OPEN

Ascertainment of Aspirin Exposure Using Structured and Unstructured Large-scale Electronic Health Record Data

Ranier Bustamante, MS, Ashley Earles, MPH,* James D. Murphy, MD,†‡ Alex K. Bryant, BS,† Olga V. Patterson, PhD,§‖ Andrew J. Gawron, MD,§‖ Tonya Kaltenbach, MD,¶# Mary A. Whooley, MD,¶# Deborah A. Fisher, MD,**†† Sameer D. Saini, MD,‡‡§§ Samir Gupta, MD, MSCS,*†‡ and Lin Liu, PhD*†‡*

Background: Aspirin impacts risk for important outcomes such as cancer, cardiovascular disease, and gastrointestinal bleeding. However, ascertaining exposure to medications available both by prescription and over-the-counter such as aspirin for research and quality improvement purposes is a challenge.

Objectives: Develop and validate a strategy for ascertaining aspirin exposure, utilizing a combination of structured and unstructured data.

Research Design: This is a retrospective cohort study.

Subjects: In total, 1,869,439 Veterans who underwent usual care colonoscopy 1999–2014 within the Department of Veterans Affairs.

Measures: Aspirin exposure and dose were obtained from an ascertainment strategy combining query of structured medication records available in electronic health record databases and unstructured data extracted from free-text progress notes. Prevalence of any aspirin exposure and dose-specific exposure were estimated. Positive predictive value and negative predictive value were used to assess strategy performance, using manual chart review as the reference standard.

Results: Our combined strategy for ascertaining aspirin exposure using structured and unstructured data reached a positive predictive value and negative predictive value of 99.2% and 97.5% for any exposure, and 92.6% and 98.3% for dose-specific exposure. Estimated prevalence of any aspirin exposure was 36.3% (95% confidence interval: 36.2%–36.4%) and dose-specific exposure was 35.4% (95% confidence interval: 35.3%–35.5%).

Conclusions: A readily accessible approach utilizing a combination of structured medication records and query of unstructured data can be used to ascertain aspirin exposure when manual chart review is impractical.

Key Words: aspirin, electronic health records, ascertainment

(*Med Care* 2019;57: e60–e64)

From the *VA San Diego Healthcare System; †University of California, San Diego; ‡Moores Cancer Center, La Jolla, CA; §VA Salt Lake City Health Care System; ‖University of Utah, Salt Lake City, UT; ¶San Francisco VA Medical Center; #University of California, San Francisco, San Francisco, CA; **Durham VA Medical Center; ††Duke University, Durham, NC; ‡‡VA Ann Arbor Healthcare System; and §§University of Michigan, Ann Arbor, MI.

Supported by Merit Review 5 I01 HX001574-03 (S.G., PI) and Quality Enhancement Research Initiative 5 IP1 HX002002-03 (T.K., Project 1 PI) from the US Department of Veterans Affairs Health Services Research & Development Service of the VA Office of Research and Development. Research reported in this publication was also supported by the National Cancer Institute of the National Institutes of Health under Award Number 1 R37 CA222866-01 (S.G., PI).

The content is solely the responsibility of the authors and views expressed in this paper are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the National Institutes of Health.

The authors declare no conflict of interest.

Reprints: Samir Gupta, MD, MSCS, UC San Diego, 3350 La Jolla Village Drive MC 111D, San Diego, CA 92161. E-mail: s1gupta@ucsd.edu.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/19/5710-e060

Aspirin is the most commonly documented medication exposure in outpatient office visits across the United States.[1] Ascertainment of aspirin exposure is of great importance for research and quality improvement initiatives due to its potential benefits for primary and secondary prevention of cardiovascular disease and cancer as well as potential risks for side effects such as gastrointestinal bleeding and hemorrhagic stroke.[2–5] For example, questions remain regarding which individuals may benefit most from aspirin for primary prevention of colorectal cancer.[6–11] In addition, monitoring the proportion of patients presenting with acute myocardial infarction who received aspirin within 24 hours of arrival is recommended as a quality metric by the National Quality Forum.[12]

Increasing availability of large-scale electronic health records (EHRs) offers the opportunity to better understand the potential risks and benefits of aspirin, and to monitor adherence to evidence-based guidelines for use. However, because aspirin is rarely prescribed and often recommended to be purchased over-the-counter, ascertainment of aspirin exposure presents a challenge.[13–16] Although prescriptions are often not made, clinical experience suggests physicians often monitor and record aspirin exposure within free-text clinical notes, as well as within structured medication records (also known as "medication lists"). Thus, strategies that include

extraction of documented aspirin use from free-text progress notes created as part of usual health care might allow for optimal ascertainment of exposure. Indeed, others have shown that sophisticated natural language processing (NLP)-based algorithms can effectively extract aspirin exposure from EHRs with reported sensitivity and specificity ranging from 95.5% to 99.0% and 91.0% to 98.9%, respectively.[4,17] However, applying NLP as part of usual epidemiologic research or quality metric assessment can be a major challenge, requiring specialized software, expertise in NLP, and substantial collaborative work. Furthermore, NLP algorithms optimized and shown effective in one clinical setting often need to be recalibrated for other settings, challenging generalizability.[18] Structured query language (SQL) search functions can be used to identify key words associated with exposure within free-text progress notes. SQL applications are widely available in most database management programs supporting EHRs, are much easier to use, and may serve as an alternative method for aspirin ascertainment from unstructured data such as free-text progress notes. Furthermore, novel strategies that leverage both unstructured data within EHRs and structured medication records may optimize ascertainment. Herein, we report results of a simple, accurate, and readily accessible method that utilizes a combination of structured medication records, as well as query of free-text progress notes for characterizing aspirin exposure that can be used for epidemiologic research and quality improvement initiatives.

METHODS

Overview

Our overarching aim was to develop and validate a strategy for identifying aspirin exposure within our use case of Veterans who underwent usual care colonoscopy within the Department of Veterans Affairs (VA). The cohort was created as part of a VA Office of Research and Development-funded study to examine predictors of colorectal cancer and high-risk polyps after baseline adenoma removal.[19] As exposure may impact colorectal cancer risk, we sought to identify a rigorous method for aspirin ascertainment. We hypothesized that an ascertainment strategy utilizing a combination of structured medication records, as well as searching for documentation of aspirin exposure within free-text progress notes would optimize identification of individuals with evidence of aspirin use at time of baseline colonoscopy. The specific aims were to characterize aspirin exposure and dose-specific exposure within the year before baseline colonoscopy for large datasets without having to rely on manual chart review.

Study Setting

The VA is one of the largest health care system in the United States, serving over 9 million Veterans annually.[20] The VA's EHR contains both structured (eg, demographics, medication records, laboratory results) and unstructured data (eg, free-text clinical progress notes, procedure notes, pathology reports) that are available for research and quality improvement initiatives through the VA Informatics and Computing Infrastructure.

Data Sources

The primary data source for this work was the VA's Corporate Data Warehouse (CDW). The CDW contains administrative billing codes, such as Current Procedural Terminology (CPT) and International Classification of Diseases, Ninth Revision (ICD9) diagnosis codes, as well as structured medication records and free-text progress notes. The cohort for this study consisted of Veterans with a CPT code for colonoscopy 1999–2014, at least 1 corresponding clinical note within 30 days of index colonoscopy, and no history of inflammatory bowel disease (see Supplementary Digital Content 1, http://links.lww.com/MLR/B701 for full list of inclusion/exclusion criteria). Date of index colonoscopy served as the anchor date for the study. We identified all structured medication records and all free-text progress notes up to 1 year before their index colonoscopy. Aspirin exposure was defined as at least 2 instances of documented aspirin exposure within free-text progress notes and/or aspirin medication fills in the year before index colonoscopy.

Aspirin Ascertainment Strategies

We developed 3 strategies for ascertaining aspirin exposure using structured and unstructured data (Table 1). The structured approach utilized several domains within the CDW. Structured data on aspirin prescriptions were stored in the VA Meds domain, a repository for pharmacy prescription data, and the Non-VA Meds domain, which captures provider-documented exposure to medications not provided by the VA. Documentation of aspirin within these domains were counted as evidence of exposure. In addition, data on aspirin allergies were stored in the allergy domain and counted as evidence of nonexposure. Use of only structured data resulted in a lower prevalence than what would be expected, which led to the inclusion of unstructured data.

For the unstructured approach, we searched a random sample of progress notes for the terms "aspirin" and "ASA" and manually reviewed these notes to identify common ways that aspirin exposure was recorded within free-text notes. We identified pairs of terms commonly associated with aspirin exposure and examined character distance between mention of aspirin and dose indicators.[21] Distance was of critical importance as SQL requires separate specification of "81mg" and "81 mg" as distinct search terms. We established SQL searches for terms consistent

TABLE 1. Overview of Structured and Unstructured Strategies for Aspirin Exposure

Structured Data		Unstructured Data		
Exposure	**Nonexposure**	**Exposure**		**Nonexposure**
VA Meds	Allergies	81	325 mg	Not
Non-VA Meds	—	81 mg	Baby	Avoid
		162	Resume	Stop
		162 mg	Yes	Allergies
		325	—	—

Med indicates medication; VA, Veterans Affairs.

with both aspirin exposure and nonexposure (Supplementary Digital Content 2, http://links.lww.com/MLR/B702). Common terms associated with aspirin exposure included "81," "162," "325," and "baby." Common terms associated with nonexposure included "avoid" and "stop."

The third strategy for ascertaining aspirin exposure included a combination of structured medication records and free-text progress notes (see Supplementary Digital Content 3, http://links.lww.com/MLR/B703 for full SQL query). To keep our structured medication definition simple and consistent, we elected to use at least 2 prescription fills in the year before index colonoscopy to ensure at least 1 refill of prescribed aspirin was filled to support evidence of aspirin exposure. For each strategy, prevalence (with 95% confidence intervals) for any aspirin exposure and dose-specific exposure were obtained. Notably, aspirin medication records pulled from structured data were always associated with documented dosage. Exposure was summarized in 2 ways: (1) a binary variable as exposed/unexposed, regardless of whether dose was extracted; and (2) a binary variable as exposed/unexposed but only if a dose was extracted.

Performance Characteristics

Performance of each strategy was assessed using positive predictive value (PPV) and negative predictive value (NPV). PPV was defined as the probability of selecting a subject who actually used aspirin among those identified as having aspirin exposure. NPV was defined as the probability of selecting a subject who did not use aspirin among those identified as not having aspirin exposure. The reference standard for performance assessment was a combination of manual chart review of free-text progress notes and pharmacy data from time of baseline colonoscopy up to 1 year before the procedure. The algorithm-based ascertainment of aspirin exposure was compared with exposure as documented in the medical chart and summarized using concordance tables. The κ statistic was calculated for aspirin exposure and dosage using combined free-text and pharmacy data. For the analysis of dose-specific exposure, PPV and NPV were specific to correct identification of presence of aspirin with the correct dose, if documented in medical records. The 95% one-sided confidence lower bounds for PPV and NPV were calculated and Bonferroni correction was used for multiple comparison adjustment to ensure an overall confidence of 95%.

To have sufficient power to judge performance, we performed an analysis of the sample sizes required for chart review based on sample estimates of PPV and NPV and their 95% one-sided confidence lower bounds (see Supplementary Digital Content 4, http://links.lww.com/MLR/B704, for sample size calculation). In our validation, we chose to select an independent random sample of at least 100 candidate cases with aspirin exposure and at least 100 candidate controls with nonexposure for chart review. During the implementation, we oversampled by 10%–20% to account for invalid charts so that we could have at least 100 candidate cases and 100 candidate controls. The sample size calculation showed that if sample PPV and NPV reach 95%, which we postulated was achievable from our preparatory work, we could confidently claim that the population PPVs and NPVs for our algorithm were both likely to be $>90\%$ with high confidence. As such, we expected that obtaining a lower bound of 90% would be sufficient for accurately describing aspirin exposure.

RESULTS

We identified 1,869,439 Veterans who underwent usual care colonoscopy 1999–2014. Table 2 summarizes performance characteristics of all 3 strategies for ascertainment of aspirin exposure within this cohort compared with manual chart review. Estimated prevalence of dose-specific exposure utilizing only structured data were 25.0% [confidence interval (CI): 24.9%–25.1%]. PPV and NPV for ascertaining dose-specific exposure reached 94.4% and 79.6%, respectively. Because the confidence lower bounds for both PPV and NPV were below 90%, we could not claim that the population-level PPVs and NPVs for our algorithm were above 90% utilizing structured data alone with high confidence.

For the unstructured approach, estimated prevalence of aspirin exposure was 31.0% (CI: 30.9%–31.1%). PPV and NPV for ascertaining aspirin exposure reached 95.1% and 97.6%, respectively. Estimated prevalence for dose-specific exposure was 29.0% (CI: 28.9%–29.1%), and PPV and NPV reached 96.0% and 98.0%, respectively. After assessing the lower bounds, we could claim that the population-level PPVs and NPVs for our algorithm were above 90% for dose, but not for exposure, with high confidence.

For the combined approach utilizing structured and unstructured data, estimated prevalence of aspirin exposure was

TABLE 2. Performance Characteristics of Aspirin Ascertainment Strategies

Ascertainment Strategy	Prevalence (95% CI)	PPV (Lower Bound*)	NPV (Lower Bound*)	Sensitivity	Specificity
Structured approach†					
Aspirin exposure	—	—	—	—	—
Aspirin dose	0.25 (0.25–0.25)	0.94 (0.88)	0.80 (0.71)	0.61	0.98
Unstructured approach					
Aspirin exposure	0.31 (0.31–0.31)	0.95 (0.89)	0.98 (0.93)	0.95	0.98
Aspirin dose	0.29 (0.29–0.29)	0.96 (0.90)	0.98 (0.93)	0.95	0.98
Combined strategy					
Aspirin exposure	0.36 (0.36–0.36)	0.99 (0.95)	0.98 (0.93)	0.96	1.00
Aspirin dose	0.35 (0.35–0.36)	0.93 (0.86)	0.98 (0.94)	0.97	0.96

*95% one-sided confidence lower bound for PPV and NPV was based on binomial exact test with Bonferroni correction.
†Aspirin medication records pulled from structured data always had a dosage, therefore structured dosage and exposure are equivalent.
CI indicates confidence interval; NPV, negative predictive value; PPV, positive predictive value.

36.3% (CI: 36.2%–36.4%). PPV and NPV for ascertaining aspirin exposure reached 99.2% and 97.5%, respectively. Estimated prevalence of dose-specific exposure was 35.4% (CI: 35.3%–35.5%), and PPV and NPV for dose-specific exposure reached 92.6% and 98.3%, respectively. After assessing the lower bounds, we could claim that the population-level PPVs and NPVs for our algorithm were above 90% for exposure, but not for dose, with high confidence. A summary of concordance tables characterizing each algorithm's performance (CDW pharmacy data only, free-text data only, and combined pharmacy and free-text data) and κ statistic for aspirin exposure and dosage using combined free-text and pharmacy data are provided in Supplementary Digital Content 5 (http://links.lww.com/MLR/B705) and Supplementary Digital Content 6 (http://links.lww.com/MLR/B706), respectively.

DISCUSSION

Ascertaining aspirin exposure using routinely recorded health care data for research and quality improvement purposes is a challenge, mainly because aspirin is often not formally prescribed or documented within structured medication records. We demonstrate that a combined strategy utilizing structured and unstructured data can successfully identify patients with aspirin exposure with high PPV (99.2%) and NPV (97.5%) when applied to national VA health care data. Performance of the combined strategy for dose-specific exposure was also very good, though the proportion of patients with dose-specific exposure was slightly lower than the estimate for any aspirin exposure, and the lower bound for PPV was just below 90%.

Our findings have several implications. First, we provide a strategy for ascertaining aspirin exposure from EHRs ripe for testing within other health care systems. We postulate the approach is likely highly generalizable because most EHRs contain both structured medication records and SQL search applications that could be used to query free-text progress notes.[17,22–25] For researchers working within the VA, we provide a readily implementable strategy for ascertaining aspirin exposure for any research or quality improvement initiative. Our approach could be easily modified to ascertain aspirin exposure for any other cohort entry point such as another procedure (eg, cardiac catheterization, coronary artery bypass surgery) or diagnosis date of interest (eg, gastrointestinal bleed, myocardial infarction).

Beyond aspirin ascertainment, our approach may also be applied to other exposures typically documented in free-text progress notes, or a combination of free-text progress notes and structured medication records. For example, our approach could be directly applied to ascertain exposure to nonsteroidal anti-inflammatory drugs or other medications that are available both over-the-counter and by prescription and/or documented within medication records (eg, proton-pump inhibitors). We speculate that exposures most amenable to a SQL strategy are medications that have a limited range of stereotypical ways of being documented. Text with higher variability, or where negation must also be carefully applied, may find greater benefit from formal NLP-based algorithms for extraction.

Our work complements prior work that has shown that NLP algorithms can be used to ascertain aspirin exposure from free-text progress notes,[4,17] but may have the advantage of being more easily implemented because many health systems may have programmers familiar with SQL, whereas NLP requires more advanced training. Our work also expands prior work implementing SQL-based search of free-text notes for exposure and outcome ascertainment.[26,27] We utilize not only inherent features to the SQL language, but supplement our methods with structured data to create a stronger algorithm.

Several limitations should be considered when interpreting our work. First, our reference standard for aspirin exposure was documentation within free-text progress notes or structured medication records. Some patients may underreport aspirin exposure and some physicians may underdocument exposure in free-text notes. Errors may have also occurred in documentation of aspirin exposure, as well as associated dose. Concerns regarding underascertainment of exposure based on the reference standard used could be addressed through another study that also includes self-report or survey data on aspirin exposure collected from patients. However, this was not feasible for the current study.

Second, we implemented our approach using data extracted from the VA CDW. While we postulate the approach is generalizable elsewhere, additional validation in other health care settings will need to be performed to confirm utility of our proposed approach outside VA data. In addition, the identifiable proportion of our cohort with dose-specific exposure was slightly lower than the proportion with any exposure, despite good performance characteristics of the algorithm. This may be the case because usual care notes do not consistently record dose, but may create a challenge for analyses where dose is of critical importance.

These limitations may be considered in light of several strengths. We utilized a large-scale EHR dataset covering over 144 VA sites across the United States over a long time period, 1999–2014. Furthermore, SQL language search functions are ubiquitous and available in most database management programs supporting EHRs, emphasizing potential generalizability. Finally, we used a methodologically rigorous approach for assessing performance characteristics of each strategy.

In conclusion, we developed and validated a highly accurate approach for characterizing aspirin exposure utilizing a combination of structured medication records and unstructured usual health care data, including documentation within free-text progress notes. This approach can be used for research and quality improvement initiatives that require documentation of aspirin exposure within the VA, and merits consideration for application to other health care systems.

REFERENCES

1. Hing E, Rui P, Palso K. National Ambulatory Medical Care Survey: 2013 State and National Summary Tables. Available at: www.cdc.gov/nchs/ahcd/ahcd_products.htm. Accessed May 18, 2018.
2. Jonsson F, Yin L, Lundholm C, et al. Low-dose aspirin use and cancer characteristics: a population-based cohort study. *Br J Cancer*. 2013;109:1921–1925.
3. Richman IB, Owens DK. Aspirin for primary prevention. *Med Clin North Am*. 2017;101:713–724.

4. Pakhomov SV, Shah ND, Hanson P, et al. Automated processing of electronic medical records is a reliable method of determining aspirin use in populations at risk for cardiovascular events. *Inform Prim Care*. 2010;18:125–133.
5. VanWormer JJ, Miller AW, Rezkalla H. Identifying opportunities to improve aspirin utilization for the primary prevention of cardiovascular disease in a regional health care system. *WMJ*. 2014;113:190–195; quiz 196.
6. Chan AT, Giovannucci EL, Meyerhardt JA, et al. Aspirin dose and duration of use and risk of colorectal cancer in men. *Gastroenterology*. 2008;134:21–28.
7. Chan AT, Arber N, Burn J, et al. Aspirin in the chemoprevention of colorectal neoplasia: an overview. *Cancer Prev Res*. 2012;5:164–178.
8. Huang WK, Chiou MJ, Yu KH, et al. The association between low-dose aspirin use and the incidence of colorectal cancer: a nationwide cohort study. *Aliment Pharmacol Ther*. 2013;38:432–439.
9. Imperiale TF. Aspirin and the prevention of colorectal cancer. *N Engl J Med*. 2003;348:879–880.
10. Patrignani P, Patrono C. Aspirin and cancer. *J Am Coll Cardiol*. 2016;68:967–976.
11. Rothwell PM, Wilson M, Price JF, et al. Effect of daily aspirin on risk of cancer metastasis: a study of incident cancers during randomised controlled trials. *Lancet*. 2012;379:1591–1601.
12. National Quality Forum. Endorsement summary: cardiovascular measures. Washington, DC. 2012. Available at: www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=69822. Accessed May 18, 2018.
13. Grenon SM, Owens CD, Alley H, et al. Posttraumatic stress disorder is associated with worse endothelial function among Veterans. *J Am Heart Assoc*. 2016;5:e003010.
14. Fedirko V, Bradshaw PT, Figueiredo JC, et al. Urinary metabolites of prostanoids and risk of recurrent colorectal adenomas in the Aspirin/Folate Polyp Prevention Study (AFPPS). *Cancer Prev Res*. 2015;8:1061–1068.
15. Liu M, Jiang M, Kawai VK, et al. Modeling drug exposure data in electronic medical records: an application to warfarin. *AMIA Annu Symp Proc*. 2011;2011:815–823.
16. Manocha D, Singh M, Mehta N, et al. Bleeding risk after invasive procedures in aspirin/NSAID users: polypectomy study in veterans. *Am J Med*. 2012;125:1222–1227.
17. Zheng C, Rashid N, Koblick R, et al. Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. *Clin Ther*. 2015;37:2048.e2–2058.e2.
18. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. 2017;24:986–991.
19. Gupta S, Liu L, Patterson OV, et al. A framework for leveraging "Big Data" to advance epidemiology and improve quality: design of the VA colonoscopy collaborative. *EGEMs (Wash DC)*. 2018;6:4.
20. US Department of Veterans Affairs. National Centers for Veterans Analysis and Statistics. Veteran Population. Available at: www.va.gov/vetdata/Veteran_Population.asp. Accessed January 8, 2018.
21. Sanchez-Pi N, Martí L, Garcia ACB. Improving ontology-based text classification: an occupational health and security application. *J Appl Logic*. 2016;17:48–58.
22. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform*. 2014;9:97–104.
23. Rule A, Rick S, Chiu M, et al. Validating free-text order entry for a note-centric EHR. *AMIA Annu Symp Proc*. 2015;2015:1103–1110.
24. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20:144–151.
25. Zhou L, Mahoney LM, Shakurova A, et al. How many medication orders are entered through free-text in EHRs?—a study on hypoglycemic agents. *AMIA Annu Symp Proc*. 2012;2012:1079–1088.
26. Lakhani P, Menschik ED, Goldszal AF, et al. Development and validation of queries using structured query language (SQL) to determine the utilization of comparison imaging in radiology reports stored on PACS. *J Digit Imaging*. 2006;19:52–68.
27. Fisk JM, Mutalik P, Levin FW, et al. Integrating query of relational and textual data in clinical databases: a case study. *J Am Med Inform Assoc*. 2003;10:21–38.