Annals of Epidemiology 25 (2015) 519–525

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Enrollment factors and bias of disease prevalence estimates in administrative claims data

Elizabeth T. Jensen MPH, PhD [a,b,*], Suzanne F. Cook PhD [c], Jeffery K. Allen MS [c], John Logie PhD [c], Maurice Alan Brookhart PhD [d], Michael D. Kappelman MD, MPH [b,e], Evan S. Dellon MD, MPH [a,b]

[a] Center for Esophageal Diseases and Swallowing, Division of Gastroenterology and Hepatology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill
[b] Center for Gastrointestinal Biology and Disease, Division of Gastroenterology and Hepatology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill
[c] World Wide Epidemiology, GlaxoSmithKline, Research Triangle Park
[d] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill
[e] Division of Pediatric Gastroenterology and Hepatology, Department of Pediatrics, University of North Carolina School of Medicine, Chapel Hill

ARTICLE INFO

Article history:
Received 12 January 2015
Accepted 11 March 2015
Available online 21 March 2015

Keywords:
Prevalence
Administrative claims data
Selection bias

ABSTRACT

Purpose: Considerations for using administrative claims data in research have not been well-described. To increase awareness of how enrollment factors and insurance benefit use may contribute to prevalence estimates, we evaluated how differences in operational definitions of the cohort impact observed estimates.

Methods: We conducted a cross-sectional study estimating the prevalence of five gastrointestinal conditions using MarketScan claims data for 73.1 million enrollees. We extracted data obtained from 2009 to 2012 to identify cohorts meeting various enrollment, prescription drug benefit, or health care utilization characteristics. Next, we identified patients meeting the case definition for each of the diseases of interest. We compared the estimates obtained to evaluate the influence of enrollment period, drug benefit, and insurance usage.

Results: As the criteria for inclusion in the cohort became increasingly restrictive the estimated prevalence increased, as much as 45% to 77% depending on the disease condition and the definition for inclusion. Requiring use of the insurance benefit and a longer period of enrollment had the greatest influence on the estimates observed.

Conclusions: Individuals meeting case definition were more likely to meet the more stringent definition for inclusion in the study cohort. This may be considered a form of selection bias, where overly restrictive inclusion criteria definitions may result in selection of a source population that may no longer represent the population from which cases arose.

Introduction

Administrative health care claims data offer the opportunity to study, at the population level, disease comorbidities, health care utilization patterns, and longitudinal studies of health outcomes. Frequently, claims data have been used in pharmacoepidemiologic studies. Because of the large number of patients included, administrative claims data have been increasingly used for studies of disease incidence and prevalence. For rare disease, claims data are one of the few resources available for assembling a sufficiently large enough cohort of cases for study. Estimates of incidence and prevalence provide a basis for research or health care service resource allocation and inform public health efforts for disease prevention.

Although numerous papers have been published on validation of disease-specific algorithms for case identification in administrative claims data [1–8], and some methodological articles present case algorithms and strategies to maximize sensitivity or specificity [9,10], there has been little discussion of how enrollment factors for the health plan benefit could influence prevalence estimates. Estimating prevalence, or more specifically, a period prevalence, in administrative claims data necessitates defining an enrollment period from which the source population arises in addition to identification of cases within the source population. Given the variability in benefit plans, this may introduce bias when estimating disease prevalence. For example, not all enrollees have a

No potential conflicts of interest relevant to this article were reported.

* Corresponding author. Center for Gastrointestinal Biology and Disease, Division of Gastroenterology and Hepatology, Department of Medicine, University of North Carolina School of Medicine, CB #7080—4140 Bioinformatics Building, 130 Mason Farm Rd, Chapel Hill, NC 27599-7080. Tel.: +1-919-428-5091; fax: +1-919-843-2508.
E-mail address: elizabeth_jensen@med.unc.edu (E.T. Jensen).

http://dx.doi.org/10.1016/j.annepidem.2015.03.008
1047-2797/

prescription drug benefit, there are differences in lengths of enrollment periods, and there are different methods for defining enrollment periods. However, the effect of these differences has, to our knowledge, never been examined on prevalence estimates.

Our primary objective was to identify factors intrinsic to use of administrative claims data that may bias estimates of disease prevalence. Specifically, our aims were to (1) assess the influence of selection of enrollment period, using a minimum enrollment versus fixed enrollment period, on prevalence estimates, (2) assess the influence of selection of continuous (without interruption) versus total enrollment (sum of continuous periods of enrollment when there was >1 enrollment period), (3) assess the influence of restriction to plans with pharmacy benefit only versus without restriction, and (4) assess the influence of restriction of the source population to patients who have evidence of having used their benefit plan.

Materials and methods

We conducted a cross-sectional study using the MarketScan administrative claims database (Truven Health Analytics, Ann Arbor, MI). This resource captures person-specific clinical utilization, expenditures, and enrollment information across inpatient, outpatient, and prescription drug services from a selection of large employers, health plans, and government and public organizations in the United States. The database includes commercial health data from approximately 100 payers. We restricted the data sample to individuals aged 0 to 64 years, as individuals of 65 years and older may have dual enrollment in both a commercial and government-sponsored Medicare insurance plan.

We used *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM) codes to characterize the disease status for several chronic gastrointestinal conditions, selected to represent a range of frequency of health care encounters and severities, namely Crohn's disease, ulcerative colitis, Barrett's esophagus, eosinophilic esophagitis, and celiac disease. Case definitions were adapted from case algorithms previously applied in an administrative claims data setting (Appendix Table A). There were no exclusions for insurance plan type; data were generated from claims arising through coverage from commercially provided insurance. No data were available on the specific insurance provider used.

We used data for individuals enrolled continuously for 6 months or more between January 1, 2009 through December 31, 2012 to allow a minimum period of time for a diagnostic code(s) for a given condition to be documented, based on the anticipated pattern of care for the individual diseases. We examined the enrollment and demographic features of patients with the conditions previously mentioned as compared to the source population and tested for statistically significant differences in the enrollment factors (Satterthwaite *t*-test for difference in mean days of enrollment; χ^2 tests for difference in proportion of more than one enrollment period and proportion with a prescription drug benefit). The mean period of enrollment was calculated from the number of contiguous days the patient was enrolled from January 1, 2009 through December 31, 2012. For patients with gaps in enrollment, the mean duration of enrollment was based on the longest single period of continuous enrollment. Changes in health plan status are generally linkable in the MarketScan data. Therefore, enrollees who change health plans when changing employment are maintained as continuous beneficiaries when there is no interruption in coverage. Roughly 95% of enrollees had only a single period of enrollment during the period of study (Appendix Table B).

To evaluate the influence of enrollment factors on prevalence estimates, we calculated the prevalence of each condition after varying criteria for inclusion in the source population (i.e., the denominator) from which the cases arose. These criteria included (1) duration of enrollment, (2) enrollment continuity, (3) prescription benefit status, and (4) use of the health insurance benefit. For evaluating the influence of duration of enrollment, we examined estimates based on inclusion of enrollees with 6 months or more and 12 months of enrollment, and then, prevalence within finite enrollment periods of 12 or 24 months. We also examined estimates when restricting the source population to those enrolled greater than or equal to the mean number of days for cases from each of the disease definitions. All analyses were based on length of continuous enrollment, with the exception of the analyses described as total enrollment. Prevalence was calculated by dividing the number of individuals from the source population meeting the case definition by the total number of individuals within the source population, as defined by these enrollment factors and definitions. Data were restricted to claims made within the period of January 1, 2009 through December 31, 2012. Wzhere a finite enrollment period was specified, prevalence was based on diagnoses within this defined period. All prevalence estimates represent a period prevalence for the enrollment period specified.

The specific parameters for the analysis included the following: (1) enrollee enrollment dates (start and end date) for characterization of enrollment period and for characterizing the mean enrollment period for cases versus the source population; (2) continuous versus total enrollment, for characterizing differences in prevalence estimates when restricting the source population to those with or without continuous enrollment for the defined period of interest (e.g., a patient could be characterized as having ≥12 months of continuous enrollment or ≥12 months of total enrollment within a given period of time); and (3) whether the enrollee had a pharmacy benefit, for assessing the influence of restricting the source population to patients with a pharmacy drug benefit. In a secondary analysis, we evaluated whether estimates observed were influenced by restriction to health plan enrollees who had evidence of having used their benefit. We characterized the enrollee as a user of their health plan benefit if there was documentation of one or more instance of an ICD-9-CM code, Common Procedural Terminology code, National Drug Code, or Healthcare Common Procedure Coding System code during the enrollment period specified for each of the prevalence definitions. In instances where there was more than one period of continuous enrollment that met the criteria for inclusion, evidence of meeting the case definition could occur in any period that met criteria for inclusion. Finally, based on our observation that the enrollment characteristics of enrollees meeting case definition were different than those not meeting case definition, we restricted the source population to patients with a minimum duration of continuous enrollment as defined by the mean duration of enrollment for a particular disease condition and, separately, enrollees with a minimum duration of enrollment as defined by the mean duration of enrollment for the source population. As this study used de-identified data, it is not considered human subjects research and was exempt from Institutional Review Board's review.

Results

There were 73,129,577 enrollees that met study inclusion criteria of continuous enrollment for at least 6 months during the period of January 1, 2009 through December 31, 2012—93.2% of these patients contributed only a single period of continuous enrollment during the study period.

A comparison of the enrollment and demographic features of cases versus the underlying source population identified differences in length of enrollment and the proportion with a drug benefit (Appendix Table B). Individuals meeting the case definition, across all disease conditions, were enrolled longer than the source population from which they arose. For example, the mean enrollment for