MINI-SENTINEL METHODS

15 COHORTS OF INTEREST FOR SURVEILLANCE PREPAREDNESS

Prepared by: Charles E. Leonard, PharmD, MSCE,[1,2] Cristin P. Freeman, MPH, MBE,[1,2] Hanieh Razzaghi, MPH,[1,2] Ryan M. Carnahan, PharmD, MS, BCPP,[3] Elizabeth A. Chrischilles, MS, PhD,[3] Susan E. Andrade, ScD,[4] Allison Naleway, PhD,[5] Robert Penfold, PhD,[6] Gregory E. Simon, MD, MPH,[6,7] Faith Barash, MD, MPH,[8] Monika Houstoun, PharmD, BCPS,[9] David Menschik, MD, MPH,[10] Patrick Archdeacon, MD[11]

Author Affiliations: 1. Center for Clinical Epidemiology and Biostatistics, Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 2. Center for Pharmacoepidemiology Research and Training, Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA; 3. Department of Epidemiology, College of Public Health, The University of Iowa, Iowa City, IA; 4. Meyers Primary Care Institute, a joint endeavor of Fallon Community Health Plan, Reliant Medical Group, and University of Massachusetts Medical School, Worcester, MA; 5. The Center for Health Research Northwest, Kaiser Permanente, Portland, OR; 6. Group Health Research Institute, Seattle, WA; 7. Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA; 8. Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, FDA, Rockville, MD; 9. Office of Surveillance and Epidemiology, Center for Drug Evaluation and Research, FDA, Silver Spring, MD; 10. Division of Epidemiology, Center for Biologics Evaluation and Research, FDA, Rockville, MD; 11. Office of Medical Policy, Center for Drug Evaluation and Research, FDA, Silver Springs, MD

Acknowledgements: The authors wish to thank Candace Fuller, PhD, MPH (Harvard Pilgrim Health Care Institute), Sean Hennessy, PharmD, PhD (Perelman School of Medicine at the University of Pennsylvania), and Sunali Goonesekera, MS (Harvard Pilgrim Health Care Institute) for contributing meaningful edits to this report; and Susan Forrow, BA (Harvard Pilgrim Health Care Institute) and Carly Comins, BS (Harvard Pilgrim Health Care Institute) for their excellent project management.

February 27, 2014

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 Methods

15 Cohorts Of Interest For Surveillance Preparedness

Table of Contents

1. BACKGROUND

Certain cohorts of persons may be more vulnerable to (i.e., at higher-than-baseline risk for) experiencing adverse effects from medical products and/or medical countermeasures. Yet, these groups are often underrepresented in clinical trials. Therefore, the safety profiles of medical products and medical countermeasures are typically less well-characterized in these individuals. Thus, it is important to develop algorithms to identify cohorts of vulnerable groups within electronic healthcare data environments, such as the Mini-Sentinel distributed database. Active surveillance could then be conducted specifically within these cohorts, providing detailed information regarding risk from their exposure to medical products and medical countermeasures.

2. PURPOSE

The workgroup set forth to recommend algorithms for the identification of vulnerable cohorts of interest to FDA. Contrary to the title of the workgroup opportunity, these included the following 18 (not 15) major groups: nursing home residents, pregnant women, live births, premature infants, persons at high risk for influenza complications, immunocompromised persons, persons with type-1 diabetes, persons with type-2 diabetes, obese persons, persons with coronary artery disease, persons with mood disorders, persons with end stage renal disease (ESRD), persons with hypertension, smokers, persons with asthma, persons with dementia, persons receiving fluoroquinolones for post-exposure prophylaxis, and first responders. The workgroup's main deliverables include the recommendations outlined in this descriptive report as well as an Excel workbook containing reviews of published and unpublished algorithms.

3. WORKGROUP CONSTITUENTS

The findings of this workgroup resulted from a collaboration between the Center for Pharmacoepidemiology Research and Training (CPeRT) at the Perelman School of Medicine of the University of Pennsylvania (as lead site), the University of Iowa, the University of Massachusetts/Meyers Primary Care Institute, the Kaiser Permanente Center for Health Research, the Group Health Research Institute, the Harvard Pilgrim Health Care Institute Mini-Sentinel Operations Center (MSOC), and FDA. Table 1 below lists workgroup members and examples of their specific expertise brought to bear.

Table 1. Workgroup constituents, roles, and expertise

Participating Site	Participant (alphabetically, by site)	Workgroup Role	Examples of Cohort-Defining Expertise
University of Pennsylvania	Cristin Freeman, MPH, MBE	Member	hypertension, immunocompromised, nursing home
	Charles Leonard, PharmD, MSCE	Lead	mood disorders, chronic kidney disease, obesity
	Hanieh Razzaghi, MPH	Member	diabetes, pregnancy-related outcomes
University of Iowa	Ryan Carnahan, PharmD,	Member	smoking, dementia, mood disorders

Participating Site	Participant (alphabetically, by site)	Workgroup Role	Examples of Cohort-Defining Expertise
	MS, BCPP		
	Elizabeth Chrischilles, MS, PhD	Member	
University of Massachusetts / Meyers Primary Care Institute	Susan Andrade, ScD	Member	pregnancy-related outcomes
Kaiser Permanente Northwest	Allison Naleway, PhD	Member	pregnancy-related outcomes, first responders
Group Health Research Institute	Robert Penfold, PhD	Member	mood disorders
	Gregory Simon, MD, MPH	Member	
Harvard Pilgrim Health Care Institute	Elizabeth Cavagnaro, BA	MSOC support	
	Carly Comins, BS		
	Susan Forrow, BA		
	Sunali Goonesekera, MS		
	Candace Fuller, PhD, MPH		
	Lisa Trebino, MS		
	Tiffany Siu Woodworth, MPH		
FDA	Patrick Archdeacon, MD	oversight and guidance	
	Faith Barash, MD, MPH		
	Carlos Bell, MPH		
	Monika Houstoun, PharmD, BCPS		
	David Menschik, MD, MPH		

4. WORKGROUP METHODS

As an overview, major workgroup activities included the: a) survey and inventory of Mini-Sentinel Investigators' and Data Partners' experience in developing algorithms to identify the aforementioned cohorts; b) review of existing (e.g., literature-reported) cohort definitions and selection of cohorts of interest; and c) development of recommendations for cohort-defining algorithms, including the documentation of rationale for their selection and technical specifications.

The survey of Mini-Sentinel Investigators and Data Partners was intended to supplement our planned literature reviews for identifying existing algorithms for cohorts. The workgroup developed a personalized solicitation that described the workgroup purpose and requested provision of published and unpublished algorithms from recipients. This solicitation was emailed by the MSOC. Responses were received and inventoried by the MSOC and the Workgroup Leader.

Workgroup members then conducted a brief review of existing literature on the 18 cohorts named in the workgroup opportunity. Information provided by solicitation respondents was also reviewed. The

intent of this initial limited review of the literature was to identify cohorts for which little or no information was available, so these cohorts could be quickly identified as poor candidates for algorithm development.

For the remaining cohorts with sufficient information available, workgroup members conducted a more comprehensive, secondary review of the literature to identify existing algorithms and characterize their validity metrics. Principal search strategies included the use of PubMed and Google Scholar, and the review of the reference sections of manuscripts identified via these methods. In rare instances in which the lead site's biomedical library did not allow for full-text access to a manuscript, an interlibrary loan procedure was utilized. Further, when detail on an algorithm was not fully presented in an identified manuscript, its lead author was contacted for supplemental information.

Results of the in-depth literature review and synthesis were cataloged in an Excel-formatted spreadsheet with one tab for each cohort of interest. Each tab includes one row-entry per relevant study and identifies the: study authors +/- PubMed identification number (or web hyperlink, in some cases); study population; calendar period under study; cohort-identifying algorithm(s); positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity of each algorithm; number of subjects (N) under study; comments; overall synthesis; and workgroup recommendation. Discussions of these findings were then presented to FDA during weekly (initially, then twice monthly) workgroup teleconferences. In some cases, refinement of the workgroup's initial recommendation was required after receiving input from FDA; refined workgroup recommendations were then presented on subsequent teleconferences.

5. WORKGROUP FINDINGS

5.1. SURVEY/INVENTORY OF MINI-SENTINEL INVESTIGATORS AND DATA PARTNERS

Mini-Sentinel Investigators and Data Partners were contacted via email during August 2012. The MSOC and Workgroup Leader received responses from eight Mini-Sentinel collaborators. Most responses included the provision of published manuscripts describing the validity of specific algorithms and/or compiled code sets (i.e., operational definitions for given cohorts) without descriptions of their validity metrics. Files accompanying these responses were cataloged by MSOC staff supporting this activity. Please note that these standalone files have limited interpretability. They were used by the workgroup's literature reviewers to ensure that all relevant data were examined.

5.2. REVIEW OF DEFINITIONS AND SELECTION OF COHORTS

As a screening step, a brief literature review was conducted for each of the 18 aforementioned cohorts. During this phase, three cohorts were identified as poor candidates for algorithm development: first responders, persons with ESRD, and persons receiving fluoroquinolones for post-exposure prophylaxis.

5.2.1. First Responders

The workgroup received clarification from FDA about the intended use of the first responders cohort. FDA indicated its preference for an algorithm that would broadly capture any demographic of first responders available within the distributed database (i.e., not solely a cohort of healthcare workers as

indicated in the workgroup opportunity) in order to conduct active surveillance on medical products used in pre- and post-exposure interventions (e.g., smallpox vaccines, fluoroquinolones for anthrax, etc.). FDA noted that algorithms that reliably captured important subgroups of first responders could also have value, if algorithms intended to capture this cohort more broadly were unlikely to have good performance metrics.

While potential algorithms for identifying first responders could include tuberculosis testing, vaccinations patterns, serologic testing and employee health visits, the workgroup could find no evidence of these approaches being implemented or validated. An alternate approach could include the use of group subscriber information (e.g., "City of Portland" members within Kaiser Permanente Northwest data), but the MSOC confirmed that subscribership data is not captured by the common data model. Regardless, even if such information were available, it would be difficult to determine the job classification of members covered by the "City of Portland" (e.g., firefighters vs. police officers). It would also be difficult to identify a specific person in the family/household with a particular job classification. Given this, the workgroup informed FDA that, at this time, there are no known automatable algorithms for identifying first responders or important subgroups of first responders within the distributed database.

5.2.2. Persons with end-stage renal disease (ESRD)

During the screening phase, the workgroup found that few studies had evaluated algorithms for identifying persons with ESRD. Much more attention has been paid to the validity of algorithms for identifying persons with chronic kidney disease (CKD) in general (see §5.3.6) or undergoing dialysis procedures in particular. Given this, in addition to the following complicating factors, the workgroup determined that the identification of a cohort of ESRD patients should be abandoned in favor of a cohort of CKD patients. First, Data Partners contributing to the distributed database likely begin to lose follow-up of ESRD (and kidney transplanted) patients from their datasets as Medicare becomes the primary payment source for these individuals (see Centers for Medicare and Medicaid Services [CMS] Product No. 10128 for more information). This was the workgroup's primary reason of concern leading to the recommendation for abandonment of the ESRD cohort. Second, no studies have examined the validity of International Classification of Diseases (ICD-9) 585.6 (*ESRD*) since its addition to the coding manual in 2005. Third, while dialysis procedure codes have a high PPV and sensitivity for identifying persons that have had a dialysis session, such sessions may not exclusively be had by persons with ESRD. For example, persons with an acute renal failure episode could undergo dialysis, including intermittent hemodialysis, continuous renal replacement therapy, and hybrid therapies such as sustained low-efficiency dialysis.

While a selected population of persons with ESRD *could* be identified within the distributed database, they would likely be non-representative of a typical ESRD population—that is, a younger population with greater access to personal resources that may preclude their reliance on Medicare and/or ESRD patients in the early months of their diagnosis while awaiting transition to Medicare. Should FDA maintain an interest in the study of an ESRD population, the workgroup alternatively recommends use of the United States Renal Data System (USRDS) dataset. This National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)-funded national data system collects, analyzes, and distributes information about ESRD in the US. Further, as an alternative to an ESRD cohort of interest, the workgroup recommends

that FDA consider the vulnerable population of persons with CKD (see §5.3.6), acknowledging that the distributed database may still under-identify late-stage patients soon after Medicare coverage begins.

5.2.3. Persons receiving fluoroquinolones for post-exposure prophylaxis (PEP)

The workgroup received clarification from FDA that the intended use of this cohort included applications in medical countermeasure surveillance. Given this, the workgroup focused their review on the identification of persons receiving fluoroquinolone prophylaxis subsequent to suspected inhalation anthrax exposure (i.e., PEP). Lessons learned from a review of this prime example may be applicable to the identification of other cohorts receiving prophylactic therapies in response to occupational exposures or bioterrorism.

Routine capture of fluoroquinolone prescriptions dispensed for PEP would not be expected within the distributed database. Should the Centers for Disease Control and Prevention (CDC) Office of Health Preparedness and Response activate a public health response subsequent to a threat, medications would likely be dispatched from the Strategic National Stockpile (SNS)—initially via 12-hour Push Packages and then via Vendor Managed Inventory. Federal, state, and local community planners would then work to ensure that SNS medicines are provided to affected persons. It is not clear that PEP prescribing would be recorded in an electronic medical record (EMR) and even less likely that billing claims would capture PEP dispensings, as SNS drugs are free. Further federal Emergency Use Authorization would likely obviate the typical "clinician prescribing-pharmacist dispensing" paradigm.

If a subset of fluoroquinolone PEP prescribing/dispensings were expected to appear in the distributed database, dependent on state and/or local community readiness procedures in response to a threat, one could conceivably focus on the identification of regimens of ciprofloxacin 500mg BID x 60 days in conjunction with anthrax vaccine adsorbed (AVA) in three subcutaneous doses at days 0, 14, and 28 as an example. This would also assume, though, that the vaccine would be recorded in the distributed database—unlikely since AVA is only available from state and local health departments via the CDC.

For these reasons, it is unlikely that the current distributed database would allow for the identification of a cohort of persons receiving fluoroquinolones for PEP. Should FDA remain interested in studying such a population, partnership with local municipalities, states, and the CDC would be critical.

5.3. DEVELOPMENT OF ALGORITHMS AND DOCUMENTATION OF SPECIFICATIONS

A more complete literature review was conducted for the following cohorts of interest. For each cohort, one or more recommendations are presented along with a rationale. A high-level summary of estimated PPVs and sensitivities for primary-recommended algorithms are presented in Table 2 below. Of note, not all cohorts undergoing complete review had sufficient evidence to support the workgroup's recommendation of an algorithm or algorithms; these are indicated accordingly in the table.

Further, as the intent of the workgroup was to recommend algorithms for identifying *cohorts* of interest in alignment with the workgroup opportunity, recommendations below were preferentially-selected (when appropriate) based on the expectation that an algorithm would have both a moderate-to-high PPV and moderate-to-high sensitivity. The workgroup does acknowledge, though, that its work may also inform the selection of algorithms for use as *health outcomes of interest (HOIs)* and/or *confounder adjustment*. Further consideration should be given before simply adopting these recommendations for

such purposes. For example, HOI algorithms may wish to further maximize PPV at the cost of sensitivity, while confounder algorithms may wish to further maximize sensitivity at the cost of PPV. On occasion, workgroup findings relevant to these considerations are commented on below, in the hopes of informing future work.

Table 2. Estimated positive predictive values (PPVs) and sensitivities of primary-recommended algorithms for identifying specific cohorts

Cohort of interest	Estimated PPV (%)	Estimated sensitivity (%)
1. Asthma		
Asthma, in adults	72	74
Asthma, in children	94	90
2. Coronary artery disease (CAD)		
CAD, general	95	55-80
Acute myocardial infarction*	50-83	80-95
Angina	no recommendation†	
3. Dementia	81	69
4. Mood disorders		
Depression	49	95
Bipolar	no recommendation†	
5. Smokers, tobacco users	76-94	19-65
6. Chronic kidney disease	71	82
7. Obese persons**	unknown	unknown
8. Hypertension	95	76
9. Immunocompromised		
Exposure to immunologics	unknown	unknown
Human immunodeficiency virus	unknown	96
Cancers of interest	no recommendation†	
10. Influenza complications	unknown	unknown
11. Nursing home residents	unknown	unknown
12. Diabetes mellitus		
Type I	96	61
Type II	no recommendation†	
Type I or Type II	>97	unknown
13. Pregnant women	no recommendation†	
14. Live births	no recommendation†	
15. Babies born prematurely	no recommendation†	

* developed with intent to maximize sensitivity while maintaining an acceptable PPV
** primary recommendation to rely on subset of persons with BMI reported or calculable within the distributed database, not via diagnosis- or medication-based code sets
† Although a complete literature review was conducted for this cohort, no recommendation could be made by the workgroup.

5.3.1. Persons with asthma

Overall, algorithms for identifying persons with asthma have high PPVs and sensitivities. Given this, the proposed algorithms below would function well in identifying a cohort of persons with asthma in which to examine a medical product-HOI relationship and as a code set for confounder adjustment. For developing a cohort of adults with asthma, the workgroup recommends an algorithm developed by Gershon *et al* (PubMed Identifier [PMID]: 20011725) in which an adult is considered asthmatic if having at least three ambulatory care visits for asthma or at least one hospitalization for asthma during a two

year window (PPV = 72%, sensitivity = 74%, when compared to expert panel diagnosis). If one wishes to maximize sensitivity, as may be desired for a confounder control algorithm, the workgroup recommends using an algorithm with at least one ambulatory care visit for asthma or at least one hospitalization for asthma ever in the past (PPV = 51%, sensitivity = 95%, when compared to expert panel diagnosis).

While algorithms by Blais *et al* (PMID: 1637489) could also be considered, their use would be fraught with challenges. For example, the common data model does not provide information on provider specialty to allow for the identification of visits to respiratory physicians. Also, Regie de l'assurance maladie du Quebec (RAMQ) Medical Services data (the dataset in which their algorithms were examined) only captures one diagnosis per medical visit, limiting generalizability to the distributed database.

For developing a cohort of children with asthma, the workgroup recommends an algorithm developed by Wakefield *et al* (PMID: 16871628) in which a child is considered a probable asthmatic if meeting the Council of State and Territorial Epidemiologists (CSTE) asthma surveillance definition *or* having at least one asthma medication dispensing event (PPV = 94%, sensitivity = 90%). This algorithm is referred to by the manuscript authors as modified CSTE model 2. Probable asthma was defined by the CSTE as having at least one inpatient, outpatient, or emergency department claim listing ICD-9 493.XX (*asthma*) as the primary/first-listed diagnosis during a 12-month period. An asthma medication was defined by a bronchodilator (including anticholinergic, sympathomimetic, or xanthine derivative), leukotriene formation inhibitor, leukotriene receptor antagonist, inhaled corticosteroid, or oral corticosteroid (confirmed via correspondence with the author on May 30, 2013 and June 10, 2013). Given that oral corticosteroids may be prescribed for other pediatric conditions, and that children with asthma treated with oral corticosteroids are likely to be concomitantly treated with a bronchodilator, the workgroup recommends that oral corticosteroids be dropped from the definition of an asthma medication. Of further note, while Wakefield *et al*'s modifications of Healthcare Effectiveness Data and Information Set (HEDIS) definitions for asthma also performed well, they did not perform as well as modifications to CSTE, hence the workgroup's recommendation based on the latter.

5.3.2. Persons with coronary artery disease (CAD)

Based on a review of reporting measures for CAD—including those published by the Physician Consortium for Performance Improvement (i.e., Physician Quality Reporting Measure for CAD), American Medical Association (i.e., CAD Algorithm for Measures Calculation), Bridges to Excellence (i.e., CAD Care Recognition) and American College of Cardiology Foundation/American Heart Association (i.e., Chronic Stable CAD Performance Measure Set)—the workgroup generally considered CAD to include: acute myocardial infarction (AMI); other acute and sub-acute forms of ischemic heart disease; old MI; angina pectoris; coronary atherosclerosis; chronic total occlusion of coronary artery; coronary atherosclerosis due to lipid rich plaque; other specified forms of chronic ischemic heart disease; and unspecified chronic ischemic heart disease. Therefore, the workgroup focused on developing algorithms for: CAD, in general, focused on the definition above; and AMI and unstable angina, in particular, as two separate sub-cohorts of potential interest.

For developing a cohort of CAD patients, the workgroup recommends the following algorithm by Solberg *et al* (PMID: 16849780) that was examined within HealthPartners administrative data: two any-position outpatient diagnoses or one any-position inpatient diagnosis for ICD-9 410.XX, 411.X, 412, 413.X, or

414.X in a 12-month period (PPV = 95%). While Solberg *et al* did not examine sensitivity, other research using a similar algorithm suggests an approximate sensitivity of 55-80%. While an algorithm by Birman-Deych *et al* (PMID: 15838413) also had an excellent PPV and good sensitivity, it was solely examined in inpatient claims within a cohort of atrial fibrillation patients, thereby making interpretation challenging. As a secondary recommendation, if an improvement to sensitivity is desired (yet at a considerable cost of PPV), the workgroup recommends another algorithm developed by Solberg *et al*: one any-position outpatient diagnosis or one any-position inpatient diagnosis for ICD-9 410.XX, 411.X, 412, 413.X, or 414.X in a 12-month period (PPV = 60%). This secondary recommendation may be more useful for a confounder control code set.

For developing a cohort of AMI patients, the workgroup first reviewed recent Mini-Sentinel work by Cutrona *et al* (PMID: 22745038). This algorithm, evaluated in the distributed database, had a PPV = 86%; yet, the algorithm's sensitivity was not examined. While Cutrona *et al*'s work should be relied upon for defining AMI as a HOI, its utility for defining a cohort of persons with AMI is unknown. Therefore, the workgroup focused on the elucidation of an algorithm with a known and high sensitivity that did not completely compromise PPV. The workgroup principally recommends a *de novo* algorithm that requires an any-position hospital discharge diagnosis of 410.XX, 427.4X, or 427.5. Based on prior work by Heckbert *et al* (PMID: 15583367) and Newton *et al* (PMID: 10210237), as reported by Metcalfe *et al* (PMID: 22742621), such an algorithm might be expected to have a PPV of about 50-83% and sensitivity of about 80-95%.

The workgroup found scant data on validated algorithms for identifying persons with unstable angina. Merry *et al* (PMID: 19337843) examined an any-position hospital discharge diagnosis for unstable angina in a Dutch dataset, finding a good PPV and moderate sensitivity, but this study relied on data that was approximately 20 years old and intermingled results from ICD-9 and ICD-10 coding systems making interpretation difficult. Therefore, the workgroup has not recommended an algorithm for this sub-cohort.

5.3.3. Persons with dementia

A 2012 systematic review by St. Germaine-Smith *et al* (PMID: 22914826) reported on eight studies that used ICD-9 codes to identify persons with Alzheimer's Disease (AD) and dementia; further, the workgroup identified three additional studies not addressed by the systematic review (i.e., Katon *et al*, Kho *et al*, Gruber-Baldini *et al*). Among the former, Bharmal *et al* (PMID: 17545733) examined ICD-9 diagnoses for dementia within Indiana Medicaid claims of persons ≥ 40 years of age and reported a PPV = 81% and sensitivity = 69%. The workgroup recommends this algorithm (requiring at least one any-position ICD-9 diagnosis for dementia on any claim type) for identifying a cohort of persons with dementia given its inclusion of codes for senile dementia, presenile dementia, senile dementia with delusional or depressive features, senile dementia with delirium, vascular dementia, alcohol-induced persistent dementia, dementia in conditions classified elsewhere (Alzheimer's dementia), Creutzfeldt-Jakob disease, Alzheimer's disease, frontotemporal dementia, senile degeneration of brain, and dementia with Lewy bodies. A modification to the above algorithm requiring at least two diagnoses does little to improve PPV and is therefore not recommended. Should one wish to maximize PPV, at the considerable cost of sensitivity, the workgroup secondarily recommends an algorithm by Quan *et al* (PMID: 18756617) that is limited to hospital discharge diagnoses alone (PPV = 96%). It is important to note that the sensitivity of this algorithm is very low (sensitivity = 32%).

5.3.4. Persons with mood disorders

Mood disorders generally encompass both depressive and bipolar disorders. The workgroup received clarification from FDA that it desired algorithms to capture persons with depression and bipolar disorder separately, if possible. Although these conditions can be difficult to disentangle, we present a discussion of each in turn.

The identification of a cohort of persons with depression is complicated by the following factors: a) incomplete recognition of depression in outpatient settings; b) heterogeneous studies examining depression algorithms; c) inconsistencies in the gold standard used by studies evaluating algorithm performance (i.e., physician diagnosis vs. depression screening tool vs. structured diagnostic assessment); d) lack of systematic studies comparing primary to secondary depression diagnoses in inpatient claims; e) purposeful miscoding of depression by some clinicians for certain patients; and f) inflated (and potentially misleading) PPVs for algorithms evaluated in samples with atypically high depression prevalences (i.e., selected cohorts). Yet, numerous publications have evaluated algorithms for identifying persons diagnosed with depression, including a Mini-Sentinel HOI Evidence Review by Townsend *et al*.

Townsend *et al*, in a subsequent publication (PMID: 22262603) to that referenced above, identified 11 studies evaluating depression algorithms and recommended the following based on work by Spettell *et al* (PMID: 12968818) within managed care data—at least two occurrences among the following criteria during a 12-month period: 1) primary/first-listed ICD-9 diagnosis for depression (296.2X [*major depressive disorder, single episode*], 296.3X [*major depressive disorder, recurrent episode*], 300.4 [*dysthymic disorder*], or 311 [*depressive disorder, not elsewhere classified*]); and 2) antidepressant prescription claim (including monoamine oxidase inhibitors, tricyclics, tetracyclics, selective serotonin reuptake inhibitors, serotonin-2 receptor antagonists, alpha-2 receptor antagonists, and miscellaneous antidepressants such as modified cyclics; excluding lithium and excluding trimipramine in persons <19 years old). The algorithm could be satisfied by either one occurrence of each component of the algorithm definition or two occurrences within either individual component. The PPV and sensitivity of this algorithm, as reported by Spettell *et al*, was 49% and 95%, respectively; kappa was 0.47, as calculated by Townsend *et al*. The workgroup recommends this algorithm for identifying a cohort of persons with diagnosed depression, yet suggests that the prescription component be updated to: include serotonin and norepinephrine reuptake inhibitors and other newer medications that are likely specific for use in depression; and exclude amitriptyline, doxepin, and bupropion as the former two are often used for sleep and the latter for smoking cessation (per expert input from Drs. Penfold and Simon). Of note, a modification of the Spettell *et al* algorithm that requires at least one occurrence within the diagnosis component slightly improved PPV (61%), but severely reduced sensitivity (52%) and is therefore not recommended by the workgroup.

A secondary definition for identifying persons with diagnosed depression can be adapted from work completed by Trinh *et al* (PMID: 21514880). This study was published shortly after Townsend *et al*'s manuscript. The best-functioning of the many algorithms evaluated by Trinh *et al* included a diagnosis of depression or an antidepressant medication without an anxiety, posttraumatic stress disorder, or pain diagnosis (PPV = 76%, sensitivity = 85%). Given the gain in PPV yet modest loss of sensitivity when compared to the Spettell *et al* algorithm discussed above, one might suggest that this algorithm be recommended as the primary rather than secondary definition. Yet, the inclusion of some hospital-

specific billing codes (in addition to ICD-9 diagnosis codes) complicates the interpretation of the algorithm's validation metrics. Therefore, the workgroup recommends that the Trinh *et al* algorithm only be used in parallel with the principal recommendation, to examine an alternate definition of persons with diagnosed depression—in particular if there is concern over the modest PPV of the principal recommendation.

Operationally, a cohort of persons with diagnosed depression is defined by Trinh *et al* as ICD-9 290.13 (*presenile dementia with depressive features*), 290.21 (*senile dementia with depressive features*), 290.43 (*vascular dementia with depressed mood*), 296.2X (*major depressive disorder, single episode*), 296.3X (*major depressive disorder, recurrent episode*), 296.82 (*atypical depressive disorder*), 296.9 (*other an unspecified episodic mood disorder*), 296.99 (*other specified episodic mood disorder*), 298.0 (*depressive type psychosis*), 300.4 (*dysthymic disorder*), 301.10 (*affective personality disorder, unspecified*), 305.8 (*nondependent antidepressant type abuse*), 305.81 (*nondependent antidepressant type abuse, continuous use*), 309.0 (*adjustment reaction with adjustment disorder with depressed mood*), 309.1 (*adjustment reaction with prolonged depressive reaction*), 311 (*depressive disorder, not elsewhere classified*), 969.0X (*poisoning by antidepressants*), E939.0 (*antidepressants causing adverse effects in therapeutic use*) or V79.0 (*screening for depression*); or one of the following antidepressant medications in a person without prior ICD-9 diagnosis of 300.00 (*anxiety state, unspecified*), 300.01 (*panic disorder, without agoraphobia*), 300.02 (*generalized anxiety disorder*), 300.09 (*other anxiety states*), 309.81 (*posttraumatic stress disorder*), or 338.X (*pain, not elsewhere classified*): amitriptyline, bupropion, citalopram, clomipramine, desipramine, doxepin, duloxetine, escitalopram, fluoxetine, fluvoxamine, imipramine, maprotiline, mirtazapine, nefazodone, nortriptyline, paroxetine, phenelzine, protriptyline, sertraline, selegiline patch, tranylcypromine, trimipramine, or venlafaxine. As with the Spettell *et al* algorithm, the workgroup further recommends excluding amitriptyline, bupropion, and doxepin from the prescription component of this definition.

Algorithms for bipolar disorder have been much less frequently evaluated than those for depression. A study by Sellgren *et al* (PMID: 21838734) examined an ICD-based algorithm for identifying bipolar patients within Swedish data and reported a PPV = 81%. Yet, this result was based on a composite algorithm that included ICD-8, ICD-9, and ICD-10 codes and greatly limits its interpretability. Another study by Unutzer *et al* (PMID: 10715498) examined diagnosis- and prescription-based algorithms within a health maintenance organization. The following three algorithms had the lowest false positive rates (i.e., <10%): 1) inpatient diagnosis of bipolar disorder; 2) outpatient diagnosis of bipolar disorder from a mental health provider; and 3) diagnosis of bipolar disorder from a non-mental health provider + a prescription for a mood stabilizer (i.e., lithium, carbamazepine, valproate) that was not associated with a seizure disorder or written by a neurologist. Given that sensitivities of these algorithms are unknown and that the common data model does not have the ability to determine prescriber specialty, the workgroup cannot recommend a particular algorithm for the identification for persons with bipolar disorder.

5.3.5. Smokers

Given complexities in identifying specific modes of tobacco exposure (e.g., smoking cigarettes vs. smoking cigars vs. pipe use vs. smokeless tobacco), the workgroup focused on the identification of tobacco users more broadly. In general, algorithms have excellent PPVs but poor sensitivities. If the desire is to optimize PPV, for example in identifying a cohort of known smokers in which to examine a

medical product-HOI relationship, the workgroup recommends the algorithm by Chen *et al* (PMID: 22904436) as examined within Kaiser's EMR. This algorithm includes the presence of any the following ICD-9, Current Procedural Terminology (CPT) level 1, and CPT level 2 codes on any claim type: 292.0 (*drug withdrawal*), 305.1 (*nondependent tobacco use disorder*), V15.82 (*personal history of tobacco use*), V65.42 (*counseling on substance use and abuse*), 99406 (*smoking and tobacco use cessation counseling visit, intermediate*), 99407 (*smoking and tobacco use cessation counseling visit, intensive*), 1000F (*tobacco use assessed*), 1001F (*tobacco use, non-smoking*), 4000F (*tobacco use cessation intervention, counseling*), 4001F (*tobacco use cessation intervention, pharmacologic therapy*). Yet, it should be noted that the small fraction of tobacco users identified via this method (i.e., carrying diagnostic codes indicative of such) could well differ from tobacco users in general.

If the desire is to maximize sensitivity, as may be prudent if evaluating tobacco use as a confounder, the workgroup recommends the following *de novo* algorithm proposed by Drs. Leonard and Carnahan—a composite of a number of other algorithms presented in the literature. This algorithm takes advantage of the identification of tobacco users via a host of ICD-9, CPT level 1, CPT level 2, Healthcare Common Procedure Coding System (HCPCS), and National Drug Coding System (NDC) codes. While one would anticipate that such an algorithm would improve upon the poor sensitivities of individual algorithms reported by others, neither sensitivity nor PPV of this *de novo* algorithm has been examined. While further improvements to sensitivity might be gained by the addition of American Dental Association (ADA) codes (i.e., D9920, D1320) to the *de novo* algorithm, a modular program (MSY3_STR122) found that these codes were very rarely used within the distributed database.

It is important to note that these algorithms may not differentiate between current and former smokers; it is critical to keep this caveat in mind when using the above algorithms. Such information is likely best gleaned from discreet EMR fields that are subject to routine updating by medical assistants, nurses, and physicians. Future attention should be paid to Data Partners contributing EMR data and the potential richness of these sources for the direct capture of tobacco use/smoking status via these discreet EMR fields, if adapted by the common data model.

5.3.6. Persons with chronic kidney disease (CKD)

As discussed previously in §5.2.2, the workgroup recommended that FDA consider the broader population of persons with CKD, instead of focusing solely on the ESRD population. While the publications discussed below were generally developed by their respective authors to identify persons with all-stage CKD (including ESRD and unspecified-stage disease), a few caveats regarding the use of these algorithms within the distributed database are worth noting. First, early stage patients may be under-identified (Ferris *et al*, PMID: 19214023). This may be driven by asymptomatic early-stage disease and/or lack of physician awareness of CKD. Second, late-stage patients (including those with ESRD/receiving chronic dialysis) may also be under-identified. This is likely related to the Data Partners contributing to the distributed database, as a vast proportion of ESRD patients will transition to Medicare—a population currently not well-captured by the distributed database. Therefore, although not formally evaluated as part of this workgroup opportunity, the workgroup hypothesizes that CKD patients identified within the distributed database using the methods below may disproportionately select those with Stage 3 through Stage 5 disease.

The validity of CKD coding is well-summarized in the following two review papers: Vlasschaert *et al* (PMID: 21184918) and Grams *et al* (on behalf of the CDC CKD Surveillance Team, PMID: 20692079). The review by Vlasschaert *et al* identified 25 studies of 13 databases in four countries, 14 of which examined ICD-9 codes for CKD. Of these, five studies (Kern *et al*, Parker *et al*, Quan'08 *et al*, Romano *et al*, So *et al*) had adequate algorithm validation statistics (PPV = 68%-82%, sensitivity = 42%-88%) and warranted further review by the workgroup. In addition, modular programs were run to determine the frequency of codes proposed by these algorithms. Among these five, the algorithms proposed by Quan '08 *et al* and Romano *et al* were examined in a general population, and therefore their results might be most generalizable to routine surveillance. Vlasschaert *et al* also reported that algorithm sensitivity has been increasing in more modern times; this is a welcomed trend given the prospective nature of the Sentinel system.

Grams *et al* reviewed 30 studies (18 based in US datasets), a number of which were previously reviewed by Vlasschaert *et al*. Of the remaining studies, four were examined in further detail by the workgroup to inform the development of a CKD algorithm. An algorithm by Ferris *et al* (PMID: 19214023) had an inadequate PPV. An algorithm by Levy *et al* (PMID: 10579743) was based on a very small sample size and reported crude agreement instead of PPV. An algorithm by Quan '02 *et al* (PMID: 12187181) performed less well than the Quan '08 *et al* algorithm referenced above. An algorithm by Stevens *et al* (PMID: 15930090) had an inadequate sensitivity. Given this, the workgroup focused on recommending an algorithm initially identified in the Vlasschaert review.

The algorithm proposed by Kern *et al* (PMID: 16584465) was evaluated in a diabetic population and included ICD-9 codes seemingly specific for acute kidney disease (in addition to those indicative of chronic disease) and was therefore eliminated from consideration. The algorithms proposed by Quan '08 *et al* (PMID: 18756617) and So et al (PMID: 17173686) can be compared head-to-head, as they included different codes evaluated in the same database (i.e., Calgary Health Region). Quan's algorithm is preferable, as it has a PPV = 71% (vs. 68%) while the sensitivities are similar. Further, the So algorithm was examined in a very small number of persons. Similarly, the algorithms proposed by Parker *et al* (PMID: 16799364) and Romano *et al* (PMID: 8277803) can be compared head-to-head, as they included different codes evaluated in the same database (i.e., CA Hospital Discharge Abstract). While the Romano algorithm had a higher PPV, it is dated (1988); the Parker algorithm is likely more reflective of current CKD coding practices. Therefore, since the Quan algorithm is more sensitive than the Parker algorithm (82% vs. 43%), the workgroup principally recommends the use of the algorithm proposed by Quan '08 *et al*. The operational definition of the Quan '08 algorithm (as reported by Vlasschaert *et al*) is an any-position, any-claim ICD-9 code of 582.XX (*chronic glomerulonephritis*), 583 (*nephritis and nephropathy, not specified as acute or chronic*), 583.0 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of proliferative glomerulonephritis*), 583.1 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of membranous glomerulonephritis*), 583.2 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of membranoproliferative glomerulonephritis*), 583.4 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of rapidly progressive glomerulonephritis*), 583.6 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of renal cortical necrosis*), 583.7 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of renal medullary necrosis*), 585.X (*chronic kidney disease*), 586.X (*renal failure, unspecified*) or 588.X (*disorders resulting from impaired renal function*).

A secondary algorithm recommendation, as originally suggested by Dr. Carnahan, combines components of the Quan and Parker algorithms. This algorithm is operationally defined as an any-position, any-claim ICD-9 code of 582.XX (*chronic glomerulonephritis*), 583 (*nephritis and nephropathy, not specified as acute or chronic*), 583.0 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of proliferative glomerulonephritis*), 583.1 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of membranous glomerulonephritis*), 583.2 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of membranoproliferative glomerulonephritis*), 583.4 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of rapidly progressive glomerulonephritis*), 583.6 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of renal cortical necrosis*), 583.7 (*nephritis and nephropathy, not specified as acute or chronic, with lesion of renal medullary necrosis*), 585.X (*chronic kidney disease*), 586.X (*renal failure, unspecified*) 588.X (*disorders resulting from impaired renal function*), 792.5 (*cloudy dialysis effluent*), V42.0 (*kidney transplant status*), V45.1 (*renal dialysis status*), V56.0 (*renal dialysis encounter*), V56.2 (*fitting and adjustment of peritoneal dialysis catheter*), V56.31 (*encounter for adequacy testing for hemodialysis*), V56.32 (*encounter for adequacy testing for peritoneal dialysis*), or V56.8 (*encounter for other dialysis*). While one could hypothesize that such an algorithm would further improve the sensitivity of the workgroup's primary recommendation, this alternate algorithm has not been evaluated in its current form.

5.3.7. Obese persons

The workgroup identified numerous studies examining ICD-9 code- and medication-based algorithms for identifying obese persons in general or among specific disease populations (N = 11), pregnant women (N = 3), and children (N = 2). These typically compared ICD-9 278.XX (*overweight, obesity and other hyperalimentation*) to body mass index (BMI) as the reference standard. Based on a review of algorithms evaluated by these studies, the workgroup does not recommend that ICD-9- and/or medication-based algorithms be used to identify a cohort of obese persons, given very poor sensitivities and variably-performing PPVs. Rather, since a decently-sized proportion of the distributed database population has at least one height recorded and at least one weight recorded (7.2 and 7.9 million persons, respectively, from nine Data Partners beginning in 2006), and that BMI defines obesity, the workgroup recommends that any study wishing to identify a cohort of obese persons rely on pre-constructed SAS macros from the Data Core for calculating BMI from height and weight.

It is important to keep in mind that calculating BMI for the subset of persons with vital signs records in the distributed database still comes with challenges. These include, but are not limited to: a) determining an allowable lag time between height and weight measurements and how this lag time may vary by child vs. adult; b) determining ranges of heights and weights that will be considered plausible, as some Data Partners do not "clean" their data; c) specifying an algorithm for BMI calculation when a person has multiple values from which to choose; d) identification and possible exclusion of women during periods of pregnancy; and e) for children, using an absolute BMI vs. z-score.

Finally, if one wishes to use a combination of ICD-9-diagnosed obesity or BMI-calculated obesity in controlling for obesity as a covariate in a study, acknowledging that this will underestimate obese persons among those without reported vitals, consider an algorithm by Quan *et al* (PMID: 18756617) that includes an any-position ICD-9 diagnosis of 278.0X, expanded to include any claim type.

5.3.8. Persons with hypertension

Numerous publications propose validated algorithms with moderate-to-high sensitivities and specificities for identifying persons with hypertension, yet most are rather dated. A notable exception is a 2006 study by Bullano *et al* (PMID: 16641668) that examined the validation statistics of different hypertension algorithms within claims of a mid-Atlantic health plan. For developing a cohort of persons with hypertension, the workgroup principally recommends Bullano *et al*'s "Rule B" algorithm that requires the presence of ≥1 medical claim for hypertension + ≥1 prescription for an antihypertensive medication (PPV = 95%, sensitivity = 76%, specificity = 93%, kappa = 0.65). This algorithm has better validation statistics than those proposed by Humphries *et al* (PMID: 10785564), Quan *et al* (PMID: 19858407), and Tessier-Sherman *et al* (PMID: 23331960).

Operationally, Bullano *et al*'s "Rule B" algorithm is defined by the presence of ICD-9 code 401.XX (*essential hypertension*) on any claim in any position, plus a medication within one of the following broadly-defined classes, within 90 days of the ICD-9 diagnosis: angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, beta blocker, calcium channel blocker, and centrally- or peripherally-acting antiadrenergic agent. Consideration could be given to expanding this class listing to include direct renin inhibitors, selective aldosterone receptor antagonists, and certain vasodilators (e.g., hydralazine), although these were not a component of the validated definition. Requiring additional occurrences of hypertensive claim diagnoses and/or antihypertensive prescriptions do not appreciably improve specificity and have a marked negative effect on sensitivity; further reliance on diagnoses alone in the absence of a prescription claim performs more poorly. Therefore, the workgroup does not recommend these latter modifications to the "Rule B" algorithm.

The workgroup considered other algorithms for identifying persons with hypertension, but favored none over that reported by Bullano *et al*. For example, Rector *et al* (PMID: 15533190) proposed the use of one hypertension diagnosis (401.XX [*essential hypertension*], 402.XX [*hypertensive heart disease*], 403.XX [*hypertensive chronic kidney disease*], or 404.XX [*hypertensive heart and chronic kidney disease*]) in any position on any claim type or an antihypertensive pharmacy claim, finding a higher sensitivity (0.90) yet lower specificity (0.60) than Bullano's "Rule B" algorithm. Further, the study by Rector *et al* relied on the suboptimal method of patient self-report as the gold standard for determining hypertension, rather than medical record review. Additional data from Robinson *et al* (PMID: 9298082), as determined from provincial Manitoban claims diagnoses of 401.XX-404.XX vs. survey responses, found that 401.XX alone contributed to 99% of all hypertensive cases found in medical and hospital data.

5.3.9. Immunocompromised persons

The workgroup received clarification from FDA that it wanted algorithms capable of capturing both broadly and separately a variety of immunocompromised groups, including persons exposed to certain "strong" immunologic therapies, persons with human immunodeficiency virus (HIV), and persons with selected cancers of interest. Given the heterogeneity of immunocompromised persons, the workgroup recommended against developing broad algorithms intended to capture all-cause immunocompromisation, although acknowledged that it would be possible to use the following recommended algorithms jointly if appropriate in a given context.

5.3.9.1. Persons exposed to immunologic therapies of interest

As persons with major cancers (and thereby treated with immunosuppressive chemotherapy and/or radiation therapies) would be identified via the approach discussed in §5.3.9.3 below, the intent of this subcohort is to identify non-cancer patients receiving immunologic treatment thought to suppress the immune system. Subsequent to an iterative discussion with FDA, the workgroup developed a listing of the following "strongly immunosuppressive" medical products, principally based on their use in transplantation/organ rejection settings or other conditions such as multiple sclerosis, rheumatoid-like arthropathies and lupus. These agents, grouped according to Drug Facts & Comparisons-designated classes, include selected immunosuppressives (alefacept, azathioprine, basiliximab, belatacept, non-ophthalmic cyclosporine, glatiramer, mycophenolate, sirolimus, and non-topical tacrolimus), selected immunomodulators (abatacept, adalimumab, anakinra, canakinumab, certolizumab, dimethyl fumarate, etanercept, fingolimod, golimumab, infliximab, lenalidomide, mitoxantrone, natalizumab, pomalidomide, rilonacept, teriflunomide, thalidomide, tocilizumab, and ustekinumab), a selected antirheumatic kinase inhibitor (tofacitinib), selected monoclonal antibodies (alemtuzumab, ofatumumab, and rituximab), and an alkylating agent (cyclophosphamide). Methotrexate, an antimetabolite with less potent immunosuppressive properties often used in less serious disease, was specifically excluded because it was thought that its inclusion would frequently result in the identification of subjects who were only minimally immunosuppressed. Operationally, prescriptions for the aforementioned agents, as identified by combinations of NDCs and HCPCSs (e.g., J codes), would be searched for within persons devoid of a malignant neoplasm or neuroendocrine tumor diagnosis (defined as ICD-9 140*-209*). Given the frequency with which new NDCs are added, a listing of agent-specific codes is not curated here as they would nearly-immediately be outdated; such a listing should be developed *de novo* at the time of subcohort development.

While this NDC/HCPCS approach has face validity, it is important to note that the validation metrics of such an algorithm is unknown.

5.3.9.2. Persons with HIV

While the workgroup acknowledges that linking to clinical cohort/registry data is the gold standard for defining an HIV population, we are able to recommend the following algorithms for use within the distributed database. Principally, the workgroup recommends the use of Antoniou *et al*'s (PMID: 21738786) preferred algorithm of "three codes in three years", among the 48 algorithms they evaluated within administrative health data of Ontario in the mid- to late-2000s—specifically selecting the algorithm with the highest specificity while maximizing sensitivity over the shortest time interval (sensitivity = 96.2%, specificity = 99.6%, kappa = 0.97). Reliance on hospital admissions or antiretroviral drug therapies did not augment the validity of administrative data-based case definitions. Operationally, this preferred algorithm should identify three occurrences of ICD-9 042 (*human immunodeficiency virus disease*) in any position on any claim type over a three year period.

Given that a chronic disease such as HIV usually requires multiple contacts with the health system to diagnose and treat, a single-visit diagnosis code is often insufficient to accurately identify cases. Therefore, the workgroup's alternate recommendation is a modification of Fultz *et al*'s (PMID: 16849965) "algorithm #2," requiring >1 outpatient or ≥1 inpatient ICD-9 code for 042 (*human immunodeficiency virus disease*) or V08 (*asymptomatic human immunodeficiency virus infection status,*

HIV positive NOS). While Fultz's published algorithm (PPV = 88%, sensitivity = 90%, specificity = 99.9%) included a Diagnosis Related Group (DRG) component, such information may not be reliably available from all Data Partners for all visit types; should it be determined that DRG capture is sufficient, Fultz's published algorithm can be used in its entirety. Further, this alternate workgroup recommendation is similar to that proposed by Nosyk *et al* (PMID: 23382898).

The workgroup suggests using the primary recommended algorithm when one anticipates sufficient follow-up time to be able to fulfill the three year requirement (at the very least). Otherwise, the workgroup suggests using the alternate recommended algorithm, acknowledging that the algorithm would likely perform best if DRG codes were reliably available from all Data Partners within the distributed database.

5.3.9.3. Persons with cancers of interest

Cancers are not typically studied as a homogenous group, given differences in the histological type and primary site of the lesion—each that often has its distinct risk factors, screening requirements, pathology, clinical manifestations, diagnostic testing, differential diagnoses, staging, treatment and prognosis, as examples. Therefore, studies examining algorithms for identifying persons with any-type cancer are scant. The workgroup recommends that FDA primarily focus on the aforementioned subcohorts of immunocompromised persons (§5.3.9.1 and §5.3.9.2) in the absence of cancer registry data. Yet, if FDA maintains an interest in developing a subcohort of persons with cancer, consideration should be given to developing subcohorts of persons with specific cancers rather than cancer in general. While the identification and recommendation of validated algorithms for specific types of cancers is outside of the scope of this workgroup, it is worth commenting on types of cancers may be of specific interest to those developing this subcohort of immunocompromised persons. The following brief discussion is not intended to be exhaustive with regard to all cancers thought to confer a high degree of immunosuppression in an affected patient, but rather to provide insight on the thought processes of the workgroup and FDA during the completion of this task.

First, consideration should be given to the identification of persons with hematopoietic cancers such as leukemias, lymphomas, myelomas. A comprehensive listing of related conditions can be found at the following National Cancer Institute Surveillance Epidemiology & End Results (SEER) weblink: http://seer.cancer.gov/seertools/hemelymph/ (a product of the SEER Hematopoietic Project). A cursory review of studies examining broad algorithms for identifying leukemias and lymphomas can be found in the workgroup's Excel workbook deliverable. Second, consideration should be given to the identification of persons with solid/organ-based tumors such as carcinomas and sarcomas found in sites such as the liver, lung, kidneys, pancreas, digestive tract, breast, reproductive organs, brain, and connective and other soft tissue, as examples. Third, consideration should be given to the identification of persons with chronic/indolent cancers such as some prostate, ovarian, thymic, parathyroid, lung, gastrointestinal stromal, and renal malignancies, as examples. Finally, consideration should be given to the identification of persons requiring extended (e.g., ≥1 year) courses of chemotherapy and radiation.

5.3.10. Persons at high risk for influenza complications

Currently, the CDC considers the following specific groups at high risk of developing influenza-related complications (http://www.cdc.gov/flu/about/disease/high_risk.htm, last updated: 11/07/2013): children <5 years; children <19 years receiving long-term aspirin therapy; adults ≥ 65 years; pregnant

women; American Indians; Alaskan Natives; persons with chronic pulmonary (e.g., asthma, chronic obstructive pulmonary disease), cardiovascular, renal, hepatic, hematological (e.g., sickle cell), metabolic (e.g., diabetes, mitochondrial disorders), morbid obesity, neurologic (e.g., epilepsy, stroke), or neurodevelopmental disease (e.g., mental retardation); and persons with a weakened immune system due to disease or medication. It is important to keep in mind that these component groups defined as "high risk" are subject to modification on an annual (or more frequent) basis and should be reevaluated at the time during which one wishes to construct this cohort. Given this, the workgroup offers advice for building such a cohort, rather than operational instructions.

The common data model allows for the easy identification of persons by age, thereby allowing for the construction of subcohorts of children <5 years and adults ≥ 65 years. Children between 5 and 19 years of age are easily identified as well, and their prescription data can be used to determine long-term aspirin therapy to the extent that such over-the-counter use is clinician-prescribed and captured by the Data Partner. Of note, the CDC does not provide an operational definition for "long term" use. Subcohorts of pregnant women will be identified subsequent to the work referred to in §5.3.13. Currently, it is not possible to routinely identify subcohorts of American Indians and Alaskan Natives, as the common data model poorly captures race/ethnicity (with the exception of a few Data Partners). The identification of immunocompromised persons is discussed in §5.3.9, yet the CDC definition of this group is likely broader than that discussed above; for example, the workgroup's definition does not include persons with all types of cancer nor persons chronically-treated with corticosteroids. The remaining medical condition-specific groups can be identified by disease-specific, ICD-based algorithms. Yet, the broadness of these CDC-defined groups likely lends itself to using wider swaths of the ICD-9 code book rather than identifying algorithms for each individual disease. For example, it may be prudent to use ICD-9 codes 490*-496* to identify a cohort of persons with *chronic obstructive pulmonary disease and allied conditions*, rather than develop specific algorithms for identifying persons with chronic bronchitis, asthma, and emphysema individually, as examples. Alternatively, some major high risk populations (such as persons with asthma, diabetes, or cardiac disease) can be identified by operational definitions proposed in influenza immunization quality measures developed by CMS, HEDIS, and or the Consumer Assessment of Healthcare Providers and Systems program.

Researchers, such as Ernst (PMID: 11186551, see manuscript's Table 1), Nakamura *et al* (PMID: 18977969, see manuscript's appendix), and Hak *et al* (PMID: 12145718, see manuscript's materials and methods section) have developed ICD-9 code sets intended to capture conditions placing persons at high risk for influenza complications in adults <50 years, adolescents and elders, respectively. While these code sets were in alignment with the Advisory Committee on Immunization Practices (ACIP) recommendations at the time of their development, they are likely incomplete based on the current ACIP definition of a person at high risk for influenza complications. That being said, these algorithms (in addition to work recently completed by the Vaccine Safety Datalink [VSD], PMID: 23129321) would be an excellent starting place for developing an operational definition. Of note, though, validation metrics of these algorithms are unknown. Regardless, the workgroup suggests modifying these algorithms, in concert with the age-based approaches described above, for defining this cohort of interest.

The workgroup identified a study by Ahmed *et al* (PMID: 16028341) that reported the validation metrics for an ICD-based algorithm to identify adults at high risk for influenza (PPV = 87%, sensitivity = 12%, and specificity = 99%). However, the ICD-9 codes used by the authors appeared less than comprehensive.

Therefore, the workgroup suggests that precedence be given to the work by the VSD, Ernst, Nakamura *et al*, and Hak *et al*.

5.3.11. Nursing home residents

The workgroup discussed a number of options for identifying a cohort of nursing home residents. In general, while the workgroup is reasonably confident that persons who have *ever* been in a nursing home can be identified using CPT codes, this approach cannot reliably be used to build specific time windows defining periods of nursing home stays. This is a major drawback and led to the workgroup's hesitation in recommending an algorithm for identifying this cohort, as it is difficult to envision the utility in simply knowing if an individual was *ever* in a nursing home (without knowing when and for how long).

If FDA maintains an interest in persons *ever* having spent time in a nursing home, specific CPT codes used to bill for the evaluation and management of patients receiving nursing facility care can be utilized. However, these codes were updated in 2005 and algorithms using such have not been validated. While validation metrics are known for algorithms using the older CPT codes for evaluation and management of patients receiving nursing facility care, and these older algorithms can be easily modified by supplanting the newer CPT codes, inferring validity metrics between old (PMIDs: 18070360, 18953230; Iwashyna TJ. *Health Services & Outcomes Research Methodology* 2003) and new algorithms should be done with caution. As a further complication, these validated algorithms were developed within Medicare data that has detailed place of service information, and such algorithms may not be transportable for use in the distributed database. With these major caveats in mind, one could consider the following operational definition for identifying a cohort of persons *ever* having been in a nursing home: one CPT code indicative of evaluation and management in a nursing home (99304, 99305, 99306, 99307, 99308, 99309, 99310, 99315, 99316, 99318). To reemphasize, use of these codes to build time windows of nursing homes stays has not been evaluated and cannot be recommended at this time without a formal validation within the distributed database. In fact, results from a preliminary modular program run suggested that CPT codes indicative of nursing home discharge evaluations were used infrequently and disproportionately less than initial assessment CPT codes.

The workgroup further discussed the capability of the common data model for identifying nursing home stays in an alternate manner—with particular interest in the care setting variable in the encounter table (coupled with admitting source and discharge status information). Yet, the institutional stay setting has limited granularity, capturing hospice, skilled nursing facility, rehabilitation center, nursing home, residential, overnight non-hospital dialysis and other non-hospital stays. While use of these common data model elements could inform the identification of periods of nursing home stays, such a method would need to be independently validated.

5.3.12. Persons with diabetes

Validation metrics of algorithms for identifying persons with types I and II diabetes can vary widely depending on the setting in which patients are identified and if laboratory measures are utilized. With this in mind, the workgroup proposes primary and secondary algorithms below.

5.3.12.1. Persons with type I diabetes

Bobo *et al* (PMID: 22920280) validated an algorithm for identifying type I diabetics within Tennessee Medicaid data, finding a PPV = 80%. Operationally, the algorithm required: 1) a primary discharge diagnosis of 250 (*diabetes mellitus*), 250.0X (*diabetes mellitus without mention of complication*), 250.1X (*diabetes with ketoacidosis*), 250.2X (*diabetes with hyperosmolarity*), 250.3X (*diabetes with other coma*), or 250.9X (*diabetes with unspecified complication*); 2) an inpatient stay with a secondary discharge diagnosis for one of these same ICD-9 codes + no diagnosis for polycystic ovary syndrome (i.e., 265.4, *polycystic ovaries*) within 120 days of the diabetes diagnosis + a confirmatory antidiabetic prescription or an additional any-setting any-position ICD-9 code for diabetes within 120 days; or 3) an outpatient visit with a primary diagnosis for one of these same ICD-9 codes + a confirmatory antidiabetic prescription or an any-position inpatient ICD-9 code for diabetes within 120 days. In any of these three scenarios, ≥1 prescription for insulin was also required within 120 days of the diabetes diagnosis, with no more than a single prescription for an oral antidiabetic drug during that interval. A single prescription for an oral agent did not serve as an exclusion criterion, because such drugs may be occasionally prescribed while awaiting the results of confirmatory testing for type I diabetes. If the aforementioned definition was not met, the individual was classified as a type II diabetic. Of note though, the study population consisted solely of a small number of pediatric, adolescent, and young adult atypical antipsychotic users aged 6-24 years.

Rhodes *et al* (PMID: 17192348) also examined a pediatric, adolescent, and young adult population, yet within the Endocrine/Diabetes or Obesity Programs at Children's Hospital in Boston, finding a PPV = 97% (sensitivity was not reported). Operationally, the algorithm required an inpatient or outpatient ICD-9 code for 250.X1 or 250.X3.

Klompas *et al* (PMID: 23193215) developed algorithms for identifying type I diabetics, without regard to patient age, within Atrius Health EMR data. Their algorithm requiring ≥2 ICD-9 diagnoses for 250.X1 or 250.X3 + a current prescription for insulin + no prescription for an oral antidiabetic agent at any time (excluding metformin) yielded a PPV = 81% and sensitivity = 32%. Twenty-one other candidate algorithms were presented by the authors in the manuscript's Table 3. Of note, among persons meeting screening criteria for potential diabetes, algorithms that maximized sensitivity (often at the cost of PPV) included individual components such as: a prescription for insulin; no record of any oral antidiabetic drug; and no record of any oral antidiabetic drug (excluding metformin). Algorithms that maximized PPV (often at the cost of sensitivity) included individual components such as: a ratio of type I to type II ICD-9 codes >0.5; a ratio of type I to type II ICD-9 codes >0.5 + prescription for insulin; a ratio of type I to type II ICD-9 codes >0.5 + prescription for glucagon; C-peptide <0.8; and a prescription for urine acetone test strips. An algorithm that maximized PPV while maintaining an acceptable level of sensitivity included a requirement for a ratio of type I to type II ICD-9 codes >0.5 + no prescription for an oral antidiabetic drug (excluding metformin)—an algorithm with a PPV = 96% and sensitivity = 61%. Klompas *et al* also developed an "optimized" algorithm, achieving a PPV = 96% and sensitivity = 100%, yet this definition included laboratory components not currently supported by the common data model (e.g., C-peptide, diabetes autoantibodies results).

For identifying a cohort of persons with type I diabetes, the workgroup recommends the following algorithm proposed by Klompas *et al*: a ratio of type I to type II ICD-9 codes >0.5 + no prescription for an oral antidiabetic drug (excluding metformin). Should requisite laboratory data be made available within

the distributed database, the Klompas *et al* "optimized" algorithm should be used. Should FDA be interested in identifying a cohort of pediatric, adolescent, and young adults with type 1 diabetes, the workgroup recommends the algorithm by Rhodes *et al*, as it was evaluated in more persons than that proposed by Bobo *et al*.

5.3.12.2. Persons with type I or type II diabetes

Most studies that validate diabetes algorithms do not distinguish between types I and II. Therefore, the workgroup presents recommendations for combined algorithms below.

O'Connor *et al* (PMID: 10178496) validated an ICD-9-based diabetes algorithm within 3,186 adult primary care patients of a staff model HMO in the Upper Midwest, using self report telephone survey as the gold standard. Within a 2-year period, diabetes was operationally defined by the presence of ≥2 ICD-9 diagnoses for 250.X (*diabetes mellitus*). This algorithm yielded a PPV = 94%, NPV = 99%, sensitivity = 91%, and specificity = 99%. While the high PPV and sensitivity makes this algorithm attractive, this validation study only evaluated outpatient codes and was rather dated. Related work by Hebert *et al* (PMID: 10624032) found similar but slightly less favorable results when examining 250.00-250.93 (*diabetes mellitus*), 357.2 (*polyneuropathy in diabetes*), 362.0-362.02 (*diabetic retinopathy*), or 366.41 (*diabetic cataract*) in any position on any claim type within Medicare claims, using self report via the Medicare Current Beneficiary Survey as the gold standard. Further, Hebert *et al* confirmed the importance of requiring at least two diagnosis codes when identifying persons with diabetes.

Solberg *et al* (PMID: 16849780) set forth to develop an algorithm to maximize PPV based on the prior work by O'Connor *et al*, seemingly within the same staff model HMO. Their Health Plan Employer Data and Information Set (HEDIS)-based algorithm required: 1) ≥2 outpatient or ≥1 inpatient ICD-9 codes from among the following, in a given calendar year: 250.XX, 357.2, 362.01, 362.02, or 366.41; or 2) a prescription for an antidiabetic medication (excluding single-agent metformin) in the same calendar year. The reference standard was medical record review. This algorithm has a PPV of >97%; sensitivity was not reported.

The workgroup recommends the above algorithm developed by Solberg *et al* as the primary method for identifying a cohort of persons with diabetes, as the author's reference standard was based on clinical data rather than self report. As a secondary recommendation, the workgroup suggests a modification to the Solberg algorithm that is more in alignment with the current HEDIS Comprehensive Diabetes Care measure definition. This secondary algorithm is operationally defined as: 1) ≥2 outpatient or ≥1 emergency department/inpatient ICD-9 from among the following, in a given calendar year: 250.XX, 357.2, 362.0X, or 366.41; or 2) a prescription for an antidiabetic medication (excluding single-agent metformin) in the same calendar year, only if no ICD-9 diagnosis of 251.8 (*other specified disorders of pancreatic internal secretion*), 256.4 (*polycystic ovaries*) or 962.0 (*poisoning by adrenal cortical steroids*) occurs in the same calendar year or year prior. Given that the validity metrics of this secondary algorithm have not been evaluated, the workgroup recommends that it only be used in parallel with the principal recommendation, to examine an alternate definition of persons with diabetes—in particular if there is concern that the principal algorithm might miss persons diagnosed with diabetes in the emergency department.

Should laboratory data, such as hemoglobin A1c and blood glucose, be made available within the distributed database, the workgroup recommends the "indicator"-based approach proposed by Zgibor

et al (PMID: 16934906). In particular, identification of a person with diabetes based on the requirement for ≥2 of the following six indicators had a PPV = 96%-97%, sensitivity 95%-97%, and specificity = 53%-90%: 1) prescription for an antidiabetic medication; 2) ICD-9 code for 250.XX on an inpatient claim; 3) ICD-9 code for 250.XX on an outpatient claim; 4) ICD-9 code for 250.XX on an emergency department claim; 5) any A1c measurement, regardless of value; and 6) blood glucose > 200 mg/dl.

5.3.13. Pregnancy cohorts, including pregnant women, live births, & babies born prematurely

The workgroup received clarification from FDA that a more comprehensive review of these populations and development of algorithms would be handled by a Medication Exposures in Pregnancy Risk Evaluation Program (MEPREP) task order being led by Dr. Darren Toh. Please refer to their work on these topics.

6. APPENDIX

Appendix Table 1. Operational definitions for algorithms recommended by the Workgroup

Cohort of interest	Recommended algorithm	Comments
1. Asthma		
Asthma, in adults		
Primary	>=3 ambulatory care visit for asthma (ICD-9 493.XX in any position) or >=1 hospitalization for asthma (ICD-9 493.XX in any position); 2-year claim period	Gershon et al
Secondary	>=1 ambulatory care visit for asthma (ICD-9 493.XX in any position) or >=1 hospitalization for asthma (ICD-9 493.XX in any position); claim period unspecified	Maximizes sensitivity (as reported by Gershon et al), as may be useful for a confounder code set
Asthma, in children		
Primary	[>=1 inpatient, outpatient, or emergency department claim listing ICD-9 493.XX as the primary/first-listed diagnosis during a 12-month period] or [>=1 asthma medication dispensing event, defined by a bronchodilator (including anticholinergic, sympathomimetic, or xanthine derivative), leukotriene formation inhibitor, leukotriene receptor antagonist, or inhaled corticosteroid]	For "probable" asthma; workgroup modification to Wakefield et al's Council of State & Territorial Epidemiologists (CSTE) modification #2
2. Coronary artery disease (CAD)		
CAD, general		
Primary	[two any-position outpatient diagnoses] or [one any-position inpatient diagnosis] for ICD-9 410.XX, 411.X, 412, 413.X, or 414.X in a 12-month period	Solberg et al
Secondary	[one any-position outpatient diagnosis] or [one any-position inpatient diagnosis] for ICD-9 410.XX, 411.X, 412, 413.X, or 414.X in a 12-month period	Maximizes sensitivity, as may be useful for a confounder code set
Acute myocardial infarction		
Primary	Any-position hospital discharge diagnosis of 410.XX, 427.4X, or 427.5	*De novo*; maximizes sensitivity while maintaining an adequate PPV
Angina	no recommendation	
3. Dementia		
Primary	Any-file ICD-9 diagnosis of 290.0, 290.1X, 290.2X, 290.3, 290.4X, 291.2, 294.1X, 046.1, 331.0, 331.1X, 331.2, or 331.82 on 1 claim, using a look-back period of 30 months; patient age >=40	Bharmal et al
Secondary	Any-position hospital discharge diagnosis of an ICD-9 code for 290.X, 294.1X, or 331.2	If an improvement on PPV is required, yet at great cost to sensitivity
4. Mood disorders		
Depression		
Primary	>=2 occurrences among the following criteria during a 12-month period: 1) primary/first-listed ICD-9 diagnosis for depression (296.2X, 296.3X, 300.4, or 311; and 2) antidepressant prescription claim: including monoamine oxidase inhibitors, tricyclics (excluding doxepin and amitriptyline), tetracyclics, selective serotonin reuptake inhibitors, serotonin and norepinephrine reuptake inhibitors, serotonin-2 receptor antagonists, alpha-2 receptor antagonists, and miscellaneous antidepressants such as modified cyclics; excluding	Workgroup modification to Spettell et al

Cohort of interest	Recommended algorithm	Comments
	lithium, excluding trimipramine in persons <19 years old, and excluding bupropion. The algorithm could be satisfied by either one occurrence of each component of the algorithm definition or two occurrences within either individual component.	
Secondary	[billing diagnosis of depression defined by ICD-9: 290.13, 290.21, 290.43, 296.2X, 296.3X, 296.82, 296.9, 296.99, 298, 300.4, 301.1, 305.8, 305.81, 309, 309.1, 311, 969, E939.0, or V79.0] or [antidepressant prescription for: citalopram, clomipramine, desipramine, duloxetine, escitalopram, fluoxetine, fluvoxamine, imipramine, maprotiline, mirtazapine, nefazodone, nortriptyline, paroxetine, phenelzine, protriptyline, sertraline, selegiline patch, tranylcypromine, trimipramine, or venlafaxine AND no diagnosis of anxiety/panic (ICD-9: 300.00, 300.01, 300.02, 300.09), PTSD (ICD-9: 309.81) or pain (ICD-9: 338)]	Workgroup modification to Trinh et al; use in parallel with primary recommendation if particular concern over modest PPV of above algorithm
Bipolar	no recommendation	
5. Smokers, tobacco users		
Primary	Presence of any the following codes on any claim type: ICD-9: 292.0, 305.1, V15.82, V65.42; CPT-I: 99406, 99407; CPT-II: 1000F, 1001F, 4000F, 4001F	Chen et al; sensitivity anticipated to be poor
Secondary	Presence of any the following codes on any claim type: ICD-9: 305.1, 649.0X, 989.84, V15.82 ; CPT-I: 83887, 99406, 99407; CPT-II: 1034F, 1035F, 4000F, 4001F, 4004F; HCPCS: C9801, C9802, G0375, G0376, G0436, G0437, G8093, G8094, G8402, G8403, G8453, G8454, G8455, G8456, G8688, G9016, S4990, S4991, S4995, S9075, S9453; NDC: nicotine replacement, varenicline, Zyban (brand only)	Leonard-Carnahan *de novo* algorithm developed by workgroup; sensitivity not known, but expected to be greater than the primary algorithm
6. Chronic kidney disease		
Primary	Any-position, any-claim ICD-9 code of 582.XX, 583, 583.0, 583.1, 583.2, 583.4, 583.6, 583.7, 585.X, 586.X, or 588.X	Quan '08 et al; likely to identify persons with Stage 3-5 disease
Secondary	Any-position, any-claim ICD-9 code of 582.XX, 583, 583.0, 583.1, 583.2, 583.4, 583.6, 583.7, 585.X, 586.X, 588.X, 792.5, V42.0, V45.1, V56.0, V56.2, V56.31, V56.32, or V56.8	Workgroup-derived combination of Quan '08 et al and Parker et al algorithms; intended to improve on sensitivity of above algorithm; likely to identify persons with Stage 3-5 disease
7. Obese persons		
Primary	Rely on pre-constructed SAS macros for calculating body mass index from height and weight	Refer to work by Mini-Sentinel Data Core
Secondary	Any-position, any-claim ICD-9 code for 278.0X	Workgroup modification to Quan et al; may be appropriate for other uses, but not cohort development
8. Hypertension		
Primary	ICD-9 code 401.XX on any claim in any position, plus a medication within one of the following broadly-defined classes, within 90 days of the ICD-9 diagnosis: angiotensin-converting enzyme inhibitor, angiotensin receptor blocker, beta blocker (systemic), calcium channel blocker, centrally- or peripherally-acting antiadrenergic agent, direct renin inhibitor, selective aldosterone receptor	Workgroup modification to Bullano et al's "Rule B" algorithm

Cohort of interest	Recommended algorithm	Comments
	antagonist, or vasodilator (i.e., hydralazine, oral minoxidil)	
9. Immunocompromised		
Exposure to immunologics		
Primary	Prescription for one of the following agents, as identified by NDC and/or HCPCSs, within persons devoid of a malignant neoplasm or neuroendocrine tumor diagnosis (defined as ICD-9 140.X - 209.X): selected immunosuppressives (alefacept, azathioprine, basiliximab, belatacept, non-ophthalmic cyclosporine, glatiramer, mycophenolate, sirolimus, and non-topical tacrolimus), selected immunomodulators (abatacept, adalimumab, anakinra, canakinumab, certolizumab, dimethyl fumarate, etanercept, fingolimod, golimumab, infliximab, lenalidomide, mitoxantrone, natalizumab, pomalidomide, rilonacept, teriflunomide, thalidomide, tocilizumab, and ustekinumab), a selected antirheumatic kinase inhibitor (tofacitinib), selected monoclonal antibodies (alemtuzumab, ofatumumab, and rituximab), and an alkylating agent (cyclophosphamide)	Intent to identify non-cancer patients receiving immunologic treatment thought to suppress the immune system
Human immunodeficiency virus		
Primary	At least 3 occurrences of ICD-9 042 in any position on any claim type over a 3-year period	Antoniou et al; still likely suboptimal to linkage to clinical cohort/registry data; use if sufficient follow-up time to be able to fulfill the 3-year requirement
Secondary	>1 outpatient or ≥1 inpatient ICD-9 code for 042 or V08 (or DRG 488-490, if available within the distributed database)	Use if follow-up time is limited; algorithm will perform better if DRG component of definition is included
Cancers of interest	no recommendation	
10. Influenza complications		
Primary	Identify Centers for Disease Control and Prevention high-risk groups current as of the time the cohort is to be constructed and build an algorithm based on advice offered within the Workgroup report	Begin with reliance on prior work by Ernst, Nakamura et al, Hak et al, and the Vaccine Safety Datalink
11. Nursing home residents	no recommendation	
12. Diabetes mellitus		
Type I		
Primary	[a ratio of type I (ICD-9 250.X1 or 250.X3) to type II (ICD-9 250.X0 or 250.X2) codes >0.5] and [no prescription for an oral antidiabetic drug (excluding metformin)]	Klompas et al
Secondary	[a ratio of type I (ICD-9 250.X1 or 250.X3) to type II (ICD-9 250.X0 or 250.X2) codes >0.5 and a prescription for glucagon] or [a ratio of type I (ICD-9 250.X1 or 250.X3) to type II (ICD-9 250.X0 or 250.X2) codes >0.5 and no prescription for an oral antidiabetic drug (excluding metformin)] or [C-peptide negative] or [diabetes autoantibodies positive] or [prescription for urine acetone test strips]	Klompas et al's optimized algorithm; use if requisite laboratory data become widely available within the distributed database, if not, drop the definition components requiring C-peptide and diabetes autoantibodies, as PPV and sensitivity are still

Cohort of interest	Recommended algorithm	Comments
		high for such a definition.
Type II	no recommendation	
Type I or Type II		
Primary	[≥2 outpatient or ≥1 inpatient ICD-9 codes from among the following, in a given calendar year: 250.XX, 357.2, 362.01, 362.02, or 366.41] or [a prescription for an antidiabetic medication (excluding single-agent metformin) in the same calendar year]	Solberg et al
Secondary	1) ≥2 outpatient or ≥1 emergency department/inpatient ICD-9 from among the following, in a given calendar year: 250.XX, 357.2, 362.0X, or 366.41; or 2) a prescription for an antidiabetic medication (excluding single-agent metformin) in the same calendar year, only if no ICD-9 diagnosis of 251.8 (*other specified disorders of pancreatic internal secretion*), 256.4 (*polycystic ovaries*) or 962.0 (*poisoning by adrenal cortical steroids*) occurs in the same calendar year or year prior	Workgroup modification to Solberg et al; use in parallel with the primary recommendation to examine alternate definition if there is concern that the primary might miss persons diagnosed with diabetes in the emergency department
Secondary	≥2 of the following six indicators: 1) prescription for an antidiabetic medication; 2) ICD-9 code for 250.XX on an inpatient claim; 3) ICD-9 code for 250.XX on an outpatient claim; 4) ICD-9 code for 250.XX on an emergency department claim; 5) any A1c measurement, regardless of value; and 6) blood glucose > 200 mg/dl	Zgibor et al; use if requisite laboratory data become widely available within the distributed database
13. Pregnant women	no recommendation	
14. Live births	no recommendation	
15. Babies born prematurely	no recommendation	