MINI-SENTINEL SYSTEMATIC REVIEWS OF VALIDATED METHODS FOR IDENTIFYING HEALTH OUTCOMES USING ADMINISTRATIVE DATA

ANALYSIS OF EVIDENCE GAPS AND LESSONS LEARNED REPORT

Prepared by: Ryan M. Carnahan, PharmD, MS, BCPP

Author Affiliations: The University of Iowa College of Public Health, Department of Epidemiology

August 11, 2011

Mini-Sentinel is a pilot project sponsored by the U.S. Food and Drug Administration (FDA) to inform and facilitate development of a fully operational active surveillance system, the Sentinel System, for monitoring the safety of FDA-regulated medical products. Mini-Sentinel is one piece of the Sentinel Initiative, a multi-faceted effort by the FDA to develop a national electronic system that will complement existing methods of safety surveillance. Mini-Sentinel Collaborators include Data and Academic Partners that provide access to health care data and ongoing scientific, technical, methodological, and organizational expertise. The Mini-Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223200910006I.

Mini-Sentinel Systematic Reviews Of Validated Methods For Identifying Health Outcomes Using Administrative Data

Analysis of Evidence Gaps and Lessons Learned Report

I. INTRODUCTION

The Food and Drug Administration Mini-Sentinel contract is a pilot program that aims to conduct active surveillance to detect and refine safety signals that emerge for marketed medical products. The program will utilize administrative data from a number of collaborating partners to examine and strengthen signals related to the safety of drugs, medical devices, and biologics. Because diagnosis codes in administrative data are not always accurate, it is important to understand the performance characteristics (i.e., predictive values, sensitivity, and specificity) of various algorithms for identifying health outcomes from the codes contained in administrative data.

The Mini-Sentinel Coordinating Center contracted with investigators to conduct systematic reviews of studies focused on the validity of algorithms for identifying 20 health outcomes of interest (HOIs) in administrative data. In the event that few validation studies were identified (less than five), the investigators also reported on non-validated algorithms that had been used to identify the health outcomes. Because of significant overlap of two outcomes (revision and excision of orthopedic medical devices), and overlap of the studies examining these outcomes, the responsible investigators provided one report to examine both outcomes. Thus, a total of 19 systematic reviews were conducted. The purpose of this document is to briefly summarize the results of these reviews, determine the gaps in evidence for performance characteristics of the algorithms for these health outcomes of interest, provide recommendations for health outcomes on which validation studies are most important for understanding performance characteristics, and describe lessons learned in the systematic review process such that future systematic reviews of these types of studies can optimize efficiency while remaining as comprehensive as is reasonable given resource limitations. The reports reviewed here can be found on the Mini-Sentinel website at the following address: http://mini-sentinel.org/foundational_activities/related_projects/default.aspx

II. SUMMARY OF RESULTS

This section briefly summarizes the key findings of the 19 systematic reviews, focusing on the number of validation studies identified and the acceptability of the performance characteristics determined in the validation studies. Further information can be found in the reports.

A. CEREBROVASCULAR ACCIDENT/TRANSIENT ISCHEMIC ATTACK

Andrade and colleagues identified 34 studies that reported on the validity of algorithms for cerebrovascular accidents (CVA) and/or transient ischemic attacks (TIA). The algorithms ranged from examining composite endpoints to studying a specific type of cerebrovascular incident. Algorithms for studying stroke or intracranial bleeds had positive predictive values (PPVs) of 80% or greater. Algorithms for TIA had PPVs of 70% or greater. The validation criteria for outcomes varied among studies, but medical record review was nearly always the method used for validation.

Because of the large number of studies and differences in the specific types of events studied, the investigators separated the studies into those studying each of four types of outcomes: composite endpoints, stroke/CVA, TIA, and intracranial bleeds/subarachnoid hemorrhage.

1. Composite Endpoints

Ten studies examined composite cerebrovascular endpoints. Most studies examined both incident and prevalent disease. The PPVs were similar for incident or prevalent disease, ranging from 57-92% for acute events and 33-96% for those examining a history of or current disease. Performance characteristics varied widely depending on the inclusiveness of the codes used in the algorithms. Algorithms using International Classification of Disease, 9th edition, Clinical Modification (ICD-9-CM) codes 430.x-438.x had the lowest PPVs, while other studies including a more restrictive set of codes, e.g., 433.x-436.x, had higher PPVs.

2. Stroke/CVA

Twenty-five studies examined stroke/CVA. Most studies used ICD-9-CM codes 434.x and 436.x, but other codes were also studied. A number of studies examined specific codes within a range of possibly relevant codes. Codes 430.x, 431.x, or 434.x generally had PPVs of 80% or better. Codes 436.x generally had PPVs of 70% or better. One study examined hospital discharge codes 430.x and 431.x separately and found a PPV of 13% for 430.x compared to 71% for 431.x.

3. TIA

Six studies examined TIA. Three of five studies examined hospitalizations or emergency department encounters with code 435.x and found PPVs of 70% or better. When outpatient codes were included, PPVs were much lower (33% in one study). Other related codes had much lower PPVs for TIA, and one study found a low PPV for code 435.9 (28%).

4. Intracranial Bleeds

Four studies examined intracranial bleeds. PPVs were 77% or better, with the lowest PPV from a study using a larger number of codes. Three studies used hospitalization diagnoses, while one examined hospitalization or emergency department diagnoses and found similar PPVs.

Assessment of Gaps

A large number of validation studies have been conducted for algorithms identifying cerebrovascular accidents and transient ischemic attacks, as well as specific types of events that fall within this outcome. The PPVs for the better algorithms generally exceeded 70% to 80%, depending on the specific outcome. Thus, it would not appear that this outcome should be a high priority for a Mini-Sentinel validation study unless validity of an algorithm in a particular subset of understudied patients was of interest. The report authors recommended that future studies might examine differences in validity of algorithms based on age and sex, differentiation of ischemic strokes caused by thrombosis versus emboli, comparisons of algorithms based on standard criteria, and comparisons of algorithms for incident versus recurrent events. Also, only a small number of studies examined ICD-10-CM codes.

B. CONGESTIVE HEART FAILURE

Saczynski and colleagues identified 35 studies that reported on the validity of algorithms for heart failure. These included inpatient and outpatient settings, incident or prevalent heart failure, and studies examining the performance characteristics based on the position of heart failure codes. The

Framingham Heart Study criteria for congestive heart failure (CHF) were the most common validation criteria and medical record review was the most common source of validation information. Most algorithms had PPVs greater than 90%. Incident CHF could be identified quite well by requiring a multi-year disease-free baseline period of eligibility.

Assessment of Gaps

The highest PPVs were from studies that required a primary hospital discharge diagnosis ICD-9-CM code of 428.x. Sensitivity of this algorithm might be suboptimal since heart failure is commonly managed in the outpatient setting. The report authors recommended future research on the performance of inpatient versus outpatient codes and in specific subpopulations of interest such as the elderly or very elderly, those with multiple chronic conditions, or in specific subgroups defined by race/ethnicity. They also recommend further research on the performance characteristics of ICD-10-CM codes. Overall, further validation studies of heart failure algorithms might be considered of low to moderate priority given the large amount of existing data in comparison to many of the other outcomes of interest.

C. ATRIAL FIBRILLATION

Jensen and colleagues identified 16 studies that reported on the validity of algorithms for atrial fibrillation. They found PPVs ranging from 56-100% (median 85%) for inpatient and outpatient codes. Studies that required an electrocardiogram to confirm atrial fibrillation tended to have PPVs in the lower end of the range. Six studies examined sensitivity and found a range from 57-95% (median 78%). Only two studies specifically sought to identify incident atrial fibrillation, and found PPVs of 62% and 77%.

Assessment of Gaps

Despite a large number of studies on this topic, only two studies sought to identify incident atrial fibrillation. Given that this is likely to be the specific outcome of interest for studies examining exposures that cause atrial fibrillation, confirmatory validation studies may be warranted. The report authors recommended the use of electronic ECG data, where available, as part of future algorithms to be studied. They also suggested future research on whether both an ICD code and an ECG should be required in an algorithm, the number of diagnoses that should be required in an algorithm, and the time period within which these diagnoses should occur to consider a person to have incident atrial fibrillation.

D. SERIOUS CARDIAC ARRHYTHMIAS

Tamariz and colleagues identified 9 studies that reported on the validity of algorithms for serious cardiac arrhythmias. Several studies examined incident arrhythmias after exposure to medications. Unfortunately, three studies did not include the specific codes used to identify the event, making it difficult to operationalize the algorithms. PPVs ranged widely, from 5-100%. The use of ICD-9-CM codes 426.x resulted in much lower PPVs. Codes 427.x and 798.x appeared to perform better, with most PPVs in the 70-80% or better range. In the studies that examined sensitivity and specificity, sensitivity was 77% or better and specificity was 84% or better. Both incident and prevalent events were studied, with the majority of studies focusing on incident arrhythmias.

Assessment of Gaps

A fair amount of evidence exists on the performance characteristics of various algorithms for identifying serious cardiac arrhythmias. Given the variability in results, algorithms, and the availability of specific algorithms, further validation studies might be considered of moderate priority. The report authors suggested that future research might make more use of pharmacy, procedure, or diagnosis related group codes. They also suggested future studies focused on high-risk groups, such as those with heart failure, or groups of different race/ethnicity. Few studies have examined ICD-10-CM codes. Lastly, the majority of studies used data on medical encounters to identify cases, though a small number examined death certificates for case confirmation or cause of death information. It can be predicted that some number of fatal arrhythmias would be missed in any study that used only administrative and claims data to identify these events because some patients would die before getting to medical attention. Future work might examine the number of events identified in administrative and claims data versus death certificates to better establish the sensitivity of administrative and claims data in identifying these events.

E. VENOUS THROMBOEMBOLISM

Harkins and colleagues identified 20 studies that reported on algorithms to identify deep vein thrombosis (DVT) or pulmonary embolism in administrative data, and the validity of the algorithms. A wide range of populations were studied and there was some variability in findings. While the performance of specific codes varied, the better algorithms using more focused sets of codes generally had PPVs greater than 70%, and some PPVs exceeded 90%. This was true for both DVT and pulmonary embolism, though pulmonary embolism algorithms generally had higher PPVs than those for DVT.

Assessment of Gaps

Given the number of studies available that generally showed relatively good performance of algorithms, these outcomes might be considered lower priority for future validation studies. The report authors noted that there was a lack of data examining the PPV of codes in high-risk populations such as those receiving orthopedic surgery. However, it could be predicted that the PPVs in these populations would actually be higher due to a greater prevalence of DVT and pulmonary embolism. The authors also suggested that more research should be conducted on the potential value of adding other types of codes to algorithms, such as procedure codes, diagnosis related groups, or pharmacy claims for anticoagulant medications, since very few studies examined how including these codes might affect PPV.

F. DEPRESSION

Townsend, et al. identified 11 studies that examined the validity of algorithms to identify depression. The validation criteria varied, including patient self-report, depression suggested or confirmed by screening or psychometric instruments, and depression diagnosed by a physician per medical records. PPVs ranged widely, from 31.5% to 98.8%. Unfortunately, the PPVs were not always clearly dependent on the type of validation of the outcome that was required. Even studies only requiring a medical record diagnosis for validation had a wide range of PPVs, from less than 50% to nearly 100%. The studies using assessments of depressive symptoms illustrated the poor sensitivity of medical diagnoses to depression. It is well recognized that depression is under-diagnosed. The use of antidepressant prescriptions to

identify depression is also suspect. Antidepressants are often prescribed for other indications such as anxiety, and depression is not always treated with antidepressants.

Assessment of Gaps

Overall, the report authors concluded that administrative data cannot reliably identify depressed persons, but that the majority of the evidence suggests that people identified by administrative data as depressed do have depression. The authors suggest several priorities for future research, including a replication study of the most promising algorithm to identify clinically diagnosed depression, assessment of this algorithm in comparison to an independent standard in settings which conduct routine screening for depression (though screening tools have limitations), and research on algorithms to identify depression in youth. Given that several studies have been conducted examining algorithms to identify clinically diagnosed depression, future validation studies might be considered of moderate priority. However, the potential importance of this outcome in youth and lack of studies in this patient subgroup may raise its priority level.

The unfortunate reality is that no algorithm using administrative data is likely to be sensitive to depression given that it is under-diagnosed in clinical practice. Another consideration is that if drugs are examined as potential causes of depression, the people taking them may have increased contact with the medical system raising the chances that their depression may be recognized. Their disease states may also contribute to depression. Studies of the relationship of exposures to this outcome may try to capitalize on data from sites that conduct routine electronic screening for depression. Results in subgroups with screening results available could be compared to those in the study groups in which only usual administrative data is available as a type of sensitivity or confirmatory analysis.

G. SUICIDE AND SUICIDE ATTEMPTS

Walkup and colleagues identified five studies examining the validity of algorithms to identify suicide or suicide attempts. The studies were very different.

Three studies examined the validity of External Cause of Injury codes (E-codes) for intentional self-harm. One examined the agreement of intention category (i.e., unintentional, intentional, assault, or undetermined) between E-codes and expert reviewers in 533 suspected drug overdoses or poisoning cases presenting to two emergency departments in urban hospitals. Agreement was found in 32.1% of cases. Another examined the proportion of 181 discharges with deliberate self-poisoning E-codes that were confirmed by the medical record or expert review. Medical records confirmed 36.5% of cases and expert review confirmed 59.5% of cases. The last study examined new users of antidepressants for treatment of depression. Intentional self-harm and suicidal intent were examined for E-codes indicating intent or undetermined intent. Intent was confirmed in 100% of 30 cases where an E-code indicated intentional self-harm or suicidal intent.

Two studies examined the ability of administrative data to correctly identify suicide. One captured suicides from hospital databases and death certificates. This study found that 59.5% of hospital stays with intentional self-harm codes that ended in death were considered suicides. This algorithm had a sensitivity of 65.0%. The other study examined the ability of hospital or emergency department records to capture suicides that were found from death certificates, coroner reports, and law enforcement reports. Their algorithms cast a broader net to examine medical encounters in suicide victims, looking for either: 1) any hospital or emergency department encounter within one day of death, or 2) the

proportion of subjects with a suicide related event during the year. The sensitivity of the first algorithm was 13.8% and the sensitivity of the second was 14.3%. This study highlights the challenges of identifying suicide through administrative healthcare data, since many victims may die prior to obtaining medical attention.

Assessment of Gaps

The authors concluded that it is difficult to recommend any single algorithm to identify suicide or suicide attempts. E-codes are inconsistently coded. When they are coded, their ability to capture outcomes may be highly dependent upon local coding practices. They recommend several priorities for future research. The first is development of expected norms of rates and distributions of E-codes by age, sex, and treatment setting to identify health care systems with acceptable levels of E-code completeness. The second is to develop standardized procedures for validating E-codes that are not limited to medical records. They also recommend gathering more information on PPVs of algorithms to capture these outcomes, including variability by treatment setting, age, and sex. Finally, they recommend research on software programs to reliably identify suicidal behaviors through text mining of electronic medical records.

Based on the limited evidence available, suicide and suicide attempts might be considered a higher priority area for further validation studies. However, the limitations of the available data need to be carefully considered in the design of such studies. E-codes are not necessarily a reliable source of self-harm data, and they will never capture those cases that do not reach medical care. Linkage of administrative data with death certificates would add substantial value in any attempt to identify completed suicides.

H. SEIZURES, CONVULSIONS, OR EPILEPSY

Kee and colleagues identified 11 studies that reported on the validity of algorithms to identify seizures, convulsions, or epilepsy. Performance characteristics varied widely, but the algorithms in adult populations generally had PPVs around 80% or higher. One algorithm that required only a diagnosis, Current Procedural Technology (CPT) code for an electroencephalogram (EEG) or anticonvulsant drug assay level, or an anticonvulsant medication prescription had a very low PPV of 32.7%. Another algorithm that required a diagnosis code or an EEG procedure code had a similarly low PPV of 21%. This poor performance would be expected given the broad net cast by these algorithms, which did not even require a diagnosis code but appeared to be attempts to sensitively identify potential cases. Four studies examined seizures or convulsions after childhood vaccinations. The PPVs in two of these studies that reported results in aggregate for all subjects were 65.4% and 39.0%. The third study examined differences in PPV in a variety of subgroups, as well as by the location of the medical encounter and time since vaccination. Generally, they found that emergency department diagnoses had excellent PPVs (>90%), inpatient diagnosis PPVs ranged from 59.7-79.1%, and outpatient codes had extremely low PPVs, particularly if the diagnosis occurred on the same day as the vaccination. A final study used only inpatient and emergency department codes to identify seizures and found an overall PPV of 94%, suggesting that this may be a more efficient method to identify cases that are likely to be real.

Overall, it seems that algorithms to determine epilepsy perform somewhat better than those to identify seizures or convulsions. This makes sense as single seizure or convulsion events can be hard to confirm while epilepsy diagnoses depend on multiple recurring events. However, restricting the outcome definition to emergency department or inpatient seizure or convulsions diagnoses appears to produce

an algorithm with an acceptable PPV, at least in young children, even if sensitivity is sacrificed to some degree.

Assessment of Gaps

Algorithms that perform quite well in many settings have been developed to identify epilepsy. While a number of studies have examined seizures or convulsions, results suggest that the PPVs for these codes are not as high as those for the epilepsy codes. In children receiving vaccines, the setting of the diagnosis and timing compared to the vaccine appears to be an important factor influencing PPVs. For the purpose of active surveillance of medical product-related HOIs, it seems likely that seizures or convulsions will be studied in relation to various exposures, as opposed to epilepsy. Given the available data and questionable performance of these codes, it may be that any study examining this outcome will need to validate the outcomes in source data to ensure that an event actually occurred. Alternatively, the study might validate a subset of outcomes and use a correction factor for the proportion of false positive cases. Any correction factor should take into account differences in performance characteristics by factors such as age, sex, setting of the diagnosis, or a history of epilepsy. It is difficult to prioritize this outcome for a validation study unrelated to a study of an exposure and outcome, as a number of studies have been done that illustrate limitations in algorithms. The major caveat is that emergency department and inpatient diagnosis codes appear to perform quite well in identifying seizures in young children who have received vaccines. Requirements of EEG codes may increase PPV but are likely to reduce sensitivity, since not all isolated seizure events will be confirmed by EEG.

I. PANCREATITIS

Moores and colleagues identified 9 studies that reported on the validity of algorithms to identify pancreatitis. PPVs ranged widely, but most studies found PPVs in the 60-80% range. Algorithms using only ICD-9-CM code 599.0 to identify acute pancreatitis generally had similar PPVs in the 60-80% range, though the total range was 40-97%. The lowest PPV was found in an outpatient detoxification program in a Department of Veterans Affairs healthcare system. It is possible that diagnostic suspicion was greater in this relatively high-risk group, or that alcohol-related gastritis may have been mistaken for pancreatitis more frequently in these patients. One study examined an inpatient pediatric population and found a PPV of 76.2%.

Assessment of Gaps

Though a number of validation studies have been conducted in a variety of populations, most studies were conducted in single centers or very restricted populations. It may be useful to conduct a validation study in a less restricted population including many health systems to get more generalizable information. There is more information on the validity of algorithms for this outcome than for a number of others, but some gaps remain. It may also be useful to determine the increases in performance that might be achieved in settings where results from laboratory tests strongly suggestive of pancreatitis could be linked with administrative data.

J. LYMPHOMA

Herman and colleagues identified only one study that reported on the validity of algorithms to identify lymphoma. The study examined four different algorithms for a number of cancers, including lymphoma. An algorithm that only required one diagnosis code had a PPV of 34.7%, the lowest of the four algorithms. An algorithm requiring two diagnosis codes within two months had a PPV of 62.8%, the highest of the four algorithms. Sensitivity ranged from 55.2-88.7%. A major limitation of this study was that the reference standard for validation of the outcome was a state cancer registry. While this would seem to be a very useful and convenient method of validation, it is highly dependent on the ability of the cancer registry to identify all cases. Cancer registries often identify cases through pathology reports, so their ability to identify all hematologic malignancies is questionable. Solid tumors are likely better identified.

Assessment of Gaps

Given that only one validation study has been conducted and it utilized a single state cancer registry, this outcome might be classified as higher priority for future validation studies. The information on sensitivity obtained from the sole validation study is likely more accurate than the information on PPV and specificity given the potential limitations of the cancer registry reference standard. Despite the fact that the PPVs were lower than for some outcomes, the validation study authors simulated a pharmacoepidemiologic study with lymphoma as the outcome and determined that little bias would result from the limitations of the algorithm given the relative rarity of lymphoma. Regardless, a study that performed medical record review to determine the validity of the outcome definition would be more conclusive. To increase efficiency of such a study, it may be useful to use a registry to identify true positive cases, but also to review the medical records of potential cases that were not identified by the registry.

K. INFECTION RELATED TO BLOOD PRODUCTS, TISSUE GRAFTS, OR ORGAN TRANSPLANTATION

Carnahan and colleagues identified only one validation study of an algorithm to identify infections in recipients of blood products, tissue grafts, or organ transplants. This study examined the validity of an algorithm to identify aspergillosis in transplant recipients. Sixteen studies were identified that studied infections, broadly or specifically defined, in these patient groups. None made any clear attempt to identify infections specifically transmitted by the blood product, tissue graft, or organ transplant. Most appeared to focus on infections related to immunosuppression in transplant recipients or related to a surgical procedure. Using administrative and claims data to determine whether infections were actually transmitted by a blood product, tissue graft, or organ transplant would seem to be very difficult. These patients are already at high risk of infection for a variety of reasons. Adding certain ICD-9-CM codes in the 996.6x range that identify infections and inflammatory reactions related to implants and grafts may prove useful for increasing the specificity of the outcome measure, though this has not been validated. The ICD-9-CM codes 999.3x may also be useful for identifying such infections, as they capture infections related to medical care not elsewhere classified. A new code for acute infection following transfusion, infusion, or injection of blood and blood products (999.34), to be implemented on October 1, 2011, may also add specificity to an infection transmitted by a transfusion, though the description does not specifically indicate that the infection was transmitted by the blood or blood product.

Assessment of Gaps

Given the near complete lack of validation studies on this topic, this might be considered a higher priority for future algorithm validation work. Special consideration should be given in considering whether the interest is to identify infection outcome measures in these patients versus evaluating whether the infection is transmitted by the blood product, tissue graft, or organ transplant. The latter is likely to be especially difficult given the clinical ambiguity of the source of many of these infections.

L. TRANSFUSION-ASSOCIATED SEPSIS OR SEPTICEMIA

Carnahan and colleagues identified no studies that reported on the validity of algorithms to identify transfusion-associated sepsis or septicemia. Since no such studies were identified, they instead reported on four validation studies of sepsis and two validation studies of transfusion codes.

For sepsis, one study examined all patients at a single hospital, two multi-site studies examined post-surgical populations [one in children and another in Department of Veterans Affairs medical centers (VAMCs)], and one evaluated only the sensitivity of codes in a group of patients from 10 hospitals in a clinical trial of a treatment for severe sepsis. The PPV of ICD-9-CM codes 038.x was 88.9% in the single center study, and the negative predictive value (NPV) was 80%. The study in children used a broader range of codes and found a PPV of 79.9%. The study in VAMCs found a PPV of 44%, sensitivity of 32%, and specificity > 99% for ICD-9-CM codes 038.x. An algorithm that used more codes had slightly better sensitivity and a similar PPV. National Surgical Quality Improvement Program data was used as the reference standard. It is possible that the data comprising this reference standard had limitations resulting in the low PPVs. It is also possible that the limited incentives for accurate billing at VAMCs affect performance characteristics of diagnostic codes. Finally, the sensitivity of a set of codes in the study of clinical trial participants was 75.4%. Overall, the ICD-9-CM codes 038.x had relatively good performance in most settings, while less information is known about other codes indicative of sepsis or septicemia.

Only one single center study and one other multi-center study were identified that validated algorithms to identify transfusion, and they focused on a single code for allogeneic red blood cell transfusions. Performance at the single center was quite good, with a sensitivity of 83% and a specificity of 97.5-100% depending on whether a subset of patients without revenue codes for transfusion were classified as true negatives or false negatives. Patients not billed for transfusion were less likely to have commercial insurance, suggesting that likelihood of reimbursement may affect coding practices. The multi-center study included discharge abstracts from 1987, and found good specificity (100%) but poor sensitivity (21% or 31% depending on the number of diagnosis fields considered in hospital discharge abstracts).

Assessment of Gaps

Two out of three studies that examined the PPV of sepsis codes found that codes 038.x performed reasonably well to identify sepsis. Sensitivity was reasonable in one study but poor in the VAMC study that also found a poor PPV. Identifying sepsis specifically related to a transfusion is a much more difficult proposition, particularly since determining such a cause can be ambiguous even in the clinical setting. A new ICD-9-CM code adopted for acute infection following transfusion, infusion, or injection of blood and blood products (999.34), to be implemented on October 1, 2011, might add value in increasing the specificity of such an outcome measure. If an algorithm is proposed that is actually likely to identify

sepsis specifically related to transfusion, a validation study might be considered a higher priority based on lack of evidence.

With regard to transfusions, a relative lack of data exists to comment on the performance characteristics of algorithms. While one single-center study found that the code for allogeneic red blood cell transfusion performed quite well, another somewhat outdated multi-center study found good specificity but poor sensitivity. Transfusions may be under-coded for a number of reasons. More current data on performance characteristics of this code would be useful. In addition to what's been studied, there are several other transfusion codes of interest. In particular, it would be helpful to understand the validity of platelet transfusion codes, since this has historically been a transfusion with a high risk of infection. Microbial growth in platelets is more likely due to room temperature storage. Rapid tests to identify such growth prior to transfusion have been developed, but it is yet to be seen whether these can eradicate the problem of platelet transfusion-associated infections. Thus, more information on the validity of a wider range of transfusion codes would be desirable.

M. TRANSFUSION-RELATED ABO INCOMPATIBILITY REACTIONS

Carnahan and colleagues identified no studies examining the validity of ABO incompatibility reaction codes. One study did perform medical record review to validate transfusion reaction codes, but none of these codes was for ABO incompatibility reactions. These reactions can occur as a result of transfusions or transplants. The lack of such codes identified in one study utilizing a database of 2.23 million hospital discharge abstracts in which a transfusion was given raises questions about the sensitivity of such codes. However, FDA researchers have identified such codes in CMS data (personal communication). Determining the sensitivity of these codes would be extremely difficult given the rarity of ABO incompatibility reactions. One might speculate that the ICD-9-CM code 999.6 for ABO incompatibility reactions would be quite specific due to limited ambiguity in the diagnosis, but this is not supported by any data.

Assessment of Gaps

Since no validation studies have been conducted for ABO incompatibility codes in administrative data, this outcome might be considered high priority for validation studies based on the criterion of available data.

N. ERYTHEMA MULTIFORME, STEVENS-JOHNSON SYNDROME, TOXIC EPIDERMAL NECROLYSIS

Schneider and colleagues identified four studies that examined the validity of algorithms to identify erythema multiforme, Stevens-Johnson syndrome, or toxic epidermal necrolysis. Two used the same cohort, so three unique validation studies were described. The report authors noted that the coding for these conditions has been modified in recent years, which reflect increasing understanding of the differences among these serious dermatologic conditions. They calculated PPVs to reflect the proportion of cases that accurately reflected the conditions represented by the algorithms at the time the studies were conducted, which also included staphylococcal scalded skin syndrome. The PPVs ranged from 61-66%. The most recent data from these studies was from 1986, suggesting a need for more current information on the performance of the relevant codes. The studies were based in large populations,

with three out of four studies using a multi-state Medicaid database and one using a regional private health plan database.

Assessment of Gaps

All studies used ICD-8 or ICD-9-CM code 695.1, which captured all the outcomes of interest. It is possible that other general codes for hypersensitivity reactions, adverse drug reactions, or allergic reactions might be used to document this diagnosis. The performance of such alternative codes to identify these outcomes is unknown. Presumably, the addition of such general codes might enhance sensitivity but would greatly reduce specificity, and it's likely that the sample size required to identify these outcomes using these codes would be too high. It may be useful to determine the performance of specific sub-codes that fall under 695.1, particularly if the goal of a study involves delineating these related but different outcomes. Since all the information on the performance of these codes is dated, it may be considered higher priority to conduct a validation study using more current data. If the performance of the codes is no better than in the past, it may be necessary to validate all events in a study of these outcomes to ensure the validity of any relationship with a drug exposure that is identified.

O. ANAPHYLAXIS, INCLUDING ANAPHYLACTIC SHOCK AND ANGIOEDEMA

Schneider and colleagues identified 4 studies that provided validation statistics for anaphylaxis, anaphylactic shock, or angioedema, out of 8 studies that conducted some type of validation. In one study, the code for anaphylactic shock (ICD-9-CM code 995.0) performed better than other codes in identifying anaphylaxis, with a PPV of 55% for identifying probable or possible cases of anaphylaxis. The PPV of this code was 57.1% in another study. The second highest PPV for codes to identify anaphylaxis was 10%, for a code that identified anaphylactic shock due to an adverse food reaction (ICD-9-CM code 995.6). Another study found that an angioedema code (ICD-9-CM code 995.1) had a PPV of 7.4% in predicting anaphylaxis. Some studies used expanded list of codes and identified a small number of cases of anaphylaxis among diagnoses that were possibly related to anaphylaxis (e.g., allergic urticaria or allergy unspecified). Thus, codes specific to anaphylaxis do not appear to capture all cases, and most are non-specific for identifying true anaphylactic reactions. Even the best potential algorithms have a lower PPV than is generally desirable. Other algorithms that include broader sets of codes appear to have much lower PPVs.

In addition to the studies reported in the review, a Google Scholar search for hypersensitivity reaction algorithms identified another study that reported validity statistics for a definition of drug-related anaphylaxis or hypersensitivity reactions. This study examined the frequency of such reactions in children and adolescents in a state emergency room hospital discharge database, and explored several algorithm iterations (West, et al. 2007). The study was based on 63 possible cases. The original algorithm in this study had a PPV of 32% for identifying drug-related anaphylaxis. Refinement of the algorithm by reclassification of certain codes increased the PPV to 56% but lowered the sensitivity. However, this PPV was calculated after dropping the requirement that the reaction was determined to be drug-related in the medical record review. One difference between this and other algorithms involved the use of E-codes. Overall, it does not change the conclusions of the report, though it raises the question of whether E-codes can improve the performance of a coded algorithm to identify anaphylaxis.

In contrast, the code for angioedema appeared to perform quite well for predicting angioedema in two studies, with PPVs of 95.3% and 90%.

Assessment of Gaps

Generally, codes for anaphylaxis appear to perform at a level that is less than that desired. Even the best performing code, anaphylactic shock, performs less than optimally. It may be necessary to confirm every case through medical record review in studies of this outcome. It may also be beneficial to include a broader set of codes to capture more cases, though care should be taken to prevent major losses in efficiency from selection of an algorithm with too many codes. Another approach would be to estimate the bias in the risk estimates based on PPVs calculated in prior validation work, and only perform validation if it is likely to substantially influence the interpretation of results. Further research may further delineate codes appropriate for identifying potentially drug-related anaphylaxis, as opposed to that from other causes. Such research might examine the value of using E-codes for this purpose.

The two studies that examined the PPV of angioedema codes found PPVs of 90% or above, so this may be a lower priority outcome for future validation studies despite the limited number of studies that have been conducted.

P. HYPERSENSITIVITY REACTIONS OTHER THAN ANAPHYLAXIS

Schneider and colleagues identified 5 studies that reported on the validity of algorithms to identify hypersensitivity reactions other than anaphylaxis, and 2 that reported validation processes but no validation statistics. Two of the studies that reported validation statistics were focused on angioedema. They found PPVs > 90%, as previously described. Of the two studies that did not provide validation statistics, one studied the outcome of angioedema in angiotensin-converting enzyme (ACE) inhibitor users and the other studied injection site reactions, allergic responses, or seizures after vaccine administration.

One study of serious allergic reactions after antibacterial drug exposure reported a PPV for anaphylaxis and resuscitation or adrenaline injection procedure codes, but not for the codes used to identify unspecified allergy or an unspecified adverse effect of a drug. Thus, it provides little useful information on identification of less severe allergic reactions. Another study examined a long list of candidate codes to predict the presence of abacavir-associated hypersensitivity reactions among new users of the drug. They found several key predictors including diagnosis of several specific symptoms commonly associated with the reaction, a claims diagnosis of adverse effect of a drug, anaphylactic shock or unspecified allergy, and discontinuing abacavir prior to 90 days of therapy. The algorithm had 95% sensitivity and 90% specificity. Unfortunately, the focus of this study on a specific drug with a fairly well characterized hypersensitivity syndrome limits its utility to inform the usefulness of algorithms to identify hypersensitivity reactions to other drugs.

The aforementioned study that used E-codes and other codes to identify emergency department visits for drug related anaphylaxis or hypersensitivity reactions (West, et al. 2007) was also included in this report. It might be used to develop hypotheses for future research, but the number of non-anaphylaxis cases was quite small. This study was identified in Google Scholar searches that identified 4 results using the terms ‘hypersensitivity icd “predictive value” 708.0,’ and 5 results when the final ICD-9-CM code was changed to 995.3. These ICD-9-CM codes represent ‘allergic urticaria’ and ‘allergy unspecified,’ respectively. It appears that 2 out of 13 cases classified as ‘other drug-related allergic reactions’ by the

algorithm actually had non-anaphylactic drug-related allergic reactions, for a PPV of 15%. These numbers are from results reported in a sensitivity analysis after dropping the requirement that a reaction was determined to be drug-related, as this was often difficult for the clinicians to determine when reviewing medical records.

Assessment of Gaps

Very few useful data are available to support the validity of algorithms to identify hypersensitivity reactions other than anaphylaxis or angioedema. This outcome might be considered higher priority for validation studies, except for the fact that this category of hypersensitivity reactions generally seems to represent a less serious set of outcomes compared to the others, e.g., urticaria is highly preferable to anaphylaxis. Research might explore the differential utility of specific codes for hypersensitivity reactions, and also consider the value added by including E-codes. Such research may also need to consider the trade-offs associated with including E-codes, which are not consistently coded across settings of care.

Q. PULMONARY FIBROSIS AND INTERSTITIAL LUNG DISEASE

Schneider and colleagues identified no studies that reported on the validity of algorithms to identify pulmonary fibrosis (PF)/interstitial lung disease (ILD). They found 5 studies that provided codes used to identify PF/ILD, though one focused exclusively on asbestosis. The algorithms for identifying PF/ILD generally used the same codes, though they differed in whether they used outpatient codes or only inpatient codes, and whether they also required codes for appropriate diagnostic procedures to confirm cases.

Assessment of Gaps

Given that no studies are available to determine the validity of algorithms to identify PF/ILD, this outcome may be considered high priority for validation studies. The ICD-9-CM codes used to identify the outcome were fairly consistent across studies. However, research might examine the impact of using only inpatient versus inpatient plus outpatient codes or requiring diagnostic procedure codes consistent with confirming the outcome. Unfortunately, the presence of a diagnostic procedure code does not mean that the procedure confirmed the diagnosis. These codes can only indicate that the procedure was conducted. To address this, the algorithm might also ensure that the diagnosis code of interest was the last related code identified during a certain time period after the index code, to support that the diagnosis was not a rule-out diagnosis later determined to be something else. One study did this, presumably because of potential diagnostic ambiguity upon initial presentation of the disease. Finally, the clinician reviewer noted that PF is typically considered a subtype of ILD, and ILD actually represents a large number of discrete diseases and conditions. The two major codes used to identify the outcome, 516.3 and 515.x, were described as too narrow and too broad, respectively, to capture all diagnoses of interest. Codes for this condition may vary by the provider specialty or health care setting.

Any studies of this outcome should carefully consider the specific codes of interest through consultation with a clinical expert in the field. If the available codes are not classified in such a way as to suggest an ability to capture all cases of interest with reasonable specificity, it may be that combining a more sensitive case identification algorithm with outcome validation is necessary.

R. ACUTE RESPIRATORY FAILURE

Schneider and colleagues identified no studies that reported on the validity of algorithms to identify acute respiratory failure. They found two studies that used administrative data to identify acute respiratory failure, both of which used ICD-9-CM code 518.8. It was noted that 518.81 may be more specific, since it refers to 'acute respiratory failure,' while 518.8 identifies a variety of conditions including chronic respiratory failure and unclassified lung diseases.

Because of the lack of studies identified in the original search, a Google Scholar search was conducted for this gap analysis on January 13, 2011, to help ensure that important studies were not missed. The search string was the following: 'respiratory failure "predictive value" 518.8.' This search identified 11 results, one of which was a study of acute respiratory distress syndrome (ARDS) that referenced 3 hospital-based studies that examined the performance characteristics of ICD-9-CM coding algorithms to identify ARDS (Moss and Mannino 2002). Two of these studies were published only as abstracts. The algorithms used were sometimes restricted to ICD-9-CM codes 518.81 and 518.82 (other pulmonary deficiency not elsewhere classified) rather than using the entire range of 518.8x. They also sometimes included 518.5 (pulmonary insufficiency following trauma or surgery). The PPVs ranged from 7% to 38%. One study found a sensitivity of 88% and specificity of 99% in a single hospital participating in a clinical trial to treat ARDS. (Thompsen, et al. 1992) The same group screened for ARDS in 5 other hospitals and found that ICD-9-CM codes had a sensitivity of 79%, specificity of 98%, and PPV of 4% in those hospitals (Thomsen and Morris 1995). Another study screened medical intensive care unit (ICU) medical records at a university hospital for patients that met ARDS criteria, and found that only 4/65 patients who met criteria had an ICD-9-CM code for the diagnosis. Only 31/65 had ARDS documented in the medical record. (Howard, et al. 2000) A final study prospectively screened 13,398 admissions at a single hospital for ARDS and found 323 patients who met consensus criteria. The clinical assessments were compared with ICD-9-CM coding. They used three ICD-9-CM algorithms that were not defined in the abstract in which the results were available. Sensitivity ranged from 28-74%. Specificity ranged from 97-100%. False positive rates ranged from 0.3% to 3%. False negative rates among the 323 patients ranged from 26% to 72%. (Rubenfeld, et al. 1998)

Assessment of Gaps

The few studies that were identified suggested that algorithms to identify ARDS, a subtype of acute respiratory failure likely of particular interest, had low PPVs. Future research might explore the PPVs of specific sub-codes to determine whether an algorithm with acceptable performance characteristics could be identified. Alternatively, the low PPVs found in research to date suggests that it may be necessary to validate cases in studies seeking to identify this outcome. An additional concern, however, is that sensitivity of ICD-9-CM codes to the condition appears to be highly variable. As one study suggested extremely low sensitivity of ICD-9-CM codes, it may be beneficial for future studies to re-examine the sensitivity of algorithms to identify this condition. Given the limited data available and questions that remain, this outcome might be considered higher priority for future validation studies.

S. ORTHOPEDIC IMPLANT REVISION AND REMOVAL

Singh and colleagues identified 5 studies that examined the validity of algorithms to identify orthopedic implant revision. None specifically evaluated implant removal. Two studies reported on total knee arthroplasty (TKA) revision and three studies on total hip arthroplasty (THA) revision.

For TKA revision, one study in Ontario found a sensitivity of 77.7%, specificity of 97.6%, PPV of 66.9%, and NPV of 98.6%. However, the reference standard in this study was physician fee-for-service claims as opposed to medical record review, so the results should be interpreted with caution. Another study using MEDPAR (Medicare part A) data found a sensitivity of 87.2% and specificity of 99.0%. This study also did not use medical record review for validation, but rather used a new code specific to revision as a reference standard for an algorithm that did not include this code. Thus, data on the validity of TKA revision algorithms is limited by questionable reference standards.

Three studies examining THA all used Medicare data from July 1995 to June 1996. Despite the use of what appears to be the same data, different samples of medical records were reviewed for each study. The PPVs for revision THA were 92% in a sample of 671 patients, 91% in a sample of 550 patients, and 71% in a sample of 374 THA patients. The last study with a lower PPV appeared to have used ICD-9-CM and not CPT codes, while the other two studies included CPT codes in the algorithm. More importantly, this study also required that the revision was on the same side as the index THA in the study. This likely explains the reduced PPV compared to the other studies.

Assessment of Gaps

With the exception of the study conducted in Ontario, all the reported studies were conducted in Medicare patients. The PPV was consistently high for revision THA except in the one study that did not appear to use CPT codes and also imposed the requirement that the revision was on the same side as an index THA. This would likely place this outcome in the low-moderate priority range for future validation studies. It would be helpful to have more information on the performance of these algorithms in non-Medicare populations, including younger patient groups. The lack of studies examining implant removal may place this particular type of revision in the higher priority range based on the criterion of available evidence.

In contrast, the codes for TKA did not appear to perform as well, and both studies used questionable reference standards for validation. Future research on algorithms to identify TKA is recommended.

III. DISCUSSION

The systematic reviews overviewed provide a wealth of information about algorithms to identify various health outcomes in administrative data. They also identify substantial gaps in evidence to support algorithms to identify a number of outcomes and many opportunities for future research. For some outcomes with lower PPVs, it may always be necessary to validate potential cases identified in administrative data, unless algorithms can be altered to enhance specificity.

A. PRIORITIZATION OF RESEARCH

Based on the amount of evidence available, consistency of findings, and the performance characteristics of algorithms studied, the outcomes can be broadly classified into a number of categories. Those HOIs with algorithms that perform consistently well across a number of studies might be considered lower priority for future validation studies. These include cerebrovascular accident; transient ischemic attack; congestive heart failure; deep vein thrombosis; pulmonary embolism; angioedema; and revision of total hip replacement. Other outcomes with a fair amount of evidence but less consistent findings on the validity of algorithms might be considered of moderate priority for future validation studies. These

outcomes include atrial fibrillation; serious cardiac arrhythmias; depression; seizures, convulsions, or epilepsy; and pancreatitis. Studies with little evidence that is relatively inconsistent, or for which algorithms performed poorly, might be considered of higher priority for future validation studies. These include suicide and suicide attempts; lymphoma; infection related to blood products, transfusion, tissue grafts, or organ transplants; transfusion-associated sepsis or septicemia; transfusion-associated ABO incompatibility reactions; hypersensitivity reactions other than anaphylaxis; pulmonary fibrosis and interstitial lung disease; total knee replacement revision; and orthopedic implant removal. Finally, two outcomes have limited evidence that suggests algorithms may perform quite poorly. This might place them in a category that is high priority for future algorithm validation studies, or one in which all potential cases should be consistently confirmed due to poor performance of algorithms. The outcomes that fall into this category are anaphylaxis and acute respiratory failure. The only evidence on the latter outcome relates to ARDS, so there may be too little evidence available to conclude for certain that a well-performing algorithm could not be designed. Codes for anaphylaxis generally had low to moderate PPVs, with the code for anaphylactic shock having the only moderately acceptable PPV.

B. ICD-10-CM ALGORITHMS

Another consideration for future research is whether validation studies have been conducted for algorithms using International Classification of Diseases, 10th edition, Clinical Modification (ICD-10-CM), codes. This coding system must be adopted in the United States by October, 2013, so research using administrative data will have to adapt accordingly. A small number of reports included studies that examined the validity of algorithms using ICD-10-CM codes. These included seizures, convulsions, or epilepsy (1 study); anaphylaxis (1 study); congestive heart failure (2 studies); and cerebrovascular accident or transient ischemic attack (2 studies). Though mapping across coding systems and generalizing ICD-9-CM algorithm validity to ICD-10-CM codes might be reasonable for some outcomes, many algorithms will need to be reconfigured and revalidated when ICD-10-CM coding is more completely implemented in the United States. It is also possible that validity during the initial period of transition may be different because of the learning curve in implementing ICD-10-CM.

C. ALGORITHMS PROVIDING HIGH LEVELS OF CONFIDENCE

It is difficult to determine the threshold of evidence required to decide that an algorithm is ready for use in high-stakes studies related to medical product safety. Also, there is no threshold positive predictive value or other validity statistic that can be chosen to make this decision. The potential bias resulting from imperfect validity is dependent on other factors such as the prevalence of the outcome. If the parameters are known, it is possible to adjust risk estimates based on factors such as the positive predictive value. Ultimately, it is inefficient to validate every case of an outcome that is identified in administrative data. Even outcome validation studies using the source data typically only validate a subset of cases and could be subject to sampling bias. At some point it is necessary to make the decision that enough evidence supports an algorithm for it to be used without another validation study.

When reviewing the evidence on the HOIs reviewed in this document, a number of factors need to be considered. For some HOIs, there may be a limited set of logical algorithms to select from, such that the choice of algorithm is relatively obvious despite its limitations in performance. For others, a variety of algorithms were studied with varying levels of validity. Algorithms with consistently high positive predictive values, above 70% for the purpose of this discussion, across a number of studies might be considered ready for application with a high level of confidence. A number of HOIs have algorithms that

meet this criterion of relatively good performance. In this author's view, the following outcomes meet these criteria: cerebrovascular accident, transient ischemic attack, ventricular arrhythmias, congestive heart failure, deep vein thrombosis, pulmonary embolism, and angioedema. Some outcomes might be considered on the cusp of having enough evidence to support use with confidence. Atrial fibrillation algorithms performed fairly well, but only two studies examined algorithms to identify incident atrial fibrillation, the likely outcome of interest in a study of drug adverse effects. Three studies found high PPVs for total hip arthroplasty revision algorithms, but all used Medicare data so more evidence in other populations may be needed. For the rest of the outcomes, algorithms may have been identified that performed well in some samples, but the strength of the evidence did not seem adequate to support conducting safety studies without accompanying algorithm validation studies. It is important to qualify this statement in that it is only one reader's view of the evidence. End-users of this information will have to make their own judgments on the quantity and quality of evidence that supports potential algorithms to identify HOIs.

D. LESSONS LEARNED

The main purpose of this section is to describe decision making processes and lessons learned during this project, primarily related to the search strategies for the reports. For a number of HOIs, MeSH terms were exclusively used to identify the HOI in PubMed despite awareness that text word searches might identify additional relevant citations. Text word searches in PubMed search the title and abstract for the stated term. This decision was simply a matter of making trade-offs between the resources available, timeline for completion, and the desire to be thorough. The increase in citations identified when text word searches were used, though variable, could be quite substantial. For some HOIs, including text word searches identified more than 1,000 additional citations. While no specific threshold was used, when text word searches would have led to substantial additional workload they were generally not used for the final search. If the additional workload was on the scale of several hundred abstracts or less, or the total number of citations was relatively smaller compared to other HOIs, text word searches were generally included. The omission of text words from a number of searches might be considered among the most substantial limitations of the reports.

Other limitations related to the indexing of citations. For example, some manuscripts are indexed in PubMed as randomized controlled trials when in fact they are not. Some manuscripts discuss a health outcome, but are not indexed with a corresponding MeSH term. In some cases this is because a manuscript discusses multiple health outcomes, and in other cases the rationale is less clear. It is important to recognize that citation indexing is limited by the potential for human error or subjective determinations of the most important topics within an article to index. In yet other cases, some manuscripts such as those evaluating comorbidity indices may perform validation studies on a large number of health outcomes. In this type of study the HOI may not even be listed in the abstract, such that it would be nearly impossible to identify the study in a PubMed search focused on that HOI. Another limitation is that there seems to be no standard convention for indexing studies that validate algorithms for identifying HOIs. In reviewing studies missed by several HOI searches but identified through other means, such as the references of reviewed manuscripts, the indexing varied widely. This issue is compounded by the fact that some studies of interest focus specifically on validation of algorithms while others include an algorithm validation study within the context of a study examining risk factors for an HOI. Thus, it seems that there is no perfect search strategy for ensuring the identification of all the studies of interest in a systematic review focused on algorithm validity that relies on citation indexing databases.

With regard to the sensitivity of the search strategy, the combined Observational Medical Outcomes Partnership (OMOP) reports identified 13 studies that examined the performance of algorithms to identify acute renal failure (Jarret, et al., Kachroo, et al.). The Mini-Sentinel PubMed search strategy, combined with MeSH terms representing acute and chronic renal failure, identified 7 of those studies, and 3 additional studies were cited in studies identified in the search. The addition of chronic kidney disease MeSH terms led to a much larger number of results (1,078 vs. 111 with 'kidney failure, acute' and 'renal insufficiency, acute' MeSH terms only), but appeared appropriate since several studies included in the OMOP reports were focused on chronic kidney disease. Of the three studies identified in the references of manuscripts in the search, one would have been identified if text word searches for acute renal or kidney failure had been included, one was improperly indexed as a randomized controlled trial and thus excluded, and one was an anomaly that seemed to meet all the criteria for the search but was not identified for reasons that could not be determined. Of the three studies that were not identified, all would have been identified by including text word search terms for the HOI.

How could the limitations of the search strategy and requirements for balancing workload with sensitivity be overcome? One possibility is to increase the number of databases searched, though this increases workload. For these reviews, Iowa Drug Information Service (IDIS) searches were conducted. Only articles relating to drugs are indexed in the IDIS database, so these searches identified much smaller numbers of citations compared to PubMed searches. No specific examination of the value gained by adding this additional search was conducted for this project. Embase searches may identify additional citations, and these were conducted for a number of HOIs for which PubMed searches resulted in a relatively small number of citations. Embase searches often result in a much larger number of citations, so there is an additional trade-off in efficiency. Embase indexes each section of manuscripts rather than focusing on the title and abstract, as is the case with PubMed. The changes in efficiency that would result from adding Embase also were not specifically examined for this project, and presumably would vary by HOI. One limitation of using Embase is that it is a proprietary database with substantial cost, in contrast to PubMed which is freely available. A final method of searching for studies that may have particular value is to use a text mining search strategy as can be conducted with Google Scholar.

Google Scholar was found to be particularly valuable in searching when few or no relevant studies were identified for an HOI. The strength of Google Scholar comes in its ability to conduct text word searches of entire manuscripts as long as they are available on the internet. This allows for very specific searches for relevant terms such as "predictive value" and "International Classification of Diseases" or "ICD." If it is highly likely that a particular code would be contained within algorithms relevant to an HOI, it might be recommended that this code be included in the Google Scholar search. This greatly enhances the specificity of such searches, which is helpful since they can sometimes produce many thousands of results. For example, at the time of this writing a Google Scholar search for 'respiratory failure "predictive value" icd' identified over 6,000 results. Though combining terms in quotes is helpful, "predictive value" is not necessarily specific since it might be used to describe screening tools, algorithms to predict outcomes, or laboratory tests, for example. The acute respiratory HOI report identified no studies which examined the performance characteristics of algorithms to identify the HOI, but it did find two studies that used International Classification of Diseases, 9th edition, Clinical Modification (ICD-9-CM) code 518.8 to identify the HOI. Consequently, this code was added to the search string. When '518.8' was added to the Google Scholar search string, only 11 results were returned. One of the results cited three studies that examined the performance characteristics of algorithms to identify acute respiratory distress syndrome, two of which were abstracts that would not be indexed in PubMed. So even though the systematic review suggested that no studies had been

conducted, it was possible to incorporate the evidence from these studies into a review of evidence gaps to inform future research. Google Scholar searches were also conducted for several other HOIs for which no relevant studies were identified, such as ABO incompatibility reactions. Future research might explore the efficiencies and number of studies identified in citation indexing databases such as PubMed compared to various approaches to searching Google Scholar. It is possible that the efficiencies of searches and the screening process could be greatly enhanced through the use of Google Scholar. If nothing else, it might be used as a final check to help ensure that easily identifiable studies were not missed due to the limitations of searching PubMed and other citation indexing databases.

A final recommendation for future systematic reviews of this kind is to get input from someone with expertise in each HOI prior to finalizing the HOI-specific search terms to ensure that no relevant terms are missed. Many searches for this project received topic expert review, but in some cases topic experts were identified later in the process of the systematic reviews and did not help with the selection of search terms. While it is not clear yet that any important search terms were missed because of this, including a topic expert early in the process is preferable. In future Mini-Sentinel or Sentinel systematic reviews of algorithms to identify HOIs, it is also recommended that the FDA's content experts be involved in developing and refining search strategies. This may be especially important if the number of search results for a particular outcome is small, suggesting limitations in scope. Their input can help ensure that the scope of the reviews includes all subtypes of an outcome that are potentially of interest.

IV. CONCLUSION

The 19 systematic reviews conducted by Mini-Sentinel investigators provide a relatively comprehensive review of the literature on the validity of algorithms to identify the health outcomes of interest. The reviews identified many useful algorithms, gaps in evidence, and suggestions for future research. Little evidence was found to support algorithms using ICD-10 codes. This overview provides a high level summary of the findings of these reports and one reviewer's suggestions on the relative prioritization of future algorithm validation studies. Ultimately, the selection of outcomes for validation studies in Mini-Sentinel will depend on the FDA's prioritization of future surveillance activities and the Agency's willingness to accept the results of prior validation studies as generalizable to Mini-Sentinel data.

The ideal methods for conducting systematic reviews to identify validated algorithms for identification of HOIs are yet to be determined. While PubMed searches are the basic standard for conducting systematic reviews, any citation indexing database has limitations. It does not appear that there is any consistent method for indexing the types of studies relevant to these reviews. Google Scholar searches may help to overcome some of these challenges. Future work might further explore ideal methods for searching Google Scholar, and the differential performance and efficiency of using various focused Google Scholar searches compared to searching PubMed or other citation indexing databases. Finally, it is recommended that topic experts be consulted early in the search process for optimal results.

Overall, the systematic review process that was enacted appears to have generally resulted in informative reports. It is important to recognize that limitations remain, however, and not all reports contain every study that might have been relevant.

V. ACKNOWLEDGEMENTS

This work was supported by a contract from the United States Food and Drug Administration, FDA HHSF2232009100061. The project would not have been possible without the valuable input and work of many people. I would particularly like to acknowledge the following individuals. Kevin Moores, PharmD, and Ronald Herman, PhD, of IDIS and Jonathan Koffel of the University of Iowa libraries provided advice or worked on designing and conducting searches, managing citations, and producing the abstract review documents. Patrick Ryan, PhD, provided helpful insights into the integrated search strategy developed by OMOP, on which our searches were built. Carol Mita of the Harvard library conducted a number of Embase searches that provided insight on the potential value of this database. Swati Sharma provided essential project management assistance. Elizabeth Chrischilles, PhD; Sean Hennessey, PhD; Darren Toh, PhD; Kimberly Lane, MPH; and Judy Racoosin, MD, MPH, provided important input on many aspects of the project through their work with the Mini-Sentinel Protocol Core. Richard Platt, MD, MSc, provided valuable advice whenever it was requested. Finally, the project would not have been possible without the hard work and input of the HOI report authors and the reviewers who generously donated their time to improving the reports.

VI. REFERENCES

1. Howard AE, Courtney-Shapiro C, Goltz M, Morris PE. Variations in the use of ICD-9 codes in medical ARDS patients (abstract). *Crit Care Med*. 2000; 28(12 suppl): A188.

2. Jarrett N, Lux L, West S. Systematic evaluation of health outcome of interest definitions in observational studies and clinical definitions for the Observational Medical Outcomes Partnership: acute renal failure: report. Available at: http://omop.fnih.org/sites/default/files/RTI%20Acute%20Renal%20Failure%20Final%20Report-110509.pdf. Accessed 08/11/2011.

3. Kachroo S, Jones N, Reynolds MW. Systematic literature review for evaluation of renal failure. Final report prepared for the Foundation of the National Institutes of Health via the Observational Medical Outcomes Partnership. Available at: http://omop.fnih.org/sites/default/files/UBC%20Systematic%20Lit%20Review%20Renal%20Failure%20Final%20Report%20OMOP%209-11-2009.pdf. Accessed 08/11/2011.

4. Moss M, Mannino DM. Race and gender differences in acute respiratory distress syndrome deaths in the United States: an analysis of multiple-cause mortality data (1979-1996). *Crit Care Med*. 2002; 30(8): 1679-85.

5. Rubenfeld GD, Caldwell ES, Steinberg KP, Hudson LD. ICD-9 codes do not accurately identify patients with the acute respiratory distress syndrome (ARDS) (abstract). *Am J Resp Crit Care Med*. 1998; 157: A680.

6. Thompsen G, Morris AH, Danino D, Ellsworth J, Wallace CJ. Accuracy of ICD-9 coding in the diagnosis of ARDS (abstract). *Am Rev Respir Dis*. 1992; 145(4 Part 2): A81.

7. Thomsen GE, Morris AH. Incidence of adult respiratory distress syndrome in the State of Utah. *Am J Respir Crit Care Med*. 1995; 152: 965-71.

8. West SL, D'Aloisio AA, Ringel-Kulka T, Waller AE, Bordley WC. Population-based drug-related anaphylaxis in children and adolescents captured by South Carolina emergency room hospital discharge database (SCERHDD) (2000-2002). *Pharmacoepi Drug Saf*. 2007; 1255-67.