experiments, which resulted in recommendations for building a risk identification and analysis system, as well as guidance for interpreting observational studies. The proceedings from the 2012 OMOP Symposium are available at http://omop.fnih.org/2012SymposiumPresentations.

In its latest experiment, the OMOP team evaluated the performance of a risk identification system for four health outcomes of interest: acute myocardial infarction, acute liver injury, acute renal failure, and gastrointestinal bleeding. For these outcomes, OMOP established a reference set of 399 test cases: 165 ‘positive controls’ that represent medical product exposures for which there is evidence to suspect an association with the outcome, and 234 ‘negative controls’ that are drugs for which there is no evidence that they are associated with the outcome. The fundamental goal of OMOP’s research is to develop and evaluate standardized algorithms that can reliably discriminate the positive controls from the negative controls, and to understand how an estimated effect from an observational study relates to the true relationship between medical product exposure and adverse events.

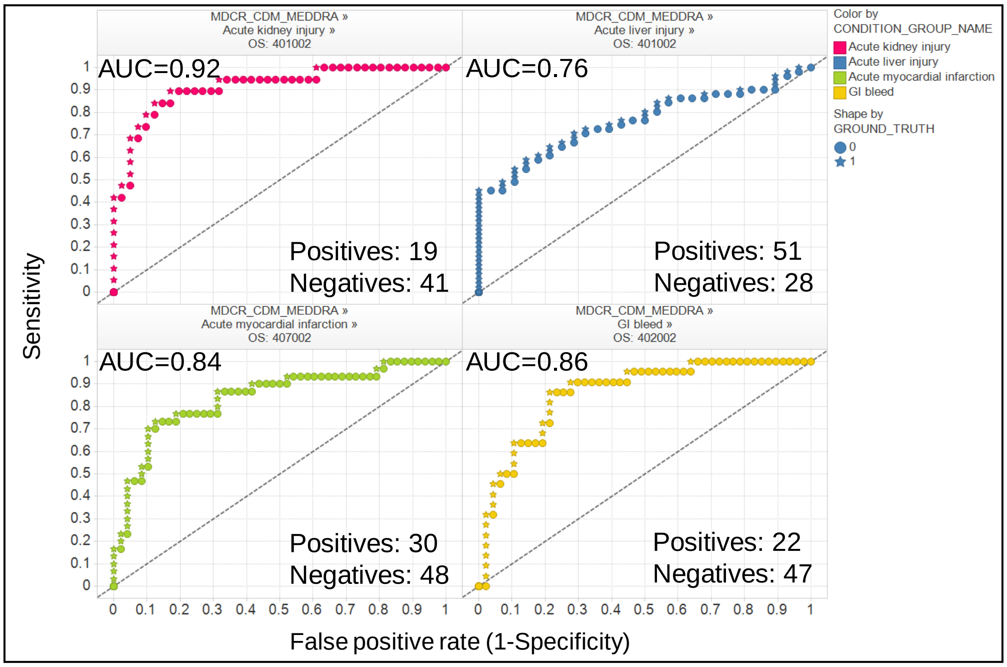
To conduct this evaluation, OMOP licensed five observational healthcare databases, representing both administrative claims and electronic health records, and covering over 70m patients with more than 150m person-years of longitudinal observation. The OMOP community designed, implemented, and tested 7 different standardized analytical methods that estimate the strength of association between any exposure and outcome to produce a relative risk and standard error. These methods included study designs commonly observed in the published epidemiology literature, such as case-control, propensity-adjusted new user cohort, and self-controlled case series. OMOP executed these methods, with various study design variants, against the five observational healthcare databases and 16 simulated datasets for all of the 399 test cases. This evaluated the degree to which the estimated effects were reliable predictors of true causal relationships and appropriately measured the magnitude of the true effects.

The OMOP experiments suggest that if methods are applied to all drug-outcome pairs and interpreted in aggregate, then results are modestly predictive of positive vs. negative classification. All methods were significantly better than random chance (suggesting that the data can be legitimately informative for decision-making), but no method was close to being perfectively predictive (suggesting that these data should not be interpreted as definitive evidence). The OMOP team sought to improve the performance by developing strategies for interpret the results, and identified four specific heuristics. First, we identified that partitioning results by outcome improved performance. Instead of applying a risk identification system to all outcomes concurrently, we found that evaluating products within a single outcome yielded higher predictive accuracy. So the question is not: 'how well does a risk identification system work?' but rather 'how well does a risk identification system for acute myocardial infarction work?' and 'how well does a risk identification system for acute liver injury work?'

Second, since outcomes have been partitioned, we found that better performance could be achieved by tailoring the analysis strategy to the outcome. Rather than applying one method across all drug-outcome pairs, one can take an empirical approach to selecting the method that produces the greatest predictive accuracy for each outcome. Third, we found that performance was improved if we restricted the scope of the risk identification system to only those drug-outcome pairs with sufficient sample size to be studied. Just as we power studies to make sure we can reliably detect a desired association, so too can be consider a risk identification system to be most reliable in the contexts where enough data are available. A consequence of this heuristic is that some products and outcomes will be deemed out-of-scope and not evaluable in this framework, but failure to restrict these underpowered scenarios can result in less reliable answers. Finally, we found that further performance gains could be achieved by optimizing the choice of analysis to the particular data source.

Combining these heuristics yielded predictive accuracy that is at least as good as most diagnostic tests regularly used in clinical practice. The ROC curves below (Figure 1) demonstrate the performance of the optimal strategy in the MarketScan® Medicare database and highlight the tradeoff between sensitivity and specificity. If users of a risk identification system applied these recommendations and sought to achieve 50% sensitivity (identify at least half of the positive effects), then they could expect to observe specificity of at least 89% (indicating that less than 10% of null effects would be identified as false positive findings). Similar results were observed across the other databases. We believe the process developed to study these four outcomes in observational data can and should be applied systematically by safety scientists for all outcomes and databases.

Figure 1: ROC curves demonstrating performance of test cases



If target sensitivity = 50%	Threshold	Specificity
Acute kidney injury	2.69	95%
Acute liver injury	1.51	89%
Acute myocardial infarction	1.59	92%
GI bleed	1.87	94%

Different stakeholders may have different tolerances for the risk of false negatives and false positives. The OMOP experiments provided empirical evidence allowing stakeholders to evaluate the appropriateness of the system, identify areas for its specific use, and investigate the consequences of decision thresholds they may choose. It also provides the broader research community a common benchmark to measure against when gauge the effectiveness of future research in improving both data and methods.

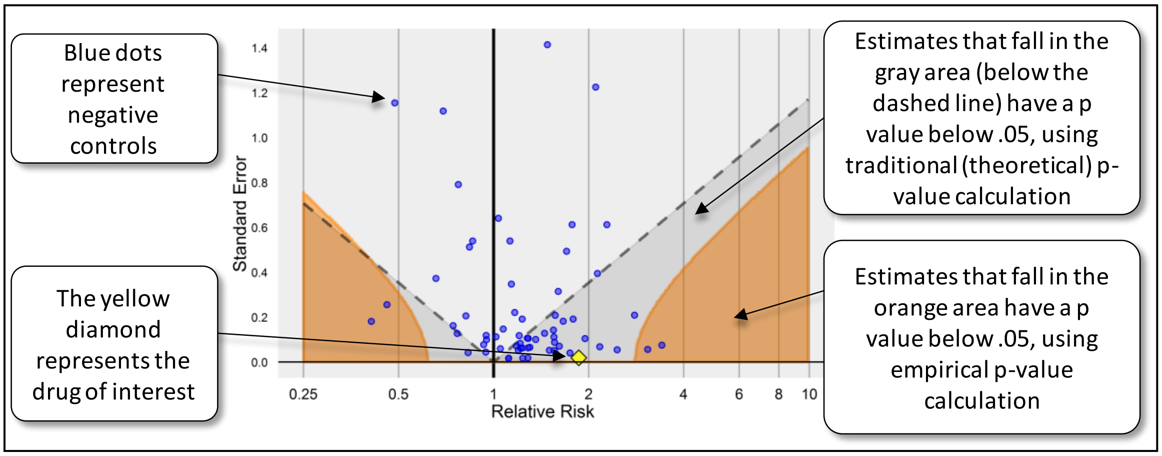
Several insights were gained from the current experiments about expected behavior of a risk identification system. We observed that self-controlled designs are optimal across all outcomes and all sources, but the specific settings are different in each scenario. All sources achieve good performance (Area under ROC curve > 0.80) for acute kidney injury, acute MI, and GI bleed, while acute liver injury has consistently lower predictive accuracy. A risk identification system should confidently discriminate positive effects with relative risk>2 from negative controls, but smaller effect sizes will be more difficult to detect. There was no evidence that any of the five data sources were consistently better or worse than others, but we did observe substantial variation in estimates across sources pointing to the need to routinely assess consistency across a network of databases. The results underscore the importance of transparency and complete specification and reporting of analyses, as all study design choices were shown to have the potential to substantially shift effect estimates. Diversity in performance and heterogeneity in estimates arose not only from different study design choices (e.g., cohort versus case-control) but also from analytic choices within study design (e.g., number of controls per case in a case-control study). We caution against generalizing these results to other outcomes or other data sources. However, we do think OMOP has now provided a well-defined procedure for how to profile a database and construct an optimal analysis strategy for a given outcome, which can be systematic, reproducible, and yield defined performance characteristics that can directly inform decision-making.

The results from the OMOP experiment can be used to go beyond the design of a risk identification system and have implications for the interpretation of observational database studies, as commonly seen in the published literature. Studies are often published where a specific design (e.g. cohort or case-control) is applied to a particular database representing a given population (e.g. administrative claims for employees with private insurance) to address a particular hypothesis that a drug is associated with an outcome. These studies will typically report an effect estimate (e.g. relative risk) and its associated confidence interval and p-value to gauge its statistical significance. The OMOP results provide an empirical basis for assessing the assumptions and appropriateness of interpreting conventional statistics from observational studies.

Traditional p-values provide an assessment of whether an observed estimate is different from ‘no effect’. They are based on a theoretical null distribution that assumes an unbiased estimator, but that assumption is commonly violated in observational data. For example, when applying a case-control design to negative controls for GI bleed we expect that 5% of the analyses will yield a p-value less than 0.05. Instead, we found that 55% of drugs were falsely identified as significant at $p<0.05$. We have developed a reproducible procedure that uses the OMOP results to produce ‘adjusted p-values’ which are calibrated to yield the desired 5% false positive rate.

We developed an empirical approach (Figure 2) to quantifying the posterior probability of a true effect, given an observed estimate and prior beliefs. Comparing the distribution of negative controls with the distribution of positive controls provides complementary information beyond the p-value, and demonstrates that $p<0.05$ doesn't guarantee a true effect exists and $p>0.05$ doesn't guarantee no effect is present. We also showed that the traditional interpretation of a 95% confidence interval (CI), that is, that the CI covers the true effect size 95% of the time, is often misleading in the context of observational database studies. We found that the coverage probability is much lower than 95% across all methods and all outcomes, and these findings were consistent across real data and simulated data. In fact, many methods have coverage probability < 50%, meaning the true effect size is outside the bounds of the confidence interval half the time. We have devised a tool to empirically adjust standard confidence intervals to yield approximately correct coverage probabilities across all method-outcome scenarios.

Figure 2: An empirical approach to null hypothesis testing



Blue dots represent negative controls

The yellow diamond represents the drug of interest

Estimates that fall in the gray area (below the dashed line) have a p value below .05, using traditional (theoretical) p-value calculation

Estimates that fall in the orange area have a p value below .05, using empirical p-value calculation

Next Steps

OMOP's research continues to reaffirm the notion that advancing the science of observational research requires an empirical and reproducible approach to methodology and systematic application. When long-term support is obtained, the OMOP research community is well-positioned to lead in this endeavor into the future.