



Published in final edited form as:

*Ann Epidemiol.* 2015 July ; 25(7): 519–525.e2. doi:10.1016/j.annepidem.2015.03.008.

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

Elizabeth T. Jensen, MPH PhD<sup>1,2</sup>, Suzanne F. Cook, PhD<sup>3</sup>, Jeffery K. Allen, MS<sup>3</sup>, John Logie, PhD<sup>3</sup>, M. Alan Brookhart, PhD<sup>4</sup>, Michael D. Kappelman, MD MPH<sup>2,5</sup>, and Evan S. Dellon, MD MPH<sup>1,2</sup>

<sup>1</sup>Center for Esophageal Diseases and Swallowing, Division of Gastroenterology and Hepatology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, NC

<sup>2</sup>Center for Gastrointestinal Biology and Disease, Division of Gastroenterology and Hepatology, Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, NC

<sup>3</sup>GlaxoSmithKline, Gillings Global School of Public Health, University of North Carolina, Chapel Hill, NC

<sup>4</sup>Department of Epidemiology, Gillings Global School of Public Health, University of North Carolina, Chapel Hill, NC

<sup>5</sup>Division of Pediatric Gastroenterology and Hepatology, Department of Pediatrics, University of North Carolina School of Medicine, Chapel Hill, NC

### 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 observed estimates, we evaluated how differences in operational definitions of the cohort impacted estimates of disease prevalence.

**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 2009–2012 to identify cohorts meeting various enrollment, prescription drug benefit, or healthcare 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.

---

**Corresponding Author:** Elizabeth Jensen, MPH PhD, CB #7080 – 4140 Bioinformatics Building, 130 Mason Farm Rd., UNC-CH, Chapel Hill, NC 27599-7080, Phone: 1-919-428-5091, Fax: 1-919-843-2508, elizabeth\_jensen@med.unc.edu.

Potential competing interests: None

Specific author contributions (all authors approved the final draft):

Jensen: Project conception, study design, data interpretation, manuscript drafting, critical revision

Cook: Study design, data interpretation, critical revision

Allen: Study design, data analyses, data interpretation, critical revision

Logie: Study design, data analyses, data interpretation, critical revision

Brookhart: Project conception, study design, data interpretation, critical revision

Kappelman: Study design, data interpretation, critical revision

Dellon: Project conception, study design, data interpretation, manuscript drafting, critical revision

**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 in the cohort. Requiring use of the 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 cohort definitions may result in selection of a study population that may no longer represent the source population.

### Keywords

Prevalence; Administrative claims data; Selection bias

---

## INTRODUCTION

Administrative healthcare claims data offer the opportunity to study, at the population level, disease comorbidities, healthcare 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. This type of epidemiologic research provides a basis for research or healthcare service resource allocation and informs public health efforts for disease prevention.

While numerous papers have been published on validation of disease-specific algorithms for case identification in administrative claims data, [1–8] and some methodological papers 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 prescription drug benefit, there are differences in lengths of enrollment periods, and there are different methods for defining enrollment periods. However, the impact 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 age 0–64 years, as individuals age 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 (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 (Supplementary 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 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 above 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; chi-square tests for difference in proportion of >1 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 (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 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 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 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 population meeting the case definition by the total number of individuals within the 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. Where 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: 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 (for example, 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 1 instance of an ICD-9 CM code, Common Procedural Terminology (CPT) code, National Drug Code (NDC), or Healthcare Common Procedure Coding System (HCPCS) code during the enrollment period specified for each of the prevalence definitions. In instances where there was >1 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 cohort 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 IRB 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 (Supplementary 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 the source population was 688 days, but mean

enrollment for cases ranged from 814 to 901 days. There were also differences between cases and the source population in the proportion with >1 continuous enrollment period, with cases less likely to have gaps in enrollment (Supplementary Table B). Similarly, individuals meeting case definition were more likely to have a prescription drug benefit in their insurance plan (Supplementary Table B).

### Variation in prevalence estimates based on enrollment periods

In general, longer enrollment periods were associated with higher prevalence relative to shorter enrollment periods, and using the mean enrollment period of cases as the minimum requirement yielded the highest prevalence estimate (Table 1). For example, for Crohn's disease, the prevalence estimate for 6 months of enrollment was 153.1 cases/100,000 and the estimate obtained when restricting to the longer enrollment period increased to 212.1 cases/100,000 (27% increase). As described above, most enrollees did not have gaps in their enrollment and the prevalence estimates obtained from using minimum total enrollment, as compared to minimum continuous enrollment, were qualitatively the same (Supplementary Table C).

Source populations using finite enrollment periods (i.e. 12 or 24 months of calendar time), where the source population and cases contributed the same length of enrollment, generated differences in prevalence estimates with prevalence higher in populations followed for longer periods (Table 2). For example, prevalence for celiac disease in the 12 months for 2010 and 2011 was 32.6/100,000 and 37.7/100,000, respectively, while the 24 month prevalence from 2010–2011 was 66.6/100,000 (43% increase). Similarly, prevalence for Barrett's esophagus in the 12 months for 2010 and 2011 was 68.3 and 76.7/100,000, respectively, while the 24-month prevalence was 147/100,000 (49% increase). Prevalence estimates also increased with more recent enrollment periods, across all disease conditions (Table 2).

### Variation in prevalence estimates based on drug benefits or benefit utilization

Requiring evidence of a prescription drug benefit for study inclusion shifted the prevalence estimates slightly higher, for both the minimum enrollment (Table 3) and finite enrollment period (Table 4) approaches. For example, for Crohn's disease, inclusion of those enrolled during the finite 12-month period (2011) resulted in an increase in prevalence of Crohn's disease from 119.4 to 121.2 for those with a drug benefit (Tables 2 and 4). Without the additional criteria of requiring a drug benefit, the prevalence among those enrolled 6 months was 153.1 cases/100,000 (Table 3). Restricting the source population to those who were users of the health care system shifted estimates higher. For example, among those enrolled 24 months, the prevalence of Crohn's disease increased from 172.4 to 186.2 with restriction of the source population to those individuals with evidence of using their insurance benefit (7% increase) (Tables 2 and 4).

Prevalence estimates obtained using the least restrictive criteria (6 months of total enrollment) were markedly different as compared to the estimates obtained using the most restrictive criteria we applied (minimum enrollment period as defined by the mean enrollment period of the cases, for those who were users of their benefit). For Crohn's

disease this represented a 45% increase (153.1 cases/100,000 versus 221.8 cases/100,000), for ulcerative colitis a 65% increase (138.2 cases/100,000 versus 228.1 cases/100,000), for eosinophilic esophagitis a 70% increase (59.5 cases/100,000 versus 101.4 cases/100,000), for Barrett's esophagus a 77% increase (133.6 cases/100,000 versus 237.0 cases/100,000), and for celiac disease a 69% increase (58.5 cases/100,000 versus 98.8 cases/100,000) in prevalence.

## DISCUSSION

There are many examples of publications in the scientific literature reporting prevalence estimates obtained from administrative claims data using a variety of approaches. These include requiring a minimum enrollment period for the study population[1–4] restricting the study population to a finite period of time where everyone in the study population has the same opportunity to become a case,[5–8] and restricting the study population to those individuals who have at least one claim in the period of interest and/or have indication of having a prescription drug benefit as part of their health plan.[11] However, there is no consensus in the literature about the optimal way to approach this, or consideration of biases introduced by using a specific methodology. In the present analysis, we illustrate how these different approaches may yield substantially different estimates.

Estimating period prevalence when using administrative claims data necessitates not only employing an appropriate, ideally validated, algorithm to identify cases of interest (which is where much of the literature is focused), but also careful consideration of possible differences in enrollment characteristics for those with and without evidence of meeting the case definition. In the present analysis, individuals meeting case definitions had longer plan enrollment, greater continuity of enrollment, and were more likely to have a prescription drug benefit than individuals in the source population. By definition, cases are also more likely to have used their benefit plan. This observations are not unexpected, and may suggest that patients with chronic health conditions preferentially select and remain enrolled in plans with more comprehensive coverage. However, we also considered that the enrollment characteristics of cases may not be attributable to the disease -- rather patients with those enrollment characteristics may be more likely to be diagnosed. If this were to be true, then estimating prevalence from a less restrictive population (those enrolled for less time, or without a drug benefit) would result in a biased estimate as the cohort would not be representative of the source population giving rise to the cases. The demographic features of the individuals meeting the case definitions may contribute to the observed differences in enrollment characteristics, however the disease conditions represented individuals at varying age and sex distributions (Supplementary Table B). In other words, the pattern of longer enrollment periods, fewer disruptions in enrollment, and increase in proportion with a drug benefit, observed for cases as compared to the source population, was consistent across disease conditions that represented similar mean ages than the source population, and higher or lower proportion of males than the source population.

There are several important implications of these findings. First, applying increasingly stringent definitions for inclusion in the study population resulted in increasing prevalence estimates. The difference in prevalence estimates obtained across definitions was significant,

with increases in prevalence estimates from 45% to 77% based purely on how the study population was defined. This was because individuals meeting case definition were more likely to meet the more stringent definition for inclusion. This may be considered a form of selection bias, where overly restrictive inclusion criteria may lead towards selection of a study population that may no longer represent the source population.

However, utilizing more liberal inclusion criteria may not necessarily be the best practice either. Shorter enrollment requirements or not restricting to individuals with a pharmacy benefit may result in under-ascertainment of true cases, as sufficient observation time is required for patients to incur the necessary number and type of claims to meet administrative case definitions. This is particularly true for case definitions that require multiple claims, and for conditions in which episodes of care may be less frequent (i.e. it will take longer amounts of time to accumulate the necessary claims). We did not examine the influence of how the number of claims required for case definition could influence the estimates observed. Our case definitions were based on definitions that had been previously applied in the literature for these conditions. The association between increased restrictive inclusionary criteria and higher prevalence estimates was consistent across disease conditions, irrespective of the number of claims required for a given case definition.

Likely, selecting the appropriate enrollment period should depend on the disease being studied. Disease-specific factors would influence this, including disease severity, patterns of care (i.e. frequency of health encounters for well and poorly controlled illness), whether the disease is episodic, chronic or limited to a single acute episode, and the relative rarity of the disease. For conditions where patients receive frequent, regular healthcare (and claims), less restrictive enrollment requirements may be appropriate as this will ensure that the study population best represents the source population. However, for patients with less frequent or more episodic care, more restrictive enrollment criteria are needed in order to minimize under-ascertainment of cases. Therefore, a detailed understanding of the expected patterns of care for the disease under study as well as clinical judgment will be critical in establishing appropriate inclusion criteria. For conditions requiring infrequent healthcare, it is likely that administrative data may not be able to accurately identify disease prevalence. In our study, we examined only chronic health conditions. The influence of enrollment characteristics on estimates may differ for acute conditions. It may be that patients with acute illness are more similar to the source population as potentially the disease status may not have influenced their health coverage or length of enrollment. Conversely, patients with more comprehensive coverage may be more likely to seek care for an acute condition. These questions should be examined in future research.

This is not simply an academic issue, as there are potential real-world implications of these data. For example, these differences could influence rare disease status, defined as less than 200,000 prevalent cases in the United States, which could impact orphan drug status and pharmaceutical development.[12] A lack of standardization in approach, across studies, could result in misleading assessments of changing disease prevalence across time, which could overemphasize or underemphasize the importance of a disease. This, in turn, could also lead to misinterpretation of the burden of disease, resulting in ill-informed resource utilization for disease research, prevention, and treatment.

Administrative claims data are also frequently used in epidemiologic studies of exposure-disease and disease-outcome. The differences we observed in enrollment characteristics of cases and non-cases may be relevant to these studies, particularly when these enrollment factors may also contribute to exposure or outcomes status, in which case enrollment factors could be an important source of confounding bias.

A limitation to this study was our inability to differentiate between enrollees who represent healthier individuals and those who have multiple sources of coverage. Estimates obtained when restricting the study population to users of their benefit increased prevalence estimates substantially. Although requiring some use of benefit may filter out those enrollees with more than one health plan, this approach may also exclude healthier patients and inflate prevalence estimates. We also do not know the quality of the coverage provided. This could have an influence on an enrollee's likelihood to seek a diagnosis and/or follow-up treatment and thus contribute to their likelihood for being identified as a case.

Additionally, we are not able to know the "true" prevalence of the selected diseases to definitively conclude which approach is ideal. There is a potential that some patients may have been misclassified as having a disease condition as ICD-9 codes are imperfect in their measure of true disease status. However, knowing the actual prevalence was of secondary importance for the aims of this study, which involved demonstrating changes in the prevalence based on definition of the study population.

There are also a number of strengths to this study. First, we utilized a large administrative claims database that is widely used by academia and industry for research purposes. This allowed us to identify large numbers of cases for a range of conditions with different patterns of care, disease severity, and calculated prevalence to evaluate how enrollment factors could influence estimates in these different settings. Irrespective of disease type, the pattern of increasing prevalence with increasing restrictiveness of the inclusion criteria persisted across disease conditions. Second, we used a number of different study population definitions, but which also allow consideration for different patterns of care. Finally, we not only assessed enrollment periods, but also pharmacy benefits and use of benefits. Our results were consistent across a number of diseases, however we limited our study to chronic conditions of the gastrointestinal system. Our observations may not be generalizable to other disease conditions or conditions that are acute.

## CONCLUSION

In conclusion, we demonstrate that the definition of the study population markedly impacts disease prevalence estimates in administrative claims data. This effect has not yet been reported in the literature, and likely impacts many published prevalence estimates. While our study design does not necessarily allow us to determine the "best" approach to defining the study population, we advise that researchers consider these issues when assembling a cohort for estimating prevalence, define the population for inclusion in a way in which reflects the appropriate patterns of care, explicitly report the rationale for their choices, and consider performing sensitivity analyses with different enrollment period definitions to examine the impact on the prevalence estimate. Overall, the minimum enrollment period



approach, with fewer restrictions in defining the study population, may lead to an underestimate of prevalence, particularly for conditions that require infrequent health care contacts. However, more restrictive approaches, using a finite enrollment period, requiring use of the benefit and/or a prescription benefit, may lead to an overestimate of prevalence, or a prevalence estimate representative of individuals with longer enrollment periods and a higher levels of benefit use and/or coverage. Although administrative claims data offer a rich opportunity for epidemiologic study of diseases, including descriptive studies of disease prevalence, recognition and documentation of the limitations and considerations for identifying an appropriate study population, a population reflective of the source population giving rise to the cases, is essential for using these data.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Financial support: This study was funded, in part, by NIH Award K23 DK090073 (ESD) and by GlaxoSmithKline (SFC, JKA and JL). MAB receives investigator-initiated research funding from the National Institutes of Health (R01 AG042845, R21 HD080214, R01 AG023178) and through contracts with the Agency for Healthcare Research and Quality's DEcIDE program and the Patient Centered Outcomes Research Institute.

## Abbreviations

<b>ICD-9 CM</b>	International Classification of Disease, Ninth Revision, Clinical Modification diagnosis codes
<b>CPT</b>	Current Procedural Terminology codes
<b>HCPCS</b>	Healthcare Common Procedure Coding System codes
<b>NDC</b>	National Drug Codes
<b>EoE</b>	Eosinophilic Esophagitis

## REFERENCES

1. Li S, Peng Y, Weinhandl ED, Blaes AH, Cetin K, Chia VM, et al. Estimated number of prevalent cases of metastatic bone disease in the US adult population. *Clinical epidemiology*. 2012; 4:87–93. [PubMed: 22570568]
2. Dellon ES, Jensen ET, Martin CF, Shaheen NJ, Kappelman MD. The Prevalence of Eosinophilic Esophagitis in the United States. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*. 2013
3. Dufour R, Joshi AV, Pasquale MK, Schaaf D, Mardekian J, Andrews GA, et al. The prevalence of diagnosed opioid abuse in commercial and Medicare managed care populations. *Pain practice : the official journal of World Institute of Pain*. 2014; 14:E106–E115. [PubMed: 24289539]
4. Prescott JD, Factor S, Pill M, Levi GW. Descriptive analysis of the direct medical costs of multiple sclerosis in 2004 using administrative claims in a large nationwide database. *Journal of managed care pharmacy : JMCP*. 2007; 13:44–52. [PubMed: 17269836]
5. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. *Dig Dis Sci*. 2013; 58:519–525. [PubMed: 22926499]

6. Khanna R, Madhavan SS, Bhanegaonkar A, Remick SC. Prevalence, healthcare utilization, and costs of breast cancer in a state Medicaid fee-for-service program. *J Womens Health (Larchmt)*. 2011; 20:739–747. [PubMed: 21417935]
7. Cosmatos I, Matcho A, Weinstein R, Montgomery MO, Stang P. Analysis of patient claims data to determine the prevalence of hidradenitis suppurativa in the United States. *Journal of the American Academy of Dermatology*. 2013; 69:819. [PubMed: 24124817]
8. Gleason PP, Alexander GC, Starner CI, Ritter ST, Van Houten HK, Gunderson BW, et al. Health plan utilization and costs of specialty drugs within 4 chronic conditions. *Journal of managed care pharmacy : JMCP*. 2013; 19:542–548. [PubMed: 23964615]
9. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *Journal of clinical epidemiology*. 2012; 65:343–349. e2. [PubMed: 22197520]
10. 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. *Pharmacoepidemiology and drug safety*. 2012; 21(Suppl 1):82–89. [PubMed: 22262596]
11. Grosse SD, Boulet SL, Grant AM, Hulihan MM, Faughnan ME. The use of US health insurance data for surveillance of rare disorders: hereditary hemorrhagic telangiectasia. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2014; 16:33–39. [PubMed: 23703685]
12. Developing products for rare diseases and conditions. Food and Drug Administration.
13. Rybnicek DA, Hathorn KE, Pfaff ER, Bulsiewicz WJ, Shaheen NJ, Dellon ES. Administrative coding is specific, but not sensitive, for identifying eosinophilic esophagitis. *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus / ISDE*. 2013
14. Amonkar MM, Kalsekar ID, Boyer JG. The economic burden of Barrett's esophagus in a Medicaid population. *The Annals of pharmacotherapy*. 2002; 36:605–611. [PubMed: 11918506]
15. Tanpowpong P, Broder-Fingert S, Obuch JC, Rahni DO, Katz AJ, Leffler DA, et al. Multicenter study on the value of ICD-9-CM codes for case identification of celiac disease. *Annals of epidemiology*. 2013; 23:136–142. [PubMed: 23313264]

**Table 1**

Prevalence estimates applying a minimum period of enrollment criteria

Disease	Prevalence definition						Percent change in estimate <sup>d</sup>
	Cases/100,000	6 months continuous <sup>a</sup> enrollment	12 months continuous <sup>a</sup> enrollment	source population <sup>b</sup>	mean continuous <sup>d</sup> enrollment of	mean continuous <sup>d</sup> enrollment of cases <sup>b</sup>	
Crohn's disease	Cases/100,000	153.1	169.1	197.4	212.1	27%	
	n	111,511	100,080	75,500	52,237		
Ulcerative colitis	Cases/100,000	138.2	156.6	194.3	218.3	36%	
	n	100,660	92,717	74,317	51,335		
EoE	Cases/100,000	59.5	67.0	84.4	97.0	38%	
	n	43,356	39,644	32,272	22,810		
Barrett's esophagus	Cases/100,000	133.6	155.1	199.5	227.0	40%	
	n	97,353	91,777	76,289	52,322		
Celiac disease	Cases/100,000	58.5	66.5	83.7	94.6	37%	
	n	42,622	39,342	32,003	22,239		
Source population	n	72,849,131	59,190,979	38,240,134	Variable <sup>c</sup>		

<sup>a</sup> Definitions based on continuous enrollment for period specified for any enrollee meeting the enrollment definition specified during 2009–2012

<sup>b</sup> Definition for inclusion in source population based on individual having mean enrollment period of source population (595 days for continuous enrollment) or cases (varies by disease condition as described in Table 2) within 2009–2012

<sup>c</sup> Crohn's disease – n=25,670,560; ulcerative colitis – n=23,993,061; EoE – n=24,480,235; Barrett's esophagus – n=24,486,080; celiac disease – n=24,474,976

<sup>d</sup> Change in estimate from least ( 6 months enrollment) to most restrictive ( mean enrollment of cases) inclusion criteria

**Table 2**

Prevalence estimates applying a finite period of enrollment criteria

Disease		Prevalence definition <sup>a</sup>					Percent change in estimate <sup>b</sup>
		2009	2010	2011	2009–2010	2010–2011	
Crohn's disease	Cases/100,000	109.7	113.5	119.4	164.2	172.4	31%
	n	31,414	36,765	42,862	30,195	40,981	
Ulcerative colitis	Cases/100,000	82.7	87.3	92.0	154.3	159.5	42%
	n	23,671	28,301	33,016	28,373	37,915	
EoE	Cases/100,000	26.7	34.5	42.0	53.4	68.0	38%
	n	7,642	11,178	15,065	9,828	16,152	
Barrett's esophagus	Cases/100,000	64.8	68.3	76.7	139.1	147.0	49%
	n	18,544	22,114	27,517	25,586	34,922	
Celiac disease	Cases/100,000	27.2	32.6	37.7	54.5	66.6	43%
	n	7,779	10,550	13,523	10,017	15,817	
Source population	n	28,629,771	32,400,533	35,896,982	18,390,346	23,764,543	

<sup>a</sup> Definitions based on having continuous enrollment for period specified for any enrollee meeting the enrollment definition specified

<sup>b</sup> Change in estimate from least (12 months of continuous enrollment) to most restrictive (24 months of continuous enrollment) inclusion criteria

**Table 3**

Prevalence estimates applying additional criteria for inclusion and finite enrollment

Disease		Prevalence definition											
		2009		2010		2011		2009-2010		2010-2011			
		Drug benefit	Users of benefit	Drug benefit	Users of benefit	Drug benefit	Users of benefit	Drug benefit	Users of benefit	Drug benefit	Users of benefit	Drug benefit	Users of benefit
Crohn's disease	Cases/100,000	110.2	127.1	115.0	133.2	121.1	138.8	163.9	175.0	173.6	186.2	186.2	186.2
	n	25,048	31,414	28,664	36,765	33,099	42,862	25,648	30,195	31,961	40,981	40,981	40,981
Ulcerative colitis	Cases/100,000	82.8	95.8	87.0	102.5	92.0	106.9	153.0	164.4	159.2	172.3	172.3	172.3
	n	18,823	23,671	21,693	28,301	25,141	33,016	23,938	28,373	29,302	37,915	37,915	37,915
EoE	Cases/100,000	27.5	30.9	35.4	40.5	42.8	48.8	54.1	56.9	69.3	73.4	73.4	73.4
	n	6,253	7,642	8,826	11,178	11,689	15,065	8,459	9,828	12,760	16,152	16,152	16,152
Barrett's esophagus	Cases/100,000	62.1	75.0	65.4	80.1	72.4	89.1	134.7	148.3	141.4	158.7	158.7	158.7
	n	14,125	18,544	16,313	22,114	19,795	27,517	21,074	25,586	26,031	34,922	34,922	34,922
Celiac disease	Cases/100,000	27.0	31.5	31.9	38.2	36.9	43.8	53.9	58.0	65.0	71.9	71.9	71.9
	n	6,127	7,779	7,958	10,550	10,076	13,523	8,427	10,017	11,965	15,817	15,817	15,817
Source population	n	22,730,936	24,710,617	24,925,485	27,598,619	27,326,188	30,887,755	16,645,359	17,258,280	18,408,647	22,010,598	22,010,598	22,010,598

<sup>a</sup> Definitions based on continuous enrollment for period specified for any enrollee meeting the enrollment definition specified

<sup>b</sup> Documentation of the enrollee having a prescription drug benefit as part of their health insurance plan

<sup>c</sup> Documentation of the enrollee having at least 1 CPT, NDC, IDC-9, HCPCS code during the enrollment period specified

**Table 4**  
Prevalence estimates applying additional criteria for inclusion and minimum enrollment

Disease	Prevalence definition <sup>d</sup>										Percent change in estimate
	6 months enrollment		12 months enrollment		mean enrollment of source population <sup>d</sup>		mean enrollment of cases <sup>d</sup>		Drug benefit <sup>b</sup>	Users of benefit <sup>c</sup>	
Crohn's disease	Cases/100,000	155.8	173.9	173.2	186.6	199.6	213.9	214.9	221.8	221.8	
	n	85,454	111,511	76,660	100,080	58,241	64,016	41,730	52,237	52,237	
Ulcerative colitis	Cases/100,000	140.1	157.0	159.9	172.8	195.7	213.3	220.5	228.1	228.1	
	n	76,877	100,660	70,762	92,717	57,106	63,833	40,863	51,335	51,335	
EoE	Cases/100,000	61.8	67.6	69.9	73.9	87.0	92.3	99.1	101.4	101.4	
	n	33,903	43,356	30,919	39,644	25,382	27,618	18,363	22,810	22,810	
Barrett's esophagus	Cases/100,000	131.2	151.8	153.2	171.1	194.9	217.1	225.2	237.0	237.0	
	n	71,966	97,353	67,794	91,777	56,880	64,960	39,893	52,322	52,322	
Celiac disease	Cases/100,000	58.6	66.5	67.0	73.3	83.6	91.2	94.2	98.8	98.8	
	n	32,172	42,622	29,648	39,342	24,395	27,306	17,139	22,239	22,239	
Source population	Cases/100,000	54,858,366	64,132,756	44,264,058	53,643,840	29,182,094	29,924,646	Variable <sup>e</sup>	Variable <sup>e</sup>	Variable <sup>e</sup>	
	n	54,858,366	64,132,756	44,264,058	53,643,840	29,182,094	29,924,646	Variable <sup>e</sup>	Variable <sup>e</sup>	Variable <sup>e</sup>	

<sup>a</sup> Definitions based on continuous enrollment for period specified for any enrollee meeting the enrollment definition specified during 2009–2012

<sup>b</sup> Documentation of the enrollee having a prescription drug benefit as part of their health insurance plan

<sup>c</sup> Documentation of the enrollee having at least 1 CPT, NDC, IDC-9, HCPCS code during the enrollment period specified

<sup>d</sup> Definition for inclusion in source population based on individual having mean enrollment period of source population (622 days for drug benefit and 660 days for users of benefit) or cases (varies by disease condition as described in Table 2) within 2009–2012

<sup>e</sup> Crohn's disease – n=19,418,589; ulcerative colitis – n=18,534,133; EoE – n=18,534,831; Barrett's esophagus – n=17,718,307; celiac disease – n=18,187,751

<sup>f</sup> Crohn's disease – n=23,549,560; ulcerative colitis – n=22,508,811; EoE – n=22,504,132; Barrett's esophagus – n=22,074,561; celiac disease – n=22,499,335