© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2010.01159.x
METHODS BRIEF

Derivation and Validation of a MEDLINE Search Strategy for Research Studies That Use Administrative Data

Carl van Walraven, Carol Bennett, and Alan J. Forster

Objective. To derive and validate a search strategy that identifies administrative database research (ADR) in the MEDLINE database.

Design. Analytical survey.

Methods. We downloaded all articles published between January 1, 2008 and October 7, 2009 in 20 top journals in internal medicine, cardiovascular medicine, public health, and health services research. These were reviewed to determine whether they were ADR (in which the study cohort, exposure, or outcome was defined using electronic data created for or during the processing of patients through their health care). We used chi-squared recursive partitioning to create a search strategy that maximized sensitivity based on publication type, MeSH headings, and text words.

Main Outcome Measures. Sensitivity and positive predictive value of the search strategy for true ADR in three samples: derivation ($n = 5,513$); internal validation ($n = 2,710$); and external validation ($n = 1,500$).

Results. The prevalence of ADR in the derivation, internal validation, and external validation samples was 2.6, 2.9, and 2.2 percent, respectively. The sensitivity of our search strategy in these samples was 90.9 percent (95 percent confidence interval [CI] 85.0–95.1), 88.5 percent (79.2–94.6), and 100 percent (99.3–100), respectively. The positive predictive value in these samples was 10.7 percent (9.0–12.6), 11.5 percent (9.1–14.4), and 3.3 percent (2.3–4.6), respectively.

Conclusion. We derived and validated a search strategy that is highly sensitive for ADR in MEDLINE.

Key Words. Administrative data uses, biostatistical methods, information technology in health

Administrative data are created as a by-product of providing patient care. These data can be used to answer research questions (i.e., “administrative database research”). Administrative database research (ADR) has several

advantages: it is relatively inexpensive (since the data have already been collected); it describes “real-world” utilization and outcomes of medical technologies; and it frequently describes health care in an entire population.

However, there is no standardized search strategy to identify ADR in electronic bibliographic databases such as MEDLINE. Without such a methodological filter, searching for ADR becomes much less efficient because one is forced to conduct a subject-specific search, which must be then manually reviewed to identify the ADR. In addition, ADR researchers may encounter subject nonspecific issues that require a search of the literature (e.g., methods of controlling for comorbidity using administrative databases). In such cases, subject-specific searches will be unhelpful.

A search strategy that is specific for ADR would make its identification much more efficient and would benefit several groups. Researchers need to identify published ADR to find research methodologies and previous ADR studies of a particular subject. Administrators need to identify ADR as comparators for processes or outcomes in their institution. Clinicians need to identify ADR to determine risk of expected outcomes for diseases and processes.

In this study, we derived and validated a search strategy to identify ADR.

METHODS

We derived and validated a search strategy for the MEDLINE electronic database because it is the most commonly used and comprehensive biomedical bibliographic database. We identified subject categories (as grouped in the Web of Science Citation Index) that we thought were most likely to publish ADR. These categories included Health Care Sciences & Services; Medicine, General & Internal; Public, Environmental, & Occupational Health; and Cardiac & Cardiovascular Systems. We identified the top five journals within each category based on their most recent journal impact factor. We downloaded all citations from these 20 journals that were published between January 1, 2008 and October 7, 2009 ($n = 17,667$, Figure 1). After excluding

Address correspondence to Carl van Walraven, M.D., M.Sc., F.R.C.P.C., Ottawa Hospital Research Institute, Institute for Clinical Evaluative Sciences, ASB 1-003 1053 Carling Avenue, Ottawa, ON K1Y 4E9; e-mail: carlv@ohri.ca. Carol Bennett, M.Sc., is with the Ottawa Hospital Research Institute, Ottawa, ON. Alan J. Forster, M.D., M.Sc., F.R.C.P.C., is with the Ottawa Hospital Research Institute, Institute for Clinical Evaluative Sciences, Ottawa, ON.

Figure 1: Study Sample

Note. Journals included in our initial sample were as follows: *Annals of Internal Medicine*; *Annual Review of Public Health*; *British Medical Journal*; *Circulation*; *Circulation Research*; *Environmental Health Perspectives*; *Epidemiologic Reviews*; *European Heart Journal*; *Health & Quality of Life Outcomes*; *Health Affairs*; *Health Technology Assessment Winchester* (England); *International Journal of Epidemiology*; *Journal of American Medical Association*; *Journal of Medical Internet Research*; *Journal of the American College of Cardiology*; *Lancet*; *Medical Care*; *Nature Clinical Practice Cardiovascular Medicine*; *New England Journal of Medicine*; and *World Health Organization Technical Report Series*.

**Excludes publication type: biography; case reports; comment; editorial; historical article; interactive tutorial; interview; letter; news; newspaper article; patient education handout; portrait.

publication types that were incompatible with ADR ($n = 9,444$), these citations ($n = 8,223$) were reviewed by either of two of us (C. v. W., C. B.) to identify those that suggested ADR (“screen-positive” citations, $n = 396$). The reviewers started out by assessing the same 300 citations and then compared results to harmonize screening methods.

The full text articles of “screen-positive” citations ($n = 396$) were then retrieved and reviewed. To be classified as ADR, studies had to define the study cohort, exposure, or outcome using electronic data that were created for or during the “processing” of patients through their health care (i.e., administrative data). Examples include datasets for health plan enrollment, service claims, hospital discharge abstracts, and medication claims.

We excluded studies that exclusively used (i.e., without *any* true administrative data) diagnosis-specific registries; birth or death registries; or electronic medical records (EMRs). For our study, we defined EMR as databases used at the point of patient care. However, we included EMR-based studies if the EMR datasets used for the research were those that are commonly available in typical administrative data repositories (e.g., medication prescriptions). Finally, we excluded studies in which administrative databases were used to identify a potential cohort that was validated with nonadministrative data (e.g., chart review).

We wanted to identify a search strategy that maximized sensitivity (i.e., probability that true ADR is identified by the search strategy) for ADR. This would allow searchers to completely identify most ADR in a particular search, the results of which could be manually reviewed to identify those citations that truly are ADR. We derived the search strategy using three groups of parameters: publication type (e.g., “Consensus Conference”); MeSH headings present in more than 2 percent of ADR studies; and text words that would strongly suggest ADR. We first identified all publication types (e.g., “Comment,” “In Vitro”) that would exclude ADR. We then identified all publication types, MeSH headings, and text words that were significantly associated with ADR status at a p -value $< .05$ using a chi-squared test (Appendix SA2).

These variables were then fed into a chi-squared recursive partitioning macro in *SAS* (%treedisc) to generate a search strategy that maximized sensitivity (Kass 1980). Recursive partitioning uses multiple variables to identify groups that are very similar based on the outcome chosen for the analysis. This was done in a derivation data set that was a 66 percent random sample from the initial sample ($n = 5,513$). The model was then tested on the internal validation group (i.e., the remaining 33 percent, $n = 2,710$). The final search strategy was that whose sensitivity in the validation group was within the 95 percent confidence interval (CI) of sensitivity in the derivation group. If this criterion was not met, we used the macro to generate another strategy using a more stringent α -error (i.e., the probability that one incorrectly concludes that two groups differ significantly when no such difference truly exists) to “grow” the tree. Decreasing the α -error decreases overfitting and increases external

validity of recursive partitioning models from the %treedisc macro by requiring a greater difference between groups (Khoshgoftaar and Allen 2001). Our final model used an α -error of 0.000001 to generate all splits.

We measured the accuracy of our search strategy using sensitivity, specificity (proportion of non-ADR not identified by the search), positive predictive value (proportion of citations identified by search that truly are ADR), and negative predictive value (proportion of citations not identified by search that are not ADR) (Table 1). We externally validated our search strategy with a stratified random sample of 1,500 English citations from MEDLINE published in 2006 that were ($n = 1,000$) and were not ($n = 500$) identified by our search strategy. The titles and abstracts were reviewed to exclude all non-ADR. Full text articles were retrieved for all other articles citations and the criteria above were applied to determined true ADR status.

Table 1: Description of All Citations in Our Initial Sample by Administrative Database Research Status

	<i>Administrative Database Research</i> (<i>N = 221, 2.7%</i>)	<i>Other Research Types</i> (<i>N = 8,002, 97.3%</i>)	<i>Overall</i> (<i>N = 8,223</i>)
Notable study types			
Comparative study	47 (21.3%)	769 (9.6%)	816 (9.9%)
Randomized controlled trial	1 (0.5%)	889 (11.1%)	890 (10.8%)
Notable MeSH headings			
Humans	218 (98.6%)	6,802 (85.0%)	7,020 (85.4%)
Cohort studies	51 (23.1%)	320 (4.0%)	371 (4.5%)
Retrospective studies	37 (16.7%)	212 (2.6%)	249 (3.0%)
Hospitalization/sn	29 (13.1%)	54 (0.7%)	83 (1.0%)
Medicare	19 (8.6%)	24 (0.3%)	43 (0.5%)
Health services	11 (5.0%)	21 (0.3%)	32 (0.4%)
accessibility/sn			
Health services/ut	9 (4.1%)	17 (0.2%)	26 (0.3%)
Medicare/sn	18 (8.1%)	5 (0.1%)	23 (0.3%)
Words present in title or abstract			
Claims	8 (3.6%)	4 (0.0%)	12 (0.1%)
Medicare	23 (10.4%)	49 (0.6%)	72 (0.9%)
Population-based or population based	19 (8.6%)	40 (0.5%)	59 (0.7%)
Long-term or long term	10 (4.5%)	124 (1.5%)	134 (1.6%)
Administrative	4 (1.8%)	1 (0.0%)	5 (0.1%)

sn, statistics and numerical data; ut, utilization.

RESULTS

Figure 1 presents how our study sample was generated. We identified 396 citations that were compatible with ADR; 221 of these were true ADR for an ADR prevalence of 2.7 percent in our initial sample.

Table 1 presents a description of our initial sample ($n = 8,223$). ADR were three times more likely to be classified as a comparative study and much less likely to be classified as a randomized controlled trial. The MeSH headings listed in Table 1—including “Cohort Studies” and “Retrospective Studies”—were notably more common in ADR. Several text words, including “claims,” “Medicare,” and “population-based,” were significantly more prevalent in ADR.

Table 2 presents our final search strategy (available, published as an Appendix SA2). Each row in our strategy describes one “limb” of the regression tree created in the analysis. The search strategy was sensitive for ADR in both the derivation (sensitivity 90.9 percent, 95 percent CI 85.0–95.1) and the internal validation groups (sensitivity 88.5 percent, 95 percent CI 79.2–94.6) (Table 3). The combination of low ADR prevalence and our search’s specificity of just 80 percent meant that approximately only 1 in 10 citations identified by our search strategy in the initial sample were true ADR (Table 3).

The validity of our search strategy was confirmed in the external sample (Table 3). The sensitivity in this sample was 100 percent (95 percent CI 89.4–100). However, the positive predictive value dropped to 3.4 percent (2.4–4.8) due to the lower prevalence of ADR in the external sample (2.3 percent, 95 percent CI 1.6–3.2).

DISCUSSION

In this study we derived and both internally and externally validated a search strategy to identify ADR in the MEDLINE database. We believe that researchers, administrators, and clinicians can confidently use this search strategy to identify ADR in MEDLINE.

ADR is a research methodology that can be applied to a wide variety of areas in research. Unfortunately, ADR does not have its own identifying heading in MEDLINE or a standard terminology to ease its identification in bibliographic databases. Our search strategy will allow researchers, administrators, and physicians to more easily identify ADR. Crossing this search strategy (which is methodologically based) with other search strategies that are

Table 2: Search Strategy to Identify Administrative Database Research

<i>Search</i>	<i>Search Term</i>
1.	Medicare/sn
2.	Hospitalization/sn
3.	2 not 1
4.	Medicare/
5.	4 not (1 or 2)
6.	(population adj based).tw.
7.	6 not (1 or 2 or 4)
8.	administrative.tw.
9.	8 not (1 or 2 or 4 or 6)
10.	Health Services/ut
11.	10 not (1 or 2 or 4 or 6 or 8)
12.	Hospitalization\$.tw.
13.	12 not (1 or 2 or 4 or 6 or 8 or 10)
14.	California/
15.	14 not (1 or 2 or 4 or 6 or 8 or 10 or 12)
16.	Health Services Accessibility/sn
17.	16 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14)
18.	Hospital Mortality/
19.	18 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16)
20.	Registries/
21.	20 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18)
22.	Cohort Studies/
23.	22 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20)
24.	Longitudinal Studies/
25.	United States/
26.	24 and 25
27.	26 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22)
28.	25 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 24)
29.	Comorbidity/
30.	29 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25)
31.	Health Care Costs/
32.	31 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29)
33.	Health Care Surveys/
34.	33 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29 or 31)
35.	"Outcome Assessment (Health Care)"/
36.	35 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29 or 31 or 33)
37.	"Quality of Health Care"/
38.	37 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29 or 31 or 33 or 35)
39.	"Costs and Cost Analysis"/
40.	39 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29 or 31 or 33 or 35 or 37)
41.	"Length of Stay"/
42.	41 not (1 or 2 or 4 or 6 or 8 or 10 or 12 or 14 or 16 or 18 or 20 or 22 or 25 or 29 or 31 or 33 or 35 or 37 or 39)

continued

Table 2. *Continued*

<i>Search</i>	<i>Search Term</i>
43.	1 or 3 or 5 or 7 or 9 or 11 or 13 or 15 or 17 or 19 or 21 or 23 or 27 or 28 or 30 or 32 or 34 or 36 or 38 or 40 or 42
44.	Limit 43 to (addresses or bibliography or classical article or clinical conference or clinical trial, all or clinical trial, phase i or clinical trial, phase ii or clinical trial, phase iii or clinical trial, phase iv or clinical trial or comparative study or congresses or consensus development conference or consensus development conference, nih or controlled clinical trial or "corrected and republished article" or dictionary or directory or duplicate publication or english abstract or evaluation studies or festschrift or government publications or guideline or in vitro or introductory journal article or journal article or lectures or legal cases or legislation or meta analysis or multicenter study or overall or periodical index or practice guideline or published erratum or randomized controlled trial or research support, nih, extramural or research support, nih, intramural or research support, non us gov't or research support, us gov't, non phs or research support, us gov't, phs or retracted publication or "retraction of publication" or "review" or "scientific integrity review" or technical report or twin study or validation studies)

sn, statistics and numerical data; ut, utilization.

subject specific based on diagnosis or other methodologies will let searchers “subsample” their searches to the ADR they want to identify.

One unique characteristic of our study was its use of recursive partitioning to create the search strategy. This methodology allowed us to identify previously unseen and complex patterns in the data to create our search strategy. This method has two potential limitations: first, it is a data-driven regression methodology and can generate models that are “overfitted” and have limited external validity. This is the reason we validated our search strategy in two separate samples. Second, the resulting search strategy has no apparent face validity to it. However, our validation analyses show that our search strategy does work, regardless of its lack of face validity.

Although not perfect, the sensitivity of our search strategy in the initial sample is in the same range as those in previous studies deriving and assessing search strategies for different study types (Wilczynski, Haynes, and Hedges Team 2004; Haynes et al. 2005). However, the positive predictive value in our initial sample was approximately 10 percent. This means that people who use this search strategy will have to carefully review citations identified by this search strategy to exclude non-ADR studies.

Our study has other limitations. The strategy applies to MEDLINE only. Wilczynski and Haynes (2005) derived a search strategy to identify prognostic studies in EMBASE, which was notably different from that used in

Table 3: Operating Characteristics of Search Strategy for Administrative Database Research (ADR) in the Derivation, Internal Validation, and External Validation Samples

	<i>ADR</i>	
	<i>Yes</i>	<i>No</i>
<i>Derivation</i>		
Identified by search		
Yes	130	1087
No	13	4283
ADR prevalence	2.6%	(2.2–3.1)
Sensitivity	90.9%	(85.0–95.1)
Specificity	79.8%	(78.7–80.8)
Positive predictive value	10.7%	(9.0–12.6)
Negative predictive value	99.7%	(99.5–99.8)
<i>Internal validation</i>		
Identified by search		
Yes	69	530
No	9	2102
ADR prevalence	2.9%	(2.3–3.6)
Sensitivity	88.5%	(79.2–94.6)
Specificity	79.9%	(78.3–81.4)
Positive predictive value	11.5%	(9.1–14.4)
Negative predictive value	99.6%	(99.2–99.8)
<i>External validation</i>		
Identified by search		
Yes	34	966
No	0	500
ADR prevalence	2.3%	(1.6–3.2)
Sensitivity	100.0%	(87.4–100)
Specificity	34.1%	(31.7–36.6)
Positive predictive value	3.4%	(2.4–4.8)
Negative predictive value	100.0%	(99.0–100)

MEDLINE. Further work is required to derive and validate a search strategy for ADR that can be used in databases other than MEDLINE. In addition, we do not know the search strategy’s accuracy in years before 2006. Finally, we derived a definition for ADR for this study because it has no standard definition. Our search strategy applies only to studies contained in our definition of ADR. Those with a more expansive definition of ADR will need to keep this in mind when using our search strategy.

In summary, we derived and validated a search strategy that is highly sensitive for ADR in MEDLINE.

ACKNOWLEDGMENTS

Disclosures: None.

Disclaimers: None.

REFERENCES

- Haynes, R. B., K. A. McKibbin, N. L. Wilczynski, S. D. Walter, S. R. WerreHedges Team. 2005. "Optimal Search Strategies for Retrieving Scientifically Strong Studies of Treatment from MEDLINE: Analytical Survey." *British Medical Journal* 330: 1179–85.
- Kass, G. V. 1980. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics* 29: 119–27.
- Khoshgoftaar, A. M., and E. B. Allen. 2001. "Controlling Overfitting in Classification-Tree Models of Software Quality." *Empirical Software Engineering* 6: 59–79.
- Wilczynski, N. L., and R. B. Haynes. 2005. "Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in EMBASE: An Analytic Survey." *Journal of the American Medical Informatics Association* 12: 481–5.
- . Hedges Team. 2004. "Developing Optimal Search Strategies for Detecting Clinically Sound Prognostic Studies in MEDLINE: An Analytic Survey." *British Medical Center for Medicine* 2: 23–8.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2. List of the Publication Types, MeSH Headings, and Text Words That Were Examined for Derivation of Search Strategy. (Those with a prevalence that differs significantly between ADR and non-ADR papers are boldfaced and were offered to the recursive partitioning model for inclusion in the search strategy.)

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.