Drug Saf (2013) 36 (Suppl 1):S27–S32
DOI 10.1007/s40264-013-0104-0

ORIGINAL RESEARCH ARTICLE

How Well Do Various Health Outcome Definitions Identify Appropriate Cases in Observational Studies?

Richard A. Hansen · Michael D. Gray · Brent I. Fox · Joshua C. Hollingsworth · Juan Gao · Peng Zeng

Abstract

Background Observational data can be useful for drug safety research, but accurate measurement of adverse health outcomes is paramount. Best practices for identifying important health outcomes of interest (HOI) are needed.

Objectives To evaluate the extent to which health outcome definitions commonly used in observational database studies identify cases that are consistent with expert panel assessment of the underlying data.

Methods Competing HOI definitions were used to identify potential cases of acute liver injury (ALI; n = 208), acute kidney injury (AKI; n = 200), and myocardial infarction (MI; n = 204) in the Truven MarketScan Lab Database (MSLR). Panelists reviewed patient-level data and answered questions about whether they believed the case actually reflected the HOI and their certainty of case classification on a 10-point scale (1 = unlikely to 10 = likely). Each patient was reviewed independently by two panelists. Case disagreements were resolved through consensus meetings. Positive predictive value (PPV) was calculated as the number of cases deemed to be true over the total number of sampled cases. Kappa statistics assessed inter-rater agreement.

Results PPV ranged from 0 to 52 % across ALI definitions, 12 to 82 % across AKI definitions, and 1 to 56 % across MI definitions. Certainty scores on the 10-point scale paralleled the PPV, with a range of mean values from 1.7 to 4.8 across ALI definitions, 3.1 to 6.0 across AKI definitions, and 2.8 to 5.7 across MI definitions. Inter-rater agreement was low to moderate (Kappa range 0.0–0.6).

Implications/Conclusions Existing HOI definitions had relatively low PPV based on expert panel review. Experts commonly disagreed on case classification. Additional work is needed to refine HOI case definitions if observational data are to be reliably used for health outcome assessment.

1 Background

Adverse drug events (ADEs) are common and costly. Improved methods for drug safety surveillance are a public health priority. Observational data sources such as electronic medical records and administrative claims offer opportunity for enhanced safety surveillance. Various approaches to identifying ADEs in observational data exist, but the literature is inconsistent in best practices for defining important health outcomes of interest (HOIs) [1].

To address inconsistent operational definitions for HOI measures, the Observational Medical Outcomes Partnership (OMOP) commissioned a series of systematic reviews of existing HOI definitions for important drug safety concerns [1]. These reviews were used to develop a library of

The research completed in this paper used Truven Health Analytics (formerly the Health Business of Thomson Reuters) data from the MarketScan Lab Supplemental database (MSLR) between January 1, 2003 and December 31, 2007.

R. A. Hansen (✉) · B. I. Fox · J. C. Hollingsworth · J. Gao
Department of Pharmacy Care Systems, Harrison School of Pharmacy, Auburn University, 020 Foy Hall, Auburn, AL 36849, USA
e-mail: rah0019@auburn.edu

M. D. Gray
HP Labs, Palo Alto, CA, USA

P. Zeng
College of Science and Mathematics, Auburn University, Auburn, AL, USA

HOI definitions. Among others, the library includes acute liver injury (ALI, http://omop.org/AcuteLiverInjury), acute kidney injury (AKI, http://omop.org/AcuteRenalFailure), and acute myocardial infarction (MI, http://omop.org/AcuteMyocardialInfarction). The OMOP definitions cover broad and narrowly defined concepts, using diagnostic codes, procedure codes, and lab tests [2].

Subsequent work conducted with the OMOP HOI definitions has revealed large variance in the prevalence of health outcomes across data sources [3, 4]. While previous studies have evaluated the performance of specific diagnostic codes or algorithms compared with medical record review [5–7], the OMOP definitions have not been previously evaluated, and in general the validity and reliability of competing measurement definitions is poorly understood. Evaluation of the performance of different HOI definitions is needed to improve definitions so that drug safety researchers using observational data can interpret results of their analyses in context of the accuracy of the HOI definition. This paper evaluates the positive predictive value (PPV) of select HOI definitions in the OMOP library through expert panel review of observational data. Specifically, alternative definitions were compared within three important HOIs, including ALI, AKI, and MI.

2 Methods

2.1 Study Design and Data Source

We conducted an expert panel review of cases meeting HOI definitions for ALI, AKI, and MI (Fig. 1) to understand how alternative definitions perform in case identification [8]. Potential cases were obtained using de-identified patient-level data from the Truven MarketScan® Lab database (MSLR) between January 1, 2003 and December 31, 2007 [9]. The MSLR data includes administrative claims from inpatient, outpatient, and pharmacy providers for more than 1.2 million persons, merged with laboratory data [10]. This includes claims from large employers, managed care organizations, hospitals, Medicare, and Medicaid programs. The MSLR claims data include International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes, Current Procedural Terminology (CPT) codes for billing procedures and services, Healthcare Common Procedural Coding System (HCPCS) codes, and Revenue Codes. These claims data are supplemented with the results of laboratory measures provided by laboratory vendors. The MSLR data were transformed to the OMOP Common Data Model which, combined with a method to standardize its content, ensures that research methods are applied systematically to produce meaningful and comparable results from disparate data sources [9]. The study was approved by the Auburn University Institutional Review Board.

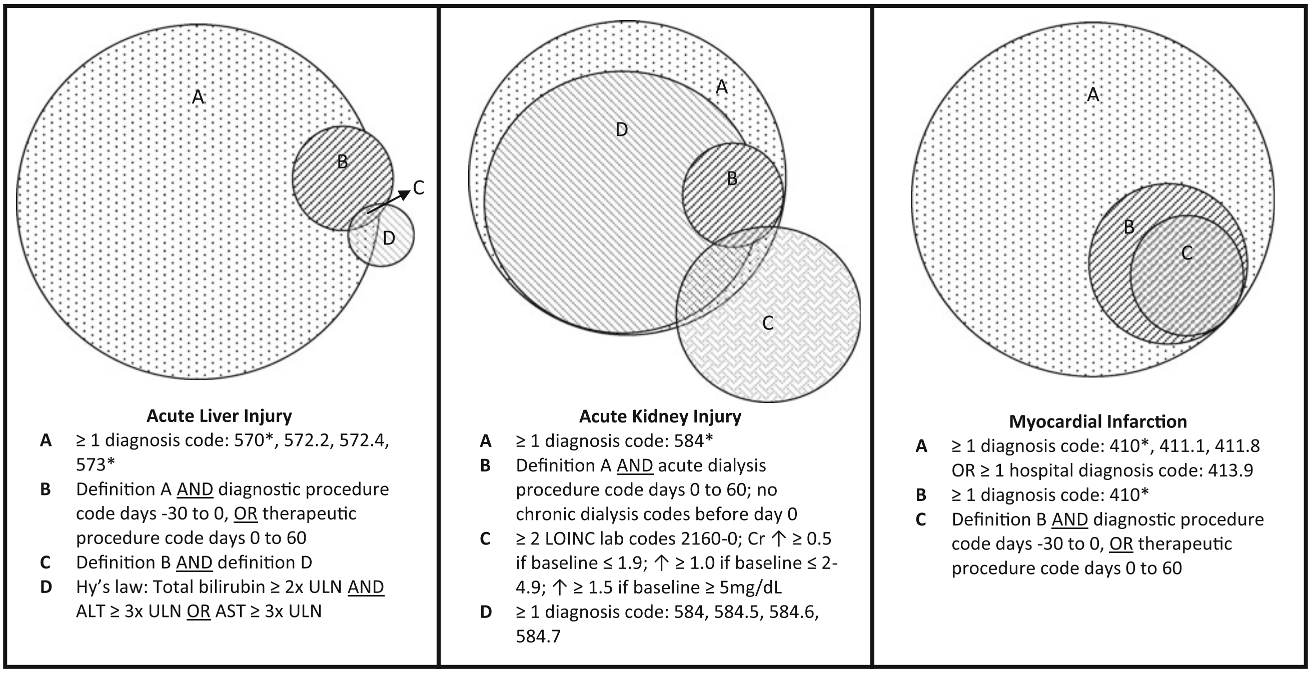
Approximately 200 patient cases meeting the competing HOI definitions were sampled per HOI. From the sampling frame of all patients that met an eligible case definition, we sampled (without replacement) all cases from the smallest sized definition cohort first and then sampled from each additional definition cohort in order from the smallest to the largest. Patients were assigned to the definition from which they were sampled, even though some patients may have satisfied more than one definition. Patient data, including demographics, diagnosis codes, healthcare encounters, procedures, and lab values and their relative timing, were then presented to expert panelists via a secure Web portal for case review.

2.2 Expert Panel Review Process

Eight panelists were recruited, paired, and then placed into one of two panels for each HOI. This allowed for independent, dual review of each case. Panelist pairs for each HOI included an expert with specialized knowledge for the HOI (e.g., nephrologist for AKI) as well as a physician panelist that had practice experience and expertise in observational data research. Panelists attended a Web-based videoconference to orient them to the case review process and Web dashboard. Orientation included a summary document containing descriptions of OMOP's HOI operational definitions, a summary of relevant diagnoses, procedures, and labs used to create filters in the Web dashboard, and a descriptive summary of population characteristics for patients identified with each definition. Panelists were asked to identify any elements related to the HOI that they believed were especially salient, and we used this feedback to ensure that the requested data were easy to find in the Web Dashboard filters. Once the case review process was launched, each panelist independently logged into the dashboard and reviewed their assigned cases (roughly 100 cases per panelist per HOI total, proportionately sampled across HOI definitions).

After reviewing patient data, panelists were asked to respond to a series of questions, including the following: (1) Do you believe this patient has the HOI (i.e., case classification) and (2) Rate the likelihood that this patient has the HOI on a 10-point scale, where 1 = definitely NOT a case and 10 = definitely a case (i.e., case certainty scale).

Once initial case reviews were complete, answers from panelist pairs were compared with respect to their determination of case classification. A consensus process was used to resolve disagreement among panelists on this question. The consensus process presented panelist pairs with their answers to question 1 (case classification) and



Acute Liver Injury

A ≥ 1 diagnosis code: 570*, 572.2, 572.4, 573*

B Definition A AND diagnostic procedure code days -30 to 0, OR therapeutic procedure code days 0 to 60

C Definition B AND definition D

D Hy's law: Total bilirubin ≥ 2x ULN AND ALT ≥ 3x ULN OR AST ≥ 3x ULN

Acute Kidney Injury

A ≥ 1 diagnosis code: 584*

B Definition A AND acute dialysis procedure code days 0 to 60; no chronic dialysis codes before day 0

C ≥ 2 LOINC lab codes 2160-0; Cr ↑ ≥ 0.5 if baseline ≤ 1.9; ↑ ≥ 1.0 if baseline ≤ 2-4.9; ↑ ≥ 1.5 if baseline ≥ 5mg/dL

D ≥ 1 diagnosis code: 584, 584.5, 584.6, 584.7

Myocardial Infarction

A ≥ 1 diagnosis code: 410*, 411.1, 411.8 OR ≥ 1 hospital diagnosis code: 413.9

B ≥ 1 diagnosis code: 410*

C Definition B AND diagnostic procedure code days -30 to 0, OR therapeutic procedure code days 0 to 60

Fig. 1 Operational definitions and definitional overlap for identification of possible health outcomes of interest. *All cases with these first three digits included. The relative size of the overlapping circles is not drawn to scale for comparison across each health outcome; *ALT* alanine aminotransferase, *AST* aspartate aminotransferase, *Cr* creatinine (serum), *ULN* upper limit of normal, *LOINC* logical observation identifier names and codes. (http://omop.org/HOI)

question 2 (case certainty scale) as well as any comments from their initial review. Consensus discussions occurred during videoconference meetings, with a shared desktop image of the dashboard so patient and case review data could be viewed by both panelists. Through mediated discussion, panelists were asked to report consensus answers.

2.3 Statistical Analysis

The PPV of each HOI definition was calculated as the number of cases deemed true by panelists divided by the total number of sampled cases. This calculation was performed at the patient level, with 2 panelist reviewers per patient. For cases in which panelists initially disagreed on case classification, consensus panels determined the final patient classification. We calculated means and standard deviations for the 10-point case certainty question using the average of the two panelists' responses when panelists agreed on case classification, and replaced the certainty value with a consensus value for cases of disagreement. Kappa statistics assessed inter-rater agreement.

3 Results

Panelist case reviews occurred between November 2011 and April 2012, reviewing 208 ALI cases (416 reviews), 200 AKI cases (400 reviews), and 204 MI cases (408 reviews). Panelists did not always agree on case classification, including disagreements for 34 (16 %) ALI cases, 78 (39 %) AKI cases, and 41 (21 %) MI cases. All disagreements were resolved successfully via a moderated consensus meeting. Kappa statistics across the panels ranged from 0.0 to 0.6.

The number of cases identified as true positives generally was low (Table 1). Across the three HOIs, 174 of 612 reviewed cases (28.4 %) were classified as real: ALI definitions yielded 37 cases from 208 reviewed (17.8 %); AKI definitions yielded 85 cases from 200 reviewed (42.5 %); and MI definitions yielded 52 cases from 204 reviewed (25.5 %). The PPV, defined as the number of screened cases that were classified as real by panelists divided by the total number of cases reviewed, ranged from 0 to 52 % across ALI definitions, 12 to 82 % across AKI definitions, and 1 to 56 % across MI definitions. Generally, the laboratory-based definitions or the more restrictive definitions yielded more true positive cases. For example, with ALI, the diagnosis code only definition identified no true positive cases, while the laboratory-based measure yielded 52 % of true positive cases. For MI, where laboratory measures from the inpatient setting were sparse, the definition including both diagnostic codes and diagnostic or therapeutic procedure codes yielded more true positive cases.

Table 1 Results of case review for acute liver injury (ALI), acute kidney injury (AKI), and acute myocardial infarction (MI)

	Aggregate by HOI definition				Panel A			Panel B		
	A	B	C	D	Panelist 1	Panelist 2	Kappa	Panelist 3	Panelist 4	Kappa
Acute liver injury										
Case prevalence; n	117,550	3,741	28	191	–	–	–	–	–	–
Cases reviewed; n	60	60	28	60	104	104	–	104	104	–
True cases (PPV); n (%)	0 (0 %)	3 (5 %)	3 (11 %)	31 (52 %)	14 (14 %)	4 (4 %)	0.1	36 (35 %)	18 (17 %)	0.6**
10-Point Certainty (M ± SD)	1.7 ± 0.8	2.4 ± 1.1	2.7 ± 1.6	4.8 ± 2.2	3.2 ± 2.1	1.7 ± 1.4	0.0	3.7 ± 2.4	3.1 ± 2.3	0.2**
Acute kidney injury										
Case prevalence; n	7,474	735	5,379	7,385	–	–	–	–	–	–
Cases reviewed; n	50	50	50	50	100	100	–	100	100	–
True cases (PPV); n (%)	6 (12 %)	16 (32 %)	41 (82 %)	22 (44 %)	38 (38 %)	88 (88 %)	0.1**	36 (36 %)	36 (36 %)	0.4**
10-point certainty (M ± SD)	3.1 ± 1.5	4.2 ± 2.1	6.0 ± 1.8	4.7 ± 1.7	3.6 ± 2.4	6.4 ± 1.6	0.0	4.3 ± 2.2	4.5 ± 2.6	0.1**
Myocardial infarction										
Case prevalence; n	21,673	7,247	6,284	–	–	–	–	–	–	–
Cases reviewed; n	68	68	68	–	102	102	–	102	102	–
True cases (PPV); n (%)	1 (1 %)	13 (19 %)	38 (56 %)	–	41 (40 %)	41 (40 %)	0.6**	30 (29 %)	11 (11 %)	0.4**
10-point certainty (M ± SD)	2.8 ± 0.9	3.6 ± 2.0	5.7 ± 2.2	–	4.0 ± 2.7	5.1 ± 2.4	0.0	4.5 ± 2.3	2.9 ± 1.8	0.1**

** $P < 0.05$

Acute liver injury—definition A: $\geq$1 diagnosis code of 570* (where "*" indicates extra digits, covering any subcategory of that ICD-9 diagnosis code), 572.2, 572.4, or 573*; definition B: definition A AND diagnostic procedure code days −30 to 0, OR therapeutic procedure code days 0–60; definition C: definition B AND definition D; definition D: Hy's law: Total bilirubin $\geq$2× the upper limit of normal (ULN) AND Alanine Aminotransferase (ALT) $\geq$3× ULN OR Aspartate Aminotransferase (AST) $\geq$3× ULN

Acute kidney injury—definition A: $\geq$1 diagnosis code of 584*; definition B: definition A AND acute dialysis procedure code days 0–60 with no chronic dialysis codes before day 0; definition C: $\geq$2 Logical Observation Identifier Names and Codes (LOINC) lab codes 2160-0 with Creatinine increase $\geq$0.5 if baseline $\leq$1.9 or increase $\geq$1.0 if baseline $\leq$2–4.9, or increase $\geq$1.5 if baseline $\geq$5 mg/dL; definition D: $\geq$1 diagnosis code of 584, 584.5, 584.6, or 584.7

Myocardial infarction—definition A: $\geq$1 diagnosis code of 410*, 411.1, or 411.8, OR $\geq$1 hospital diagnosis code of 413.9; definition B: $\geq$1 diagnosis code of 410*; definition C: definition B AND diagnostic procedure code days −30 to 0, OR therapeutic procedure code days 0–60

The certainty scores on the 10-point scale paralleled the PPV, with a range of mean values from 1.7 to 4.8 across ALI definitions, 3.1 to 6.0 across AKI definitions, and 2.8 to 5.7 across MI definitions. These values generally reflect a fair degree of uncertainty in case classification.

4 Discussion

Using existing HOI operational definitions, we identified potential cases of ALI, AKI, and MI in the MSLR database and presented patient-level data to expert panelists. Based on independent, dual expert review of potential cases identified by existing HOI definitions, only 17.8 % of ALI cases, 42.5 % of AKI cases, and 25.5 % of MI cases were determined to be real. There was a fair degree of disagreement among panelists in case classification (16–39 % of cases), with relatively low Kappa statistics. The high degree of false positive cases across HOIs is concerning, highlighting the importance of improving methods for HOI measurement in observational data sources.

The diagnosis only-based definitions generally yielded the fewest true cases, while definitions including laboratory values or procedures yielded more positive cases. The inability of diagnostic codes to accurately identify events like ALI, AKI, and MI has been previously documented in past validation studies [5–7]. This suggests that future work might focus on these more restrictive definitions using data sources containing laboratory measurements. However, this comes at a cost in sample size and may not translate into improved ability to detect drug-outcome relationships. For example, OMOP experiments have illustrated a reduction in performance (lower area under the curve) for more specific HOI definitions as compared with broad, diagnosis-only based definitions [11]. Additional assessment of how tradeoffs in sample size and PPV of measurement definitions should be conducted.

Previous research by Vessey and Doll [12, 13] has illustrated that people developing an HOI often have a predisposing cause. In the context of drug safety research, those with a predisposing cause might be less likely than others to be using the drug of interest. Therefore, the cleanest HOI definition is one that represents idiopathic disease, or rather disease of unknown origin. While our expert panel review was not explicitly designed to identify idiopathic cases of each HOI, panelists were asked to give particular attention to cases that could be attributed to factors other than a drug exposure (e.g., pre-existing viral hepatitis and ALI) and consider whether the outcome definition was measuring these factors or a new, potentially drug-related, acute injury. If the HOI could be explained by other factors, it was not considered a case. Further, since we allowed the expert panelists to see drug exposure data for drugs that were indicated, used off-label, or contraindicated for the HOI, it is possible that panelists might have used knowledge of a drug exposure when classifying cases. These considerations, and the potential biases they might introduce when investigating drug-outcome associations, need to be considered in future assessments.

Our study is limited by the underlying validity and completeness of our observational claims data source and lack of standardization among panelists in terms of how they viewed the HOIs. We did not go to medical records to adjudicate the expert classification. This could lead to case misclassification, and might explain why we generally identified fewer positive cases when compared with results from prior validation studies [5–7]. Further, we only reviewed a sample of cases identified in a single data source. These data, and the coding practices used to generate these data, might differ from other healthcare databases leading to different results. This important limitation previously has been illustrated by Katz et al. [14], when comparing the prevalence of coded signs, symptoms, laboratory tests, and diagnostic procedures across possible cases of ALI in different data sources. Because the panelist review process was time consuming, we were limited in the number of cases that could be reviewed. We therefore focused only on cases identified by competing HOI definitions. Because we did not review cases that did not meet one of our existing HOI definitions, we cannot calculate sensitivity or specificity of the definitions in screening the population. Finally, the MSLR data used in our evaluation were converted from their original form to a Common Data Model [9]. While this data conversion is important to allow for efficient replication of drug safety analyses across different data sources and large populations, it is unclear what impact the data conversion might have on HOI measurement.

5 Conclusions

Through dual, independent case review with disagreements resolved by consensus discussions, expert panelists' review found the PPV ranged from 0 to 52 % across ALI definitions, 12 to 82 % across AKI definitions, and 1 to 56 % across MI definitions. The purely diagnosis-based definitions yielded the lowest proportion of cases, while the definitions incorporating laboratory measures or procedures yielded more cases. Outcome definitions need to be improved and better studied for observational data to be reliably used in drug safety surveillance and other health outcomes research.

Acknowledgments This project was supported by the Foundation for the National Institutes of Health, grant HANSE11OMOP. We thank the following expert panelists: Jon Duke, MD, MS, Regenstrief Institute; Ben Powers, MD, MHS—St Luke's Health System; Mei Sheng Duh, MPH, SCD, Analysis Group; T. Craig Cheetham, PharmD, MS, Kaiser Permanente; Scott Sanoff, MD, MPH, University of Virginia; Andrew Bomback, MD, MPH, Columbia University; Gregory Hess, MD, MBA, MSc, University of Pennsylvania; Themistocles Assimes, MD, PhD, Stanford University. A preliminary version of this work was presented as a poster at the OMOP Symposium, June 28, 2012, and at the Annual Meeting of the American Pharmacists Association, March 3, 2013. Dr. Hansen has received consulting support from Takeda Pharmaceuticals, Novartis, and Daiichi Sankyo for research unrelated to the content of this paper, and he has provided expert testimony for Allergan. Drs. Gray, Fox, Hollingsworth, and Zeng, as well as Ms. Gao declare no potential conflicts of interest.

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.

References

1. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010;153(9):600–6.
2. Observational Medical Outcomes Partnership (2010) [cited 2012 October 8]. http://omop.org/HOI.
3. Ryan PB, Madigan D. Observational Medical Outcomes Partnership (OMOP) methods evaluation. OMOP 2011 Symposium; 2011 January 11; Washington, DC.
4. Racoosin JA, Ryan PB. Implications of health outcomes of interest definitions: acute liver injury case study. OMOP 2011 Symposium; 2011 January 11; Washington, DC.
5. Jinjuvadia K, Kwan W, Fontana RJ. Searching for a needle in a haystack: use of ICD-9-CM codes in drug-induced liver injury. Am J Gastroenterol. 2007;102(11):2437–43.
6. Vlasschaert ME, Bejaimal SA, Hackam DG, et al. Validity of administrative database coding for kidney disease: a systematic review. Am J Kidney Dis. 2011;57(1):29–43.
7. Metcalfe A, Neudam A, Forde S, et al. Case definitions for acute myocardial infarction in administrative databases and their impact on in-hospital mortality rates. Health Serv Res. 2013;48(1):290–318.
8. Fox BI, Hollingsworth JC, Gray MD, et al. Developing an expert panel process to refine health outcome definitions in observational data. J Biomed Inform 2013;46(5):795–804.
9. Overhage JM, Ryan PB, Reich CG, et al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.

10. Observational Medical Outcomes Partnership. OMOP Collaborator—Thonmson Reuters. [March 7, 2012]. http://omop.org/CDMvocabV4.
11. Ryan PB. Lessons for building a risk identification and analysis system. Observational Medical Outcomes Partnership Symposium, June 28, 2012. Bethesda, MD. http://omop.org/2012OMOPmeeting.
12. Vessey MP, Doll R. Investigation of relation between use of oral contraceptives and thromboembolic disease. Br Med J. 1968; 2(5599):199–205.
13. Vessey MP. Learning how to control biases in studies to identify adverse effects of drugs. JLL Bulletin: Commentaries on the history of treatment evaluation. 2006. Accessed 26 June 2013. http://www.jameslindlibrary.org/illustrating/articles/learning-how-to-control-biases-in-studies-to-identify-adverse-ef–2.
14. Katz AJ, Ryan PB, Racoosin JA, et al. Assessment of case definitions for identifying acute liver injury in large observational databases. Drug Saf 2013;36(8):651–61.