Drug Saf (2013) 36 (Suppl 1):S181–S193
DOI 10.1007/s40264-013-0111-1

ORIGINAL RESEARCH ARTICLE

Alternative Outcome Definitions and Their Effect on the Performance of Methods for Observational Outcome Studies

Christian G. Reich · Patrick B. Ryan · Martijn J. Schuemie

Abstract

Background A systematic risk identification system has the potential to test marketed drugs for important Health Outcomes of Interest or HOI. For each HOI, multiple definitions are used in the literature, and some of them are validated for certain databases. However, little is known about the effect of different definitions on the ability of methods to estimate their association with medical products.

Objectives Alternative definitions of HOI were studied for their effect on the performance of analytical methods in observational outcome studies.

Methods A set of alternative definitions for three HOI were defined based on literature review and clinical diagnosis guidelines: acute kidney injury, acute liver injury and acute myocardial infarction. The definitions varied by the choice of diagnostic codes and the inclusion of procedure codes and lab values. They were then used to empirically study an array of analytical methods with various analytical choices in four observational healthcare databases. The methods were executed against predefined drug-HOI pairs to generate an effect estimate and standard error for each pair. These test cases included positive controls (active ingredients with evidence to suspect a positive association with the outcome) and negative controls (active ingredients with no evidence to expect an effect on the outcome). Three different performance metrics where used: (i) Area Under the Receiver Operator Characteristics (ROC) curve (AUC) as a measure of a method's ability to distinguish between positive and negative test cases, (ii) Measure of bias by estimation of distribution of observed effect estimates for the negative test pairs where the true effect can be assumed to be one (no relative risk), and (iii) Minimal Detectable Relative Risk (MDRR) as a measure of whether there is sufficient power to generate effect estimates.

Results In the three outcomes studied, different definitions of outcomes show comparable ability to differentiate true from false control cases (AUC) and a similar bias estimation. However, broader definitions generating larger outcome cohorts allowed more drugs to be studied with sufficient statistical power.

Conclusions Broader definitions are preferred since they allow studying drugs with lower prevalence than the more precise or narrow definitions while showing comparable performance characteristics in differentiation of signal vs. no signal as well as effect size estimation.

The OMOP research used data from Truven Health Analytics (formerly the Health Business of Thomson Reuters), and includes MarketScan® Research Databases, represented with MarketScan Lab Supplemental (MSLR, 1.2 m persons), MarketScan Medicare Supplemental Beneficiaries (MDCR, 4.6 m persons), MarketScan Multi-State Medicaid (MDCD, 10.8 m persons), MarketScan Commercial Claims and Encounters (CCAE, 46.5 m persons). Data also provided by Quintiles® Practice Research Database (formerly General Electric's Electronic Health Record, 11.2 m persons) database. GE is an electronic health record database while the other four databases contain administrative claims data.

C. G. Reich (✉)
AstraZeneca PLC, 35 Gatehouse Drive,
Waltham, MA 02451, USA
e-mail: reich@omop.org

P. B. Ryan
Janssen Research and Development LLC, Titusville, NJ, USA

M. J. Schuemie
Department of Medical Informatics, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands

C. G. Reich · P. B. Ryan · M. J. Schuemie
Observational Medical Outcomes Partnership, Foundation for the National Institutes of Health, Bethesda, MD, USA

1 Introduction

Large-scale observational data have tremendous potential and are increasingly adopted in observational epidemiological studies of benefit or risk of medical treatments due to their large sample sizes, high generalizability to medical practice and low cost relative to randomized clinical trials. However, the data used in these studies are collected for a different purpose: reimbursement (administrative claims data) and healthcare delivery (electronic healthcare records). In order to study outcomes with such data, patient cohorts have to be defined for the case and control populations using codes that represent clinical facts, such as diagnostic, procedure, drug prescriptions or dispensing event and lab test codes. This is easier said than done, as the congruence of the codes and clinical reality is influenced by financial pressures to maximize remuneration, the semantic precision of the codes themselves and the quality of medical records.

These outcome definitions are usually contrived by the authors of studies through selection of codes and interpretation of their use in medical practice or in light of clinical guidelines. It is assumed that the accuracy of an observational study estimating the effect between medical treatment and outcome is dependent on the amount of misclassification of their definitions [1], and performance criteria of information retrieval are applied to quantify them: sensitivity and specificity or positive and negative predictive value. Typically, the researcher has to make a tradeoff between narrow and precise clinical definitions resulting in more accurate cases, or broader definitions with a higher sensitivity. However, these parameters are hard to establish for a given cohort, as they require costly and technically challenging validation or source record verification studies, where cases ascertained by the definition are manually checked against the medical records of the patients. As a result, for any one outcome multiple alternative definitions exist in the literature with a drastically varying degree of validation and unknown reproducibility across different databases, as shown by us and others [2, 3]. This creates a substantial challenge for the interpretation of study reports, which could have tremendous impact on the indication of treatments in medical practice.

Though the statistical relationship between misclassification rate and accuracy of study results is well founded [1], the choice of outcome definitions might have an influence on the study result in other ways: narrow definitions may produce a smaller number of cases with an increase in the confidence interval of the study, and the outcome chosen for the study may not coincide exactly with the true biological effect of the medication. An example for the latter effect is a narrow definition of Myocardial Infarction excluding other ischemic events, while in reality the biological mechanism of action of the treatment might cause either effect.

Instead of attempting to quantify these parameters and adding to the complexity of study design and interpretation, we decided to empirically test the performance of alternate outcome definitions directly for their ultimate purpose: the ability to determine the strength of the association between medical treatment and outcome. As part of the Observational Medical Outcomes Partnership (OMOP) experiments [2], we applied a set of test cases of drugs with known association to three Health Outcomes of Interest (HOIs) in multiple alternate definitions, as well as a set of negative control drugs, to a library of methods in four select administrative claim and electronic healthcare record (EHR) databases. We tested the ability to discriminate between signal and negative control as well as the amount of bias introduced through the definitions and estimated the Minimal Detectable Relative Risk (MDRR) the definitions allowed to determine in a given database.

2 Methods

2.1 Experiment Design

The study was conducted as part of the larger OMOP experiment [2], For the purpose of this experiment, three Health Outcomes of Interest (HOIs) were chosen for their variety of alternative definitions: Acute Kidney Injury (AKI), Acute Liver Injury (ALI) and Acute Myocardial Infarction (AMI). We conducted a comprehensive literature review to inform our choices for these definitions based on the experience in epidemiological studies [4]. The result of this review can be found at http://omop.org/HOI. We found that there is very little consensus in the literature about outcome definitions. We therefore decided to create a library of HOI definitions that cover a wide representation of the prior work for testing in the OMOP experiment: broad and narrow selections of diagnosis codes, combination of diagnosis codes with diagnostic or therapeutic procedures and lab values. See Appendix for details of the alternative definitions.

2.2 Acute Kidney Injury

AKI or acute renal failure is a rapid loss of kidney function resulting in the inability to produce sufficient amount of urine, leading to the accumulation of toxic substances in the circulation. It is caused by low blood supply (pre-renal), damage to the kidney tissue itself (intrinsic) or obstruction of the urinary tract (post-renal). In the clinic, AKI is diagnosed on the basis of characteristic laboratory findings, such as elevated blood urea nitrogen and creatinine or the cessation of urine production [5]. These findings are the foundation for outcome definitions used in observational studies [6]. Figure 1

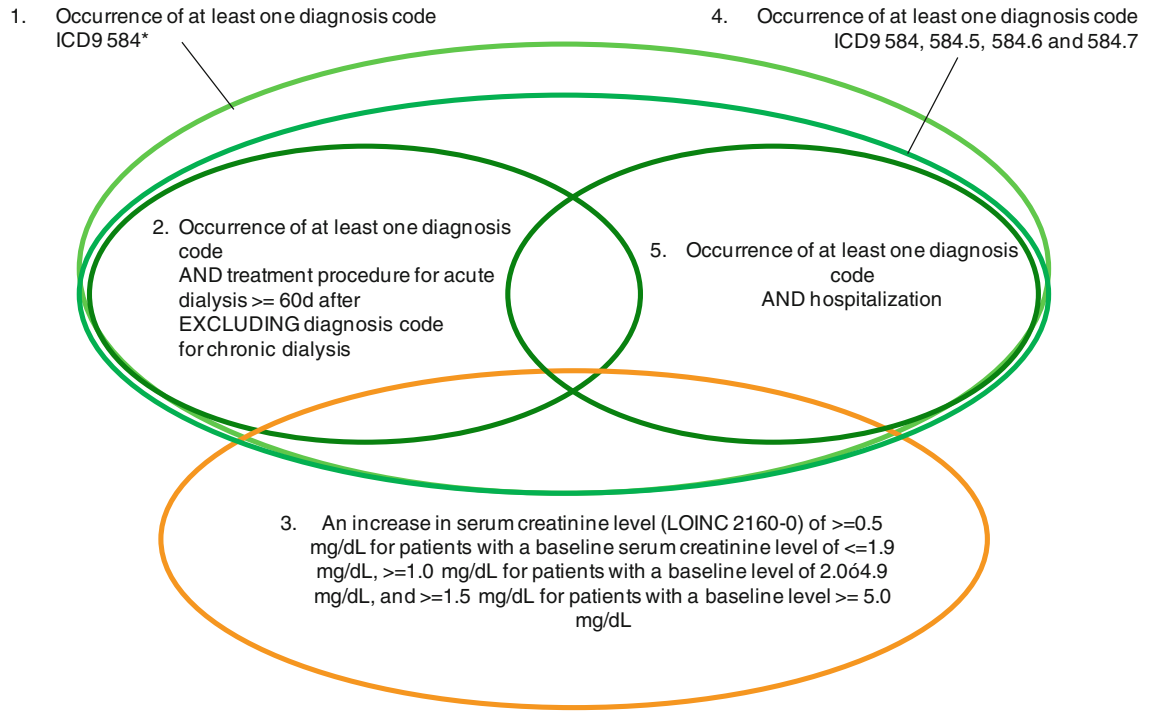

Fig. 1 Alternative definitions of AKI. The Venn diagram indicates the relationship of these definitions to each other. *AKI* acute kidney injury, *ICD-9* International Classification of Diseases, Ninth Revision, *LOINC* Logical Observation Identifiers Names and Codes

shows the alternative definitions and their relationship to each other. The definitions either rely on diagnostic or procedure codes (green ovals) or based on serum creatinine lab findings (orange ovals).

2.3 Acute Liver Injury

ALI or acute liver failure is the loss of liver function of synthesis, metabolization and excretion of a wide range of substances in the absence of a preexisting liver condition, leading to impaired coagulation and encephalopathy [7]. The variety of definitions is larger than in AKI, as the clinical diagnosis is based on physical examination, laboratory tests and current and past medical history, and these parameters are subjective clinical findings and rarely available in observational databases. Figure 2 shows the alternative definitions and their relationships to each other. The definitions are based on diagnostic codes (green ovals), diagnostic codes in combination with procedure codes (orange), diagnostic codes with transaminase elevation in the lab results (red) and purely lab results according to Hy's law (blue).

2.4 Acute Myocardial Infarction

AMI or heart attack occurs when myocardial ischemia, a diminished blood supply to the heart, exceeds a critical threshold and overwhelms myocardial cellular repair mechanisms designed to maintain normal function and homeostasis. Diagnosis is based on the clinical history, EKG and blood test results, especially creatine kinase (CK), CK-MB fraction, and troponin I and T levels.

Figure 3 shows the alternative definitions and their relationships to each other. The definitions are based on diagnostic codes (green ovals), diagnostic codes in combination with procedure codes (orange), and based on EKG and lab values (red).

2.5 Testing of Methods

A variety of analytical methods is used in risk identification of medical products [8]. A method is defined as a generic design of how cohorts are identified and how an association between treatment and outcome is evaluated. As part of the OMOP experimental framework, a method library was established for the purpose of empirical testing for their performance [9]. All specific analysis choices, such as the pre-exposure control period or the length of the post-exposure time-at-risk were parameterized.

The tests were conducted using four observational healthcare databases to allow evaluation of performance across different populations and data capture processes during the years 2003–2008: MarketScan Medicare Supplemental Beneficiaries (MDCR, 4.6 m patients); MarketScan Multi-State Medicaid (MDCD, 10.8 m patients); MarketScan Commercial Claims and Encounters (CCAE,

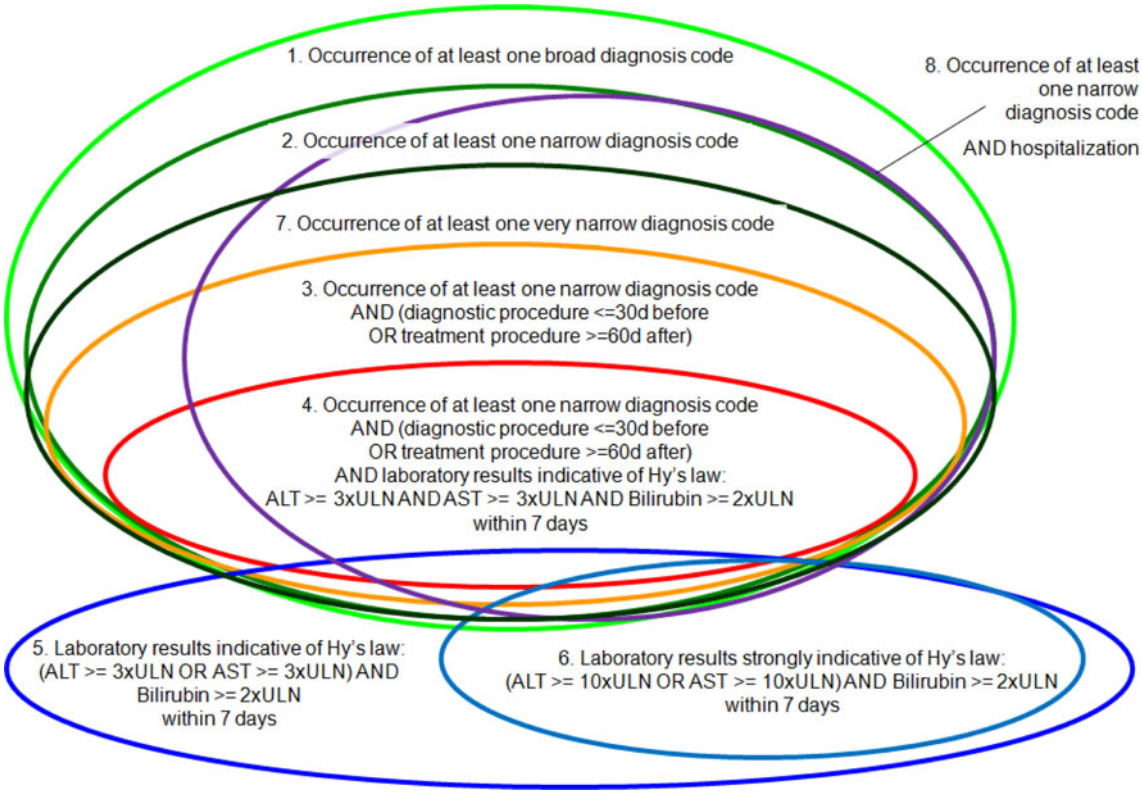

Fig. 2 Alternative definitions of ALI. The Venn diagram indicates the nested relationship of these definitions. *ALI* acute liver injury, *ALT* alanine aminotransferase, *AST* aspartate aminotransferase, *ULN* upper limit of normal

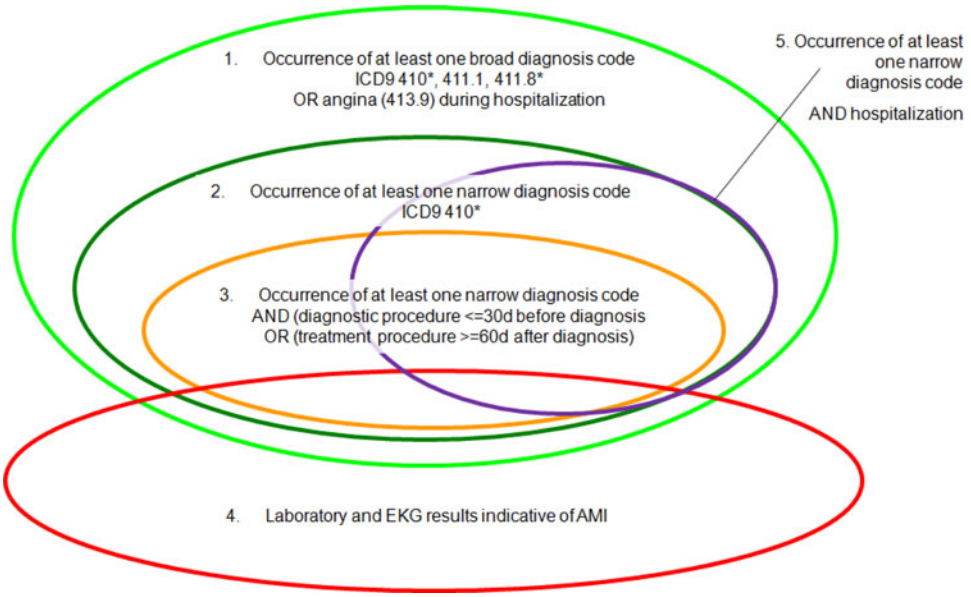

Fig. 3 Alternative definitions of AMI. The Venn diagram indicating the nested relationship of the five alternative definitions. An *asterisk* indicates a wildcard, i.e. any code with or without additional digits is included in the definition. *AMI* acute myocardial infarction, *ICD-9* International Classification of Diseases, Ninth Revision, *EKG* electrocardiogram

46.5 m patients); and General Electric Centricit (GE, 11.2 m patients) database. GE is an EHR database, the other three databases contain administrative claims data. The data are described in more detail elsewhere [10].

The methods were executed against a set of drug-outcome pairs to generate an effect estimate and standard error for each pair and parameter combination. The test cases included positive controls—active ingredients with evidence to

Table 1 Optimal method/analysis choice combination for best average AUC across alternative HOI definitions for each database. Analysis choices have a six-digit analysis ID. For details see http://omop.org/Research

Database	Acute kidney injury	Acute liver injury	Acute myocardial infarction
MDCR	**OS: 401002**	**OS: 401002**	**OS: 407002**
	Study design: Self-controlled cohort	Study design: Self-controlled cohort	Study design: Self-controlled cohort
	Exposures to include: All occurrences	Exposures to include: All occurrences	Exposures to include: All occurrences
	Outcomes to include: All occurrences	Outcomes to include: All occurrences	Outcomes to include: First occurrence
	Time-at-risk: Length of exposure + 30d	Time-at-risk: Length of exposure + 30d	Time-at-risk: Length of exposure + 30d
	Include index date in time-at-risk: No	Include index date in control period: No	Include index date in time-at-risk: No
	Control period: Length of exposure + 30d		Control period: Length of exposure + 30d
	Include index date in control period: No		Include index date in control period: No
MDCD	**OS: 408013**	**OS: 409013**	**OS: 407004**
	Study design: Self-controlled cohort	Study design: Self-controlled cohort	Study design: Self-controlled cohort
	Exposures to include: First occurrence	Exposures to include: First occurrence	Exposures to include: All occurrences
	Outcomes to include: First occurrence after exposure	Outcomes to include: First occurrence	Outcomes to include: First occurrence
	Time-at-risk: All time post-exposure start	Time-at-risk: All time post-exposure start	Time-at-risk: Length of exposure + 30d
	Include index date in time-at-risk: No	Include index date in time-at-risk: No	Include index date in time-at-risk: No
	Control period: All time prior to exposure start	Control period: All time prior to exposure start	Control period: 365d prior to exposure start
	Include index date in control period: No	Include index date in control period: No	Include index date in control period: No
CCAE	**OS: 404002**	**OS: 403002**	**OS: 408013**
	Study design: Self-controlled cohort	Study design: Self-controlled cohort	Study design: Self-controlled cohort
	Exposures to include: All occurrences	Exposures to include: All occurrences	Exposures to include: First occurrence
	Outcomes to include: All occurrences	Outcomes to include: All occurrences	Outcomes to include: First occurrence after exposure
	Time-at-risk: Length of exposure + 30d	Time-at-risk: Length of exposure + 30d	Time-at-risk: All time post-exposure start
	Include index date in time-at-risk: No	Include index date in time-at-risk: Yes	Include index date in time-at-risk: No
	Control period: Length of exposure + 30d	Control period: Length of exposure + 30d	Control period: All time prior to exposure start
	Include index date in control period: Yes	Include index date in control period: No	Include index date in control period: No
GE	**SCCS: 1949010**	**OS: 409002**	**ICTPD: 3016001**
	Outcomes to include: All occurrences	Study design: Self-controlled cohort	Control period: −1080d to −361d before exposure start
	Prior distribution: normal	Exposures to include: First occurrence	Time-at-risk: 60d from exposure start
	Variance of the prior: Determined through cross-validation	Outcomes to include: First occurrence	Use control period in expected calculation: Yes
	Time-at-risk: 30d from exposure start	Time-at-risk: Length of exposure + 30d	Use 1mo prior to exposure in expected calculation: Yes
	Include index date in time-at-risk: Yes	Include index date in time-at-risk: No	Use 1d prior to exposure in expected calculation: No
	Apply multivariate adjustment on all drugs: Yes	Control period: Length of exposure + 30d	
	Required observation time: 180d	Include index date in control period: No	

MDCR MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *MDCD* MarketScan Multi-state Medicaid, *GE* GE Centricity

suspect a positive association with the outcome—and negative controls—active ingredients with no evidence to expect a causal effect with the outcome. The full set of test cases and its construction is described elsewhere [11]. For each HOI and database, the method/analysis choice combination with the highest average AUC across all definitions of an HOI was used (Table 1). These top performing analyses all involved a Self-Controlled Cohort design, except for AMI in the GE database where a Self-Controlled Case Series approach is the best performer (for details of method performance see [8]). For every database and HOI we restricted the analysis to those drug-outcome pairs where a relative risk of at least 1.25 could be detected with 95 % confidence and 80 % power, based on the age-by-gender-stratified drug and outcome prevalence estimates [12, 13].

2.6 Metrics

The alternative HOI definitions were applied to evaluate two performance characteristics. One characteristic was

discrimination, i.e. the ability of methods to distinguish between positive and negative controls. The relative risk estimates were used to compute an Area Under the Receiver Operator Characteristics Curve (AUC) [14]. An AUC of 1 indicates a perfect distinction of positive test cases from negative ones. An AUC of 0.5 is equivalent to random guessing. The second performance characteristic was bias, i.e. the difference between the true and the estimated relative risk. Since the precise relative risk of outcomes known to be caused by a drug is not known for positive controls, we used only the negative test cases to calculate the bias based on the assumption that the true relative risk must be equal to 1.

3 Results

3.1 Prevalence of Alternative HOI Definitions

Table 2 shows the prevalence of the alternative HOI definitions in the four databases studied. Prevalence estimates varied substantially for each outcome and between alternative definitions of the same HOI. The degree of these differences depends mostly on the availability of certain codes in a given database. For example, definitions #1 and #3 for ALI differ by a factor of about 4,500 due to the fact that in addition to diagnostic codes, #3 requires procedures for the diagnosis of ALI (liver biopsy or liver imaging) or for the therapy of ALI (liver transplantation in fulminant cases). However, GE is a database of outpatient EHR records, and those procedures are typically carried out in hospitalized patients.

Some definitions produce empty cohorts. For example, AKI definition #2 requires hemodialysis, a procedure usually carried out in specialty practices not typically covered by the GE Centricity EHR system, and definition #3 requires results of serum creatinine tests, not available in the 3 claims databases MDCR, MDCD and CCAE. The relative size of the population for each definition is comparable between databases. However, MDCR numbers are generally higher than in any of the other databases, reflecting the higher prevalence of these diseases in the population characterized by higher age and morbidity.

3.2 Test Cases with Sufficient Predictive Power in Alternative HOI Definitions

Depending on the prevalence of outcome and drug in each age and gender stratum of each database, the alternative

Table 2 Prevalence of alternative HOI definitions in the four studied databases

HOI and definition #	MDCR		MDCD		CCAE		GE	
	Patients	%	Patients	%	Patients	%	Patients	%
Acute kidney injury								
1	201,769	4.43	97,873	0.91	114,472	0.25	12,553	0.11
2	5,741	0.13	10,929	0.10	6,353	0.01	–	–
3	–	–	–	–	–	–	226,942	2.02
4	201,769	4.43	97,873	0.91	114,472	0.25	12,553	0.11
5	183,298	4.02	87,150	0.81	99,070	0.21	212	0.002
Acute liver injury								
1	264,122	5.80	274,437	2.55	1,235,711	2.66	186,677	1.66
2	159,370	3.50	141,236	1.31	626,239	1.35	46,687	0.42
3	9,931	0.22	9,478	0.09	37,598	0.08	41	0.000
4	–	–	–	–	–	–	–	–
5	–	–	–	–	–	–	6,452	0.06
6	–	–	–	–	–	–	2,067	0.02
7	7,933	0.17	18,249	0.17	42,738	0.09	2,738	0.02
8	121,332	2.66	104,235	0.97	399,037	0.86	76	0.00
Acute myocardial infarction								
1	665,396	14.61	221,780	2.06	731,792	1.58	140,336	1.25
2	186,878	4.10	64,411	0.60	149,586	0.32	44,654	0.40
3	161,353	3.54	51,109	0.47	132,391	0.28	3,892	0.03
4	–	–	–	–	–	–	–	–
5	171,877	3.77	55,700	0.52	135,680	0.29	225	0.002

HOI health outcome of interest, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *MDCD* MarketScan Multi-State Medicaid, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

Table 3 Number of positive and negative test cases for each HOI as defined and as available in each database based on a minimal detectable relative risk cut-off of 1.25

HOI definition	MDCR		MDCD		CCAE		GE		Defined	
	−	+	−	+	−	+	−	+	−	+
Acute kidney injury										
1	41	19	42	19	34	19	6	8	64	24
2	8	9	13	10	3	7				
3			1	4			44	19		
4	41	19	42	19	34	19	6	8		
5	40	19	40	19	34	19				
Acute liver injury										
1	28	51	29	58	32	63	26	48	37	79
2	25	48	27	52	32	57	14	32		
3	12	21	9	20	12	29				
4										
5							3	11		
6							1	2		
7	10	19	12	26	12	29	1	4		
8	24	46	25	49	30	54				
Acute myocardial infarction										
1	48	30	50	29	46	33	39	25	66	36
2	41	24	33	21	37	26	25	16		
3	41	24	30	21	36	26	3	2		
4										
5	41	24	32	21	36	26				

HOI health outcome of interest, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *MDCD* MarketScan Multi-State Medicaid, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity, *Defined* test cases defined for the HOI irrespective of availability in databases

definitions for each HOI have different numbers of negative and positive test cases with sufficient power to detect a RR of 1.25 (Table 3). For some HOIs with low prevalence numbers there were only few or even no test cases with enough power. For example, in GE there are only two positive (Lisinopril and Ibuprofen) and one negative drug test case (Fluticasone) with sufficient power to participate in the performance evaluation for Acute Liver Injury definition 6.

3.3 Discrimination of Alternative HOI Definitions

Figure 4 shows an example of the Receiver Operator Characteristics Curves and the AUC values for the alternative definitions of Acute Kidney Injury testing the test cases with sufficient power available in CCAE. Note that with only three negative and seven positive test cases as in AKI 2 the AUC loses precision. Figure 5 shows the comparison of all alternative HOI definitions across the four databases using the optimal method/analysis choice combination for the HOI and database. In general, the AUCs are fairly constant within each database, but start fluctuating if the number of test cases is very low. However, when comparing these AUC numbers it is important to keep in mind that they are generated from a set of test cases that is different for each outcome definition due to the availability of sufficient samples.

3.4 Bias

In order to gain insight into the bias inherent in the effect estimates when using different HOI definition, we plotted the estimates of the effect sizes of the negative controls in Fig. 6 produced, again using the method-analysis choice combination with the best AUC as described in Ryan et al. [8] Because we assume the true effects sizes for these negative controls is 1, we would like the estimates to be closely centered around one, which is the case for the majority of estimates. We furthermore see that the distribution of estimates does not change dramatically with different definitions of the HOI.

In sum, the alternative HOI definitions we studied resulted in very different cohort sizes, which had a substantial effect on the availability of positive and negative test cases that could be studied with sufficient power.

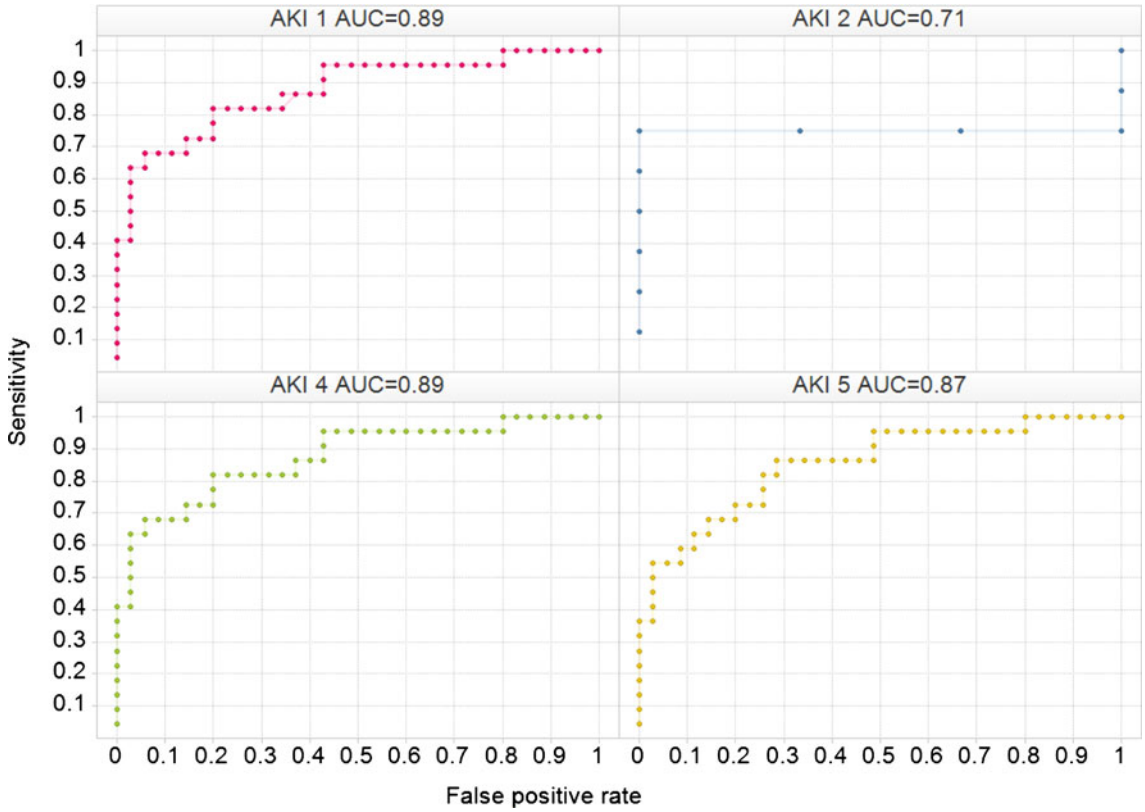

Fig. 4 Receiver-Operator Characteristics Curve and Area Under the Curve for Acute Kidney Failure alternate definitions in the CCAE database. *CCAE* MarketScan Commercial Claims and Encounters, *AKI* acute kidney injury, *AUC* area under the curve

Despite this, the alternative definitions faired comparably in the ability to discriminate signal from noise, or positive from negative drug-outcome test cases in the methods tested. They also did not differ notably in the bias generated by methods, as measured for negative test cases with an expected relative risk of 1. In particular, narrow definitions, which are believed to retrieve cases more accurately reflecting the outcome, are not generally superior to broader, more inclusive definitions. On the contrary, narrow definitions generated fewer cases, which reduced the sample size a method has to work with, which in turn increases the confidence interval of a point estimate and reduces power.

4 Discussion

The wide variety of ways to define a clinical outcome in the literature poses the question which of these alternative definitions is the best choice for drug outcome studies. Generally, the assumption is that the quality of an outcome definition will influence the amount of misclassification, i.e. the degree by which a definition fails to identify outcome cases (false negative) or retrieves non cases (false positives). Definitions with the smallest rate of misclassification are said to produce the most accurate study results [15].

Since the number of cases is generally much smaller than the number of non cases, the emphasis is usually put on ensuring the highest possible positive predictive value (PPV) for the case group only (case ascertainment). In order to determine this PPV, source record verification or validation studies are undertaken to compare the cases retrieved through the definition with the content of the patient charts. However, source record verification is a very laborious and expensive undertaking and not all sources of clinical data allow access to protected patient information. As a putative gold-standard, it can also introduce bias with the potential of distorting the classification it tries to measure [16, 17].

Our results show that in the three HOIs alternative definitions had an influence on the performance of methods, but this effect was not systematic or substantial. The reasons for this somewhat counterintuitive finding remain unclear and require more research. We would speculate several possible explanations:

1. Bias due to misclassification is a function of the sensitivity and specificity of the outcome definition and the frequency of drug and outcome in the database.

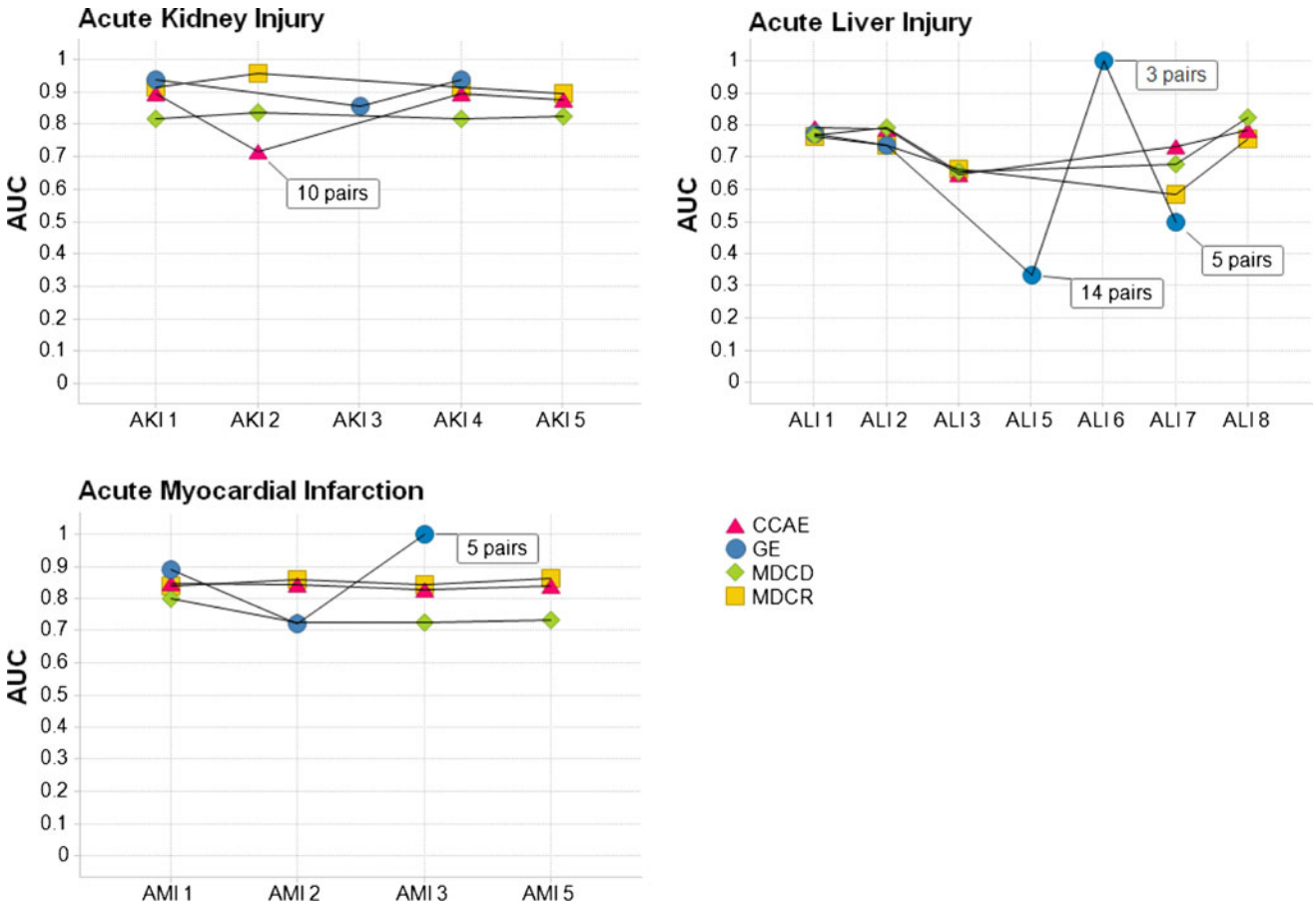

Fig. 5 AUC values of alternative definitions for each of the three HOI across the four databases. Definitions with only few test cases are labeled. *Upper left* diagram *red* markers are equivalent to Fig. 4. *HOI* health outcome of interest, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

However, the dependency on these parameters could have reached a plateau, i.e. the amount of bias is not sensitive to changes of those frequencies given the drugs and outcomes we studied.

2. The improved classification and the reduced bias that comes with a precise narrow definition might be offset by the loss in sample size and precision of the method.
3. The etiology of the outcome makes it somewhat robust against misclassification, if the false positive cases are biologically related to the studied outcome and therefore not all that "false". For example in the case of AMI, the platelet aggregation and vasoconstriction caused by COX-2 inhibitors cause myocardial infarction, but other ischemic cardiovascular events such as unstable angina, cardiac thrombus or resuscitated cardiac arrest can be equally triggered by the same mechanism [18]. So, administrative claims submitted for patients with acute symptoms that were admitted to rule out an acute myocardial infarction might nominally have the wrong diagnostic code, but reflect acute ischemic events triggered by a drug just as patients with the confirmed diagnosis.

We want to emphasize that our findings are based on a very small set of outcomes in internal medicine, all of which are very acute events requiring hospitalization and intensive treatment. These findings need to be reproduced in a wider selection of therapeutic areas, healthcare delivery mechanisms, database types and complexity of making the diagnoses, in order to be considered for generalization. We also did not conduct a validation study, but if the general trends of our findings are upheld, we would postulate an alternative approach for selecting the best outcome definitions. This approach would test a set of candidate definitions using negative and positive test cases and select the one with the highest performance in dis-

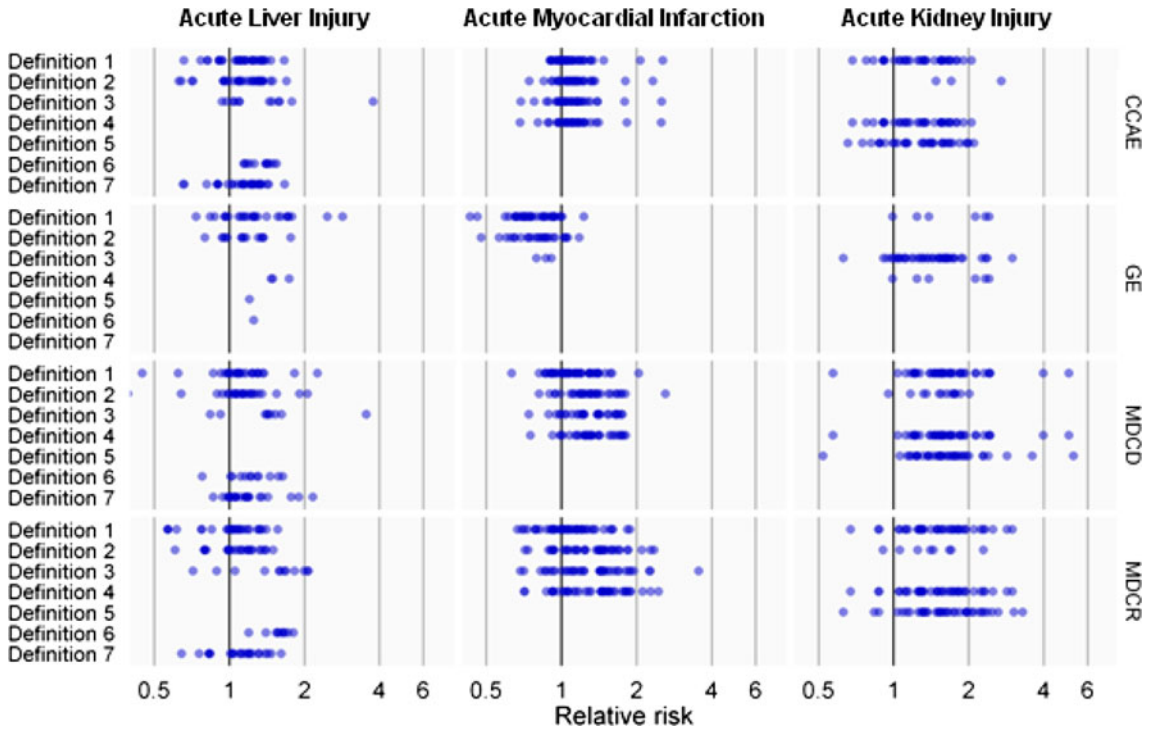

Fig. 6 Bias of alternative definitions for each of the three HOI across the four databases. Each *dot* shows the estimated effect estimate for one negative control drug-outcome pair, where the true effect estimate is assumed to be 1. Only estimates where there was enough power to detect a RR of 1.25 are shown. *HOI* health outcome of interest, *MDCD* MarketScan Multi-state Medicaid, *MDCR* MarketScan Medicare Supplemental Beneficiaries, *CCAE* MarketScan Commercial Claims and Encounters, *GE* GE Centricity

crimination, the least bias and the largest resulting sample size for optimal power.

5 Conclusions

Our studies have shown that narrow definitions do not lead to better performance of methods differentiating between signal and noise or estimating the effect size of medical treatments leading to outcomes of interest. Instead, broader definitions allow studying drugs with lower prevalence because the improved sample size increases the power of the statistics.

Acknowledgments The Observational Medical Outcomes Partnership is funded by the Foundation for the National Institutes of Health (FNIH) through generous contributions from the following: Abbott, Amgen, AstraZeneca, Bayer Healthcare Pharmaceuticals, Biogen Idec, Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Janssen Research and Development, Lundbeck, Merck & Co., Novartis Pharmaceuticals Corporation, Pfizer, Pharmaceutical Research Manufacturers of America (PhRMA), Roche, Sanofi-Aventis, Schering-Plough Corporation, and Takeda. Dr. Reich is an employee of AstraZeneca. Drs. Ryan and Schuemie are employees of Janssen Research and Development. Dr. Schuemie received a fellowship from the Office of Medical Policy, Center for Drug Evaluation and Research, Food and Drug Administration. Dr. Schuemie has previously received a grant from FNIH.

This article was published in a supplement sponsored by the Foundation for the National Institutes of Health (FNIH). The supplement was guest edited by Stephen J.W. Evans. It was peer reviewed by Olaf H. Klungel who received a small honorarium to cover out-of-pocket expenses. S.J.W.E has received travel funding from the FNIH to travel to the OMOP symposium and received a fee from FNIH for the review of a protocol for OMOP. O.H.K has received funding for the IMI-PROTECT project from the Innovative Medicines Initiative Joint Undertaking (http://www.imi.europa.eu) under Grant Agreement no 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

Appendix: Health Outcome of Interest Definitions Used in the Study

HOI	#	Definition
Acute liver injury	1	Occurrence of at least one diagnostic code ICD-9-CM: • 277.4 'Disorders of bilirubin excretion' • 570* 'Acute and subacute necrosis of the liver'[a] • 572.2 'Hepatic coma (hepatorenal syndrome)' • 572.4* 'Hepatorenal syndrome'[a] • 573* 'Other disorders of the liver, including chemical or drug induced'[a] • 576.8 'Other specified disorders of biliary tract' • 782.4 'Jaundice, unspecified, not of newborn' • 789.1* 'Hepatomegaly'[a] • 790.4* 'Nonspecific elevation of transaminase or lactic dehydrogenase levels'[a] • 794.8* 'Abnormal liver function test results'[a]
	2	Occurrence of at least one diagnostic code ICD-9-CM: • 570* 'Acute and subacute necrosis of the liver'[a] • 572.2 'Hepatic coma (hepatorenal syndrome)' • 572.4* 'Hepatorenal syndrome'[a] • 573* 'Other disorders of the liver, including chemical or drug induced'[a]
	3	Definition Acute Liver Injury #2 AND Occurrence of at least one diagnostic procedure code within 30 days prior to diagnostic code[a] AND Occurrence of at least one therapeutic procedure code within 60 days after the diagnostic code[b]
	4	Definition Acute Liver Injury #2 AND Occurrence of at least one diagnostic procedure code within 30 days prior to diagnostic code[b] AND Occurrence of at least one therapeutic procedure code within 60 days after the diagnostic code[b] AND Definition Acute Liver Injury #5
	5	Indicative of Hy's law: Occurrence of the following lab tests and test results within 7 day period: Alanine aminotransferase $\geq$3 times the upper limit of normal as defined in data source, or 40 IU/L if not available OR Aspartate aminotransferase $\geq$3 times the upper limit of normal as defined in data source, or 40 IU/L if not available} AND Total bilirubin $\geq$2 times the upper limit of normal as defined in data source, or 1.2 mg/dL if not available
	6	Strongly indicative of Hy's law: Occurrence of the following lab tests and test results within 7 day period: Alanine aminotransferase $\geq$10 times the upper limit of normal as defined in data source, or 40 IU/L if not available OR Aspartate aminotransferase $\geq$10 times the upper limit of normal as defined in data source, or 40 IU/L if not available AND Total bilirubin $\geq$2 times the upper limit of normal as defined in data source, or 1.2 mg/dL if not available
	7	Occurrence of at least one diagnostic code ICD-9-CM: • 570* 'Acute and subacute necrosis of the liver'[a] • 572.4* 'Hepatorenal syndrome'[a] • 573.0 'Chronic passive congestion of liver' • 573.1 'Hepatitis in viral diseases classified elsewhere' • 573.4 'Hepatic infarction'
	8	Definition of Acute Liver Failure #2 AND Hospitalization at date of diagnostic code

continued

HOI	#	Definition
Acute kidney injury	1	Occurrence of at least one diagnostic code ICD-9-CM: • 584* 'Acute renal failure'[a]
	2	Definition Acute Kidney Injury #1 AND Occurrence of at least one therapeutic ICD-9 Procedure code within 60 days after diagnostic code: • 39.95 'Hemodialysis' • 54.98 'Peritoneal dialysis' AND Excluding any diagnostic code for chronic dialysis status anytime before the diagnostic code ICD-9-CM: • V45.1 'Renal dialysis status' • V56.0 'Encounter for dialysis and dialysis catheter care' • V56.31 'Encounter for adequacy testing for hemodialysis' • V56.32 'Encounter for adequacy testing for peritoneal dialysis' • V56.8 'Other dialysis'
	3	Occurrence of the lab test code LOINC 2160-0 and the following test results: • Serum creatinine $\geq$0.5 mg/dL for patients with a baseline level of $\leq$1.9 mg/dL • Serum creatinine $\geq$1.0 mg/dL for patients with a baseline level of 2.0–4.9 mg/dL • Serum creatinine $\geq$1.5 mg/dL for patients with a baseline level $\geq$5.0 mg/dL The baseline level is defined as the occurrence of most recent lab test result any time prior to the elevated test result
	4	Occurrence of at least one diagnostic code ICD-9-CM: • 584 'Acute renal failure' 584.5 'Acute renal failure with lesion of tubular necrosis' • 584.6 'Acute renal failure with lesion of renal cortical necrosis' • 584.7 'Acute renal failure with lesion of renal medullary (papillary) necrosis'
	5	Definition of Acute Kidney Injury #1 AND Hospitalization at date of diagnostic code
Acute myocardial infarction	1	Occurrence of at least one broad diagnostic code ICD-9-CM: • 410* 'Acute myocardial infarction'[a] • 411.1 'Intermediate coronary syndrome' • 411.8 'Other acute coronary occlusion' • 413.9 'Other and unspecified angina pectoris' on or during hospitalization
	2	Occurrence of at least one narrow diagnostic code ICD-9-CM: • 410* 'Acute myocardial infarction'[a]
	3	Definition Acute Myocardial Infarction #2 AND Occurrence of at least one diagnostic procedure code within 30 days prior to diagnostic code[a] OR Occurrence of at least one therapeutic procedure code within 60 days after the diagnostic code[c]
	4	American College of Cardiology/European Society of Cardiology Consensus Definition: Occurrence of the following lab test results: • Occurrence of the lab test 'Blood Troponin' LOINC codes 42757-5 or 10839-9, and lab test results $\geq$ the upper limit of normal in two sequential measurements, or $\geq$2 times the upper limit of normal in one measurement, and falling in a subsequent measurement. The ULN is defined as the 99th percentile of a non-MI control group, or 0.5 ng/mL if not available AND • Occurrence of the lab test 'Creatinine Phosphokinase MB Isozyme (CK-MB)' LOINC codes 49551-5 or 13969-1, and lab test results $\geq$ the upper limit of normal in two sequential measurements, or $\geq$2 times the upper limit of normal in one measurement, and falling in a subsequent measurement. The ULN is defined as the 99th percentile of a non-MI control group, or 6 ng/mL if not available AND Occurrence of an EKG test with the LOINC codes 11524-6, 8601-7, 18843-3, 18844-1, 18810-2, 8625-6 or 8634-8 within 10 days prior or after the lab test result and any of the following readings • Any Q wave in leads V1 through V3, Q wave $\geq$5 to 30 ms in leads I, II, aVL, aVF, V4, V5, or V6 OR ST segment elevation: New or presumed new ST segment elevation at the J point in two or more contiguous leads with the cut-off of 0.2 mV in leads V1, V2, or V3 and $\geq$0.1 mV in other leads (contiguity in the frontal plane is defined by the lead sequence aVL, I, inverted aVR, II, aVF, III) • ST segment depression OR T wave abnormalities
	5	Definition Acute Myocardial Infarction #2 AND Hospitalization at date of diagnostic code

[a] An asterisk indicates a wildcard, i.e. any code with or without additional digits is included in the definition

[b] A detailed list of all codes are available at http://omop.org/AcuteLiverInjury

[c] A detailed list of all codes are available at http://omop.org/AcuteMyocardialInfarction

References

1. Barron BA. The effects of misclassification on the estimation of relative risk. Biometrics. 1977;33(2):414–8.
2. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Int Med. 2010;153(9):600–6.
3. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf. 2012;21(Suppl 1):82–9.
4. Stang PE, Ryan PB, Dusetzina SB, Hartzema AG, Reich C, Overhage JM, et al. Health outcomes of interest in observational data: issues in identifying definitions in the literature. Health Outcomes Res Med. 2012;3(1):e37–44.
5. Kellum JA, Bellomo R, Ronco C. Definition and classification of acute kidney injury. Nephron Clinical Pract. 2008;109(4):c182–7.
6. James M, Pannu N. Methodological considerations for observational studies of acute kidney injury using existing data sources. J Nephrol. 2009;22(3):295–305.
7. Katz AJ, Ryan PB, Racoosin JA, Stang PE. Assessment of case definitions for identifying acute liver injury in large observational databases. Drug Saf. 2013;36(8):651–61.
8. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0108-9.
9. Observational Medical Outcomes Partnership Methods Library; 2012. http://omop.org/HOI [cited 2012 December 13].
10. Overhage JM, Ryan PB, Schuemie MJ, Stang PE. Desideratum for Evidence Based Epidemiology. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0102-2
11. Ryan PB, Schuemie MJ, Welebob E, Duke J, Valentine S, Hartzema AG. Defining a reference set to support methodological research in drug safety. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0097-8
12. Armstrong B. A simple estimator of minimum detectable relative risk, sample size, or power in cohort studies. Am J Epidemiol. 1987;126(2):356–8.
13. Reich C, Ryan PB, Suchard MA. The impact of drug and outcome prevalence on the feasibility and performance of analytical methods for a risk identification and analysis system. Drug Saf (in this supplement issue). doi:10.1007/s40264-013-0112-0
14. Cantor SB, Kattan MW. Determining the area under the ROC curve for a binary diagnostic test. Med Decis Making. 2000;20(4):468–70.
15. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977;105(5):488–95.
16. Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. Am J Epidemiol. 1993;137(11):1251–8.
17. Evans JM, MacDonald TM. Misclassification and selection bias in case-control studies using an automated database. Pharmacoepidemiol Drug Saf. 1997;6(5):313–8.
18. Mukherjee D, Nissen SE, Topol EJ. Risk of cardiovascular events associated with selective COX-2 inhibitors. JAMA. 2001;286(8):954–9.