Evaluating Methods to Estimate United States Multiple Sclerosis Prevalence from Administrative Health Claims Data

Edward Neuberger

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2019

Committee:

David Veenstra

Emily Beth Devine

Program Authorized to Offer Degree:

Pharmacy

University of Washington

Abstract

Evaluating Methods to Estimate United States Multiple Sclerosis Prevalence from Administrative Health Claims Data

Edward Neuberger

Chair of the Supervisory Committee:
Professor David Veenstra
Department of Pharmacy

Introduction

The prevalence of Multiple Sclerosis (MS) in the United States (US) historically has been poorly described, in part because disease activity is intermittent, making detection using administrative health claims (AHC) data challenging.[1] In a recent study Wallin and colleagues inflated a 3-year cumulative prevalence estimate by 37% to 47% to estimate a 10-year cumulative prevalence to account for case under-ascertainment in AHC data.[2] However, their new estimate is approximately double that of previous analyses.[3-5] The objective of my study was to assess the validity of the 3-year vs. 10-year inflation factor using a US AHC dataset between 2008 and 2017.

Methods

Patients with MS were identified in AHC data from IBM® MarketScan® Research Databases – one of the largest samples of employer-sponsored health plan data comprised of over 250 million unique patients since 1995.[6] Multiple Sclerosis patients were identified based on inpatient,

outpatient, and disease-modifying therapy (DMT) outpatient pharmaceutical claims using a validated algorithm developed by Culpepper and colleagues.[7] Cumulative prevalence was estimated annually over ascertainment periods of 3 years (2015 – 2017) and 10 years (2008 – 2017), and the 2017 cumulative prevalence was compared between the two ascertainment periods. To ensure the I was implementing the algorithm correctly, I compared my 2008 through 2010 results to those obtained in the same analysis conducted by the algorithm developers. Lastly, because the algorithm was validated using DMTs approved through 2010, the effect of including DMTs approved through 2017 was assessed.

Results

The 2017 10-year cumulative prevalence (251/100,000) was 7.6% higher than the 3-year cumulative prevalence (233/100,000). Using the algorithm with only DMTs approved in the US through 2010, the 2017 10-year cumulative prevalence (240/100,000) was 10.1% higher than the 3-year cumulative prevalence (218/100,000). Compared with Culpepper's MarketScan® analysis (2008 – 2010), I identified 5,115 fewer cumulative cases (67,728 total) in the same time period, although the proportion of cases in each age group was not significantly different (Chi-square heterogeneity test, $p = 0.35$).

Conclusion

I found the 3-versus-10-year ascertainment period cumulative prevalence difference for 2017 to be 30 to 40 percentage points smaller than that identified by Culpepper et al. for 2010. The different datasets and time periods used may contribute in part to this discrepancy, but the reason for this difference remains unclear. Until future work elucidates a more robust 3-vs-10-year difference, this method should be used cautiously to extrapolate prevalence estimates.

Introduction:

The prevalence of multiple sclerosis (MS) in the United States has been poorly described, potentially limiting informed resource allocation and policy decisions.[1] In response to this evidence gap, in early 2018 the MS Prevalence Working Group funded by the National Multiple Sclerosis Society performed a strengths, weaknesses, opportunities and threats (SWOT) analysis of approaches to estimate MS prevalence and identified use of administrative health claims (AHC) database analyses as the most feasible approach, although not without challenges.[8] A major limitation of AHC databases is that individuals who are not insured and engaged with the healthcare system are not captured. For chronic conditions such as MS for which contact with the health care system may vary over time or with disease activity, case ascertainment can be difficult and limited observation periods can underestimate prevalence.[9,10]

One method to address this limitation is the use of cumulative prevalence; individuals meeting the case definition in any year during the ascertainment period are considered a case thereafter. Additionally, the use of longer ascertainment periods improves the ability to capture more cases, with lifetime prevalence being the metric of interest (the proportion of a population that at some point in life develops MS). While there is no published literature on the ascertainment period required to approximate lifetime prevalence in MS, investigators in systemic lupus erythematosus (another chronic disease with variable health care utilization over time and disease activity) found a minimum 10-year ascertainment to be sufficient to approach lifetime prevalence in AHC datasets.[10]

In order to calculate cumulative prevalence, a reliable algorithm is necessary to first identify MS cases in AHC datasets. Culpepper et al.[7] developed an AHC case identification algorithm. The authors reviewed previously validated administrative case definitions for MS to create multiple candidate algorithms composed of MS-specific International Classification of Diseases (ICD) diagnosis codes and disease-modifying therapies (DMTs). The 10 candidate algorithms varied in number of inpatient, outpatient, and DMT encounters required to define an individual as a case, as well as the time periods over which to meet the criteria. Their algorithms ranged from including at minimum 2 inpatient ICD codes to at minimum 5 outpatient ICD codes, with and without DMT claims, over ascertainment periods of 1 or 2 years. These algorithms were each applied separately in three AHC datasets (Veterans Affairs, Kaiser Permanente Southern California, and the Canadian Manitoba province) in the population of individuals with at least one health care encounter associated with a diagnosis code for MS. The performance of each algorithm was compared to medical record review as the "gold standard". The preferred algorithm, based on superior test characteristics and consistency across datasets and age groups, defined MS cases as individuals with three or more MS-related encounters from any combination of inpatient, outpatient, or DMT claims within a one-calendar-year time period. This algorithm had the highest crude accuracy scores (84.3% to 90.1%) and the highest or second highest Youden J statistic (0.60 – 0.69) overall among the three datasets. Sensitivity and specificity ranged from 86% to 92% and 66% to 83%, respectively, across the datasets. Positive predictive value was greater than 95% in all three datasets and negative predictive value ranged from 41% to 57%. In an independent validation dataset of the Canadian Saskatchewan general population (individuals with or without an MS claim), sensitivity and specificity were 96% and 99%, respectively.

Considering the limitations of short ascertainment periods, Culpepper et al. quantified the difference in cumulative prevalence estimates for 2010 using 3-year (2008 – 2010) and 10-year (1999 – 2010) periods in the Veterans Health Administration (VA) and the Canadian Manitoba province (Manitoba) AHC datasets; the notion being that the cumulative prevalence found in a short 3-year ascertainment period could be used to approximate that of a longer 10-year period, which would in turn approximate a lifetime prevalence estimate. They found that the 10-year cumulative prevalence estimate was higher than the 3-year estimate by 37% (95% CI: 13%, 66%) and 47% (95% CI: 23%, 76%) in the VA and Manitoba datasets, respectively (Table 1).[7] They then supported these findings in a separate IQVIA™ (formerly Intercontinental Marketing Services) PharMetrics Plus Health Plan Claims Dataset (PharMetrics); they found the 9-year (2007 – 2015) cumulative prevalence estimate was higher than the 3-year (2013 – 2015) estimate by approximately 39% (95% CI: 13%, 71%). The authors additionally identified an annual growth rate in MS cumulative prevalence of 2.1% in the VA (from 2010 to 2014) and 2.5% in the PharMetrics (from 2010 to 2015) datasets.

Table 1. Cumulative Prevalence Estimates Using Different Ascertainment Periods.[7]

Dataset (year estimated)	Cumulative Prevalence (per 100,000 population)*		Relative Difference (95% CI)
	3-year	*10-year*	
VA (2010)	178	243	37% (13%-66%)
Manitoba (2010)	199	293	47% (23%-76%)
	3-year	*9-year*	
PharMetrics (2015)	155	215	39% (13%-71%)

* Approximated from figures provided in Culpepper et al.

Wallin et al. estimated the prevalence of MS in the entire 2017 US population using the cumulative prevalence method (Figure 1).[2] The authors estimated 3-year (2008 – 2010) cumulative prevalence

using private insurance data from Optum and Truven Health, Medicare and Medicaid data from the Centers for Medicare & Medicaid (CMS), and Veteran data from the VA.[2] They found the 3-year (2008 – 2010) cumulative prevalence estimate for the combined datasets to be 199.8/100,000, corresponding to 470,053 people with MS in the United States in 2010. To account for the limited ascertainment period, Wallin et al. used the 3-versus-10-year ascertainment period cumulative prevalence relative difference (3-vs-10-year difference) identified by Culpepper et al. to inflate their 3-year estimate into a 10-year estimate for 2010. Adjusting for demographics and the uninsured population, they increased the 3-year estimate by 37% as a lower bound and 47% as an upper bound to calculate a 10-year estimate of 265.1 (95% CI: 264.3, 265.8) to 309.2 (95% CI: 308.1, 310.1) per 100,000 population, corresponding to 623,437 and 727,344 persons, respectively. Applying Culpepper et al. annual cumulative prevalence growth rates, they estimated the 2017 US MS prevalence was between 337.9 and 362.6 per 100,000 population, corresponding to 851,749 and 913,925 persons, respectively. Importantly, these estimates approximately double previous estimates, and have been cited widely.[11-13]

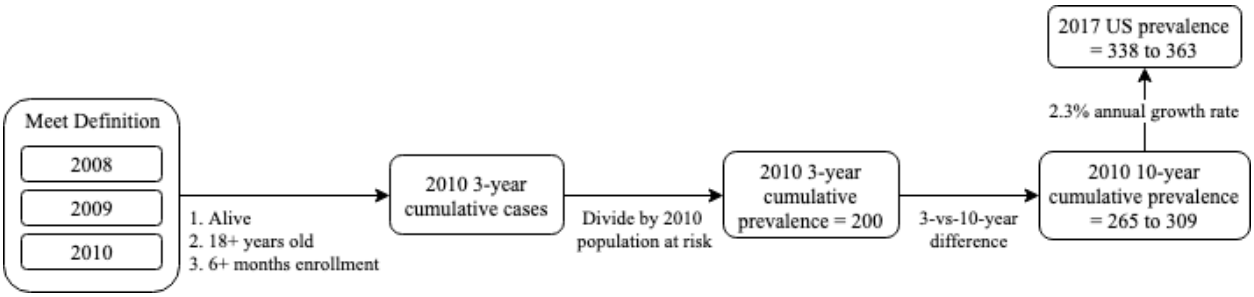

Figure 1. Wallin et al. Estimation of 2017 US Multiple Sclerosis Prevalence (per 100,000).

There are some concerns with the Wallin et al. estimate, however. While the Culpepper case identification algorithm appears to accurately identify MS cases, the adjustment methods used by Wallin et al. to inflate the 3-year ascertainment period cumulative prevalence to a 10-year estimate may not be appropriate. The 3-vs-10-year difference found by Culpepper et al. may be

overestimated, as the 39% difference identified using the PharMetrics dataset could not be reproduced in an independent analysis, which found a ~7% difference using the same years and dataset (E. Yang, personal communication, 2019).

While a difference between 3-year and 10-year ascertainment period cumulative prevalence estimates is expected because the longer period provides greater opportunity for cases to be detected, the significant discrepancy from the independent PharMetrics analysis is puzzling. Further investigation into the 3-vs-10-year difference is needed. The objective of this study was to independently quantify the difference between 3-year and 10-year ascertainment period cumulative prevalence using an alternate AHC dataset.

Methods:

Study design and setting

I conducted a retrospective claims database analysis using the IBM® MarketScan® Research Databases (formerly Truven®). I analyzed the MarketScan® Commercial Claims and Encounters (CCAE) and Medicare Supplemental and Coordination of Benefits (MDCR) Research Databases using claims data from January 2008 through December 2017. MarketScan® contains de-identified, person-specific records from approximately 350 payers covering over 250 million cumulative lives.[6] The databases link healthcare data from hospital, outpatient medical, and outpatient pharmaceutical claims for the population consisting of insured employees and their dependents, early retirees, Consolidated Omnibus Budget Reconciliation Act (COBRA) continuees, and Medicare-eligible retirees with employer-provided Medicare Supplemental plans

in the United States. Annual patient-level data are organized for Commercial and Medicare Supplemental databases in identical tables described in Appendix A.

Data Management

In MarketScan® each inpatient and outpatient claim include the patient-specific unique enrollee identification number, associated ICD-9 or ICD-10 diagnostic code (340 or G35 for MS), and date of the encounter. The inpatient admissions claims tables contain a principal diagnosis variable as well as 15 additional variables for any associated diagnoses contributing to the encounter. The outpatient services claims tables contain four variables for any associated diagnosis influencing the outpatient encounter. The outpatient pharmaceutical claims tables contain the National Drug Code (NDC) for the prescription dispensed. The variables I used and the tables they are found in are summarized in Appendix B.

I implemented the algorithm using SAS version 9.4 and performed Chi-square tests in the statistical package, R.

Multiple Sclerosis Case Identification Algorithm

To identify MS patients, I used the Culpepper case identification algorithm, which defines a MS case as an individual with three or more MS-related inpatient, outpatient, and/or disease-modifying therapy (DMT) encounters in any combination within a one-calendar-year time period. The diagnosis codes and DMT NDCs I used to identify inpatient, outpatient, and outpatient pharmaceutical encounters are outlined in Appendices D and E.

Inpatient Encounters

I identified inpatient encounters from the inpatient admissions claims tables as any overnight hospitalization with an ICD diagnosis code for MS included as the principal diagnosis for the admission. I collapsed inpatient encounters that overlapped or were consecutive within 24 hours of each other into a single unique encounter.

Outpatient Encounters

I identified outpatient encounters from the outpatient services claims tables as any outpatient visit containing an ICD code for MS in any of the four associated diagnoses variables (with the exception of the 2008 tables, which only contain 2 variables for diagnosis codes). I did not include outpatient encounters occurring on the same day as another, or within an inpatient admission period (between an inpatient encounter admission and discharge date) as unique outpatient encounters.

Disease-modifying Therapy Encounters

I identified DMT encounters from the outpatient pharmaceutical claims tables as any prescription claim for any DMT approved by the United States Food and Drug Administration through 2017: interferon betas 1b and 1a, peginterferon beta-1b, glatiramer acetate, natalizumab, alemtuzumab, daclizumab, ocrelizumab, fingolimod, and dimethyl fumarate. Disease-modifying therapies are those that reduce relapse frequency, delay progression of disability, or limit new disease activity as seen on magnetic resonance imaging (MRI). The MarketScan® RED BOOK™ Supplement contains drug information linked to the Outpatient Pharmaceutical tables by NDC. I used RED BOOK™ to identify all NDC numbers for all approved DMTs and found all prescription claims

containing any of the NDC numbers. I did not include duplicate claims occurring on the same date for the same NDC as unique DMT encounters. Additionally, I excluded natalizumab encounters if the enrollee also contained a diagnosis code for inflammatory bowel disease (IBD), for which the medication is also approved, in the inpatient admissions or outpatient services tables.

Culpepper et al. developed and validated the algorithm using DMTs approved through 2010. To isolate the effect of including DMTs approved in the US since 2010 in my analysis, I also applied the algorithm using only NDCs for DMTs approved through 2010: glatiramer acetate (COPAXONE®), interferon beta-1b (BETASERON®), interferon beta-1a (AVONEX® and REBIF®), and natalizumab (TYSABRI®).

Implementation of the Case Identification Algorithm

I reproduced the case identification methods of Wallin et al. as closely as possible to determine the appropriate 3-vs-10-year difference. Since the detailed programming code they used was not publicly available, I first ensured I was applying their case identification algorithm the same way they did. As part of the Wallin et al. analysis, they estimated cumulative prevalence for each year between 2008 and 2010 in MarketScan® and provided their 2010 3-year cumulative case results stratified by age group and sex. I compared my estimates from the same years to theirs.

I used Chi-square tests for heterogeneity using a significance level of 0.05 to compare cumulative case results between my analysis and that of Wallin et al. I compared proportions of total cases in each age group (18-24, 25-34, 45-54, 55-64, 65-74, and ≥75 years) overall and in each age group stratified by sex. My algorithm implementation steps are outlined in Appendix F.

Three-year and 10-year Cumulative Prevalence

After observing similar results to the Wallin et al. MarketScan® analysis, I used MarketScan® data from January 2008 through December 2017. My 3-year ascertainment period included cases from 2015 through 2017 and my 10-year ascertainment period was from 2008 through 2017. I estimated cumulative prevalence for each year since the first year of each ascertainment period.

To be included in the numerator, enrollees needed to be in the denominator of persons at risk, which was all enrollees at least 18 years of age with at least 6 months of active enrollment data in that year. For example, in a 3-year ascertainment period from 2008 to 2010, cumulative prevalence in 2008 was all persons meeting the case definition in 2008 divided by persons at risk in 2008, cumulative prevalence in 2009 was all persons meeting the case definition in 2008 or 2009 divided by persons at risk in 2009, and cumulative prevalence in 2010 was all persons meeting the case definition in 2008, 2009, or 2010 divided by persons at risk in 2010.

After calculating 3-year (2015 – 2017) and 10-year (2008 – 2017) cumulative prevalence for 2017, I calculated the 3-vs-10-year difference as the relative increase from the 3-year estimate to the 10-year estimate.

Results:

Implementation of the Case Identification Algorithm

My 3-year (2008 – 2010) cumulative case results stratified by age and sex are compared to those of Wallin et al. in Table 2. My cumulative case results yielded similar proportions among age groups overall and when stratified by sex. However, although the Chi-square test for heterogeneity indicated the case proportions were not statistically significantly different from the results of

Wallin et al., they identified 5,115 more cases than I did from the same database over the same time period. My annual cumulative prevalence estimates from 2008 to 2010 are compared to the Wallin et al. MarketScan® analysis in Figure 2.

Table 2. 2010 3-year Cumulative Cases by Age Group and Sex

	My Analysis			Wallin et al.		
Age Group	**Female**	**Male**	**Total**	**Female**	**Male**	**Total**
18-24	883	314	1,197	911	331	1,242
25-34	5,514	1,735	7,249	6,023	1,867	7,890
35-44	12,066	3,581	15,654	13,076	3,873	16,949
45-54	16,876	4,948	21,869	18,330	5,361	23,691
55-64	12,560	3,960	17,038	13,768	4,241	18,009
65-74	2,555	938	3,539	2,848	1,006	3,854
75+	848	334	1,182	894	313	1,207
Total	51,302	15,810	67,728	55,851	16,992	72,843
p-value	0.94	0.76	0.34	Reference	Reference	Reference

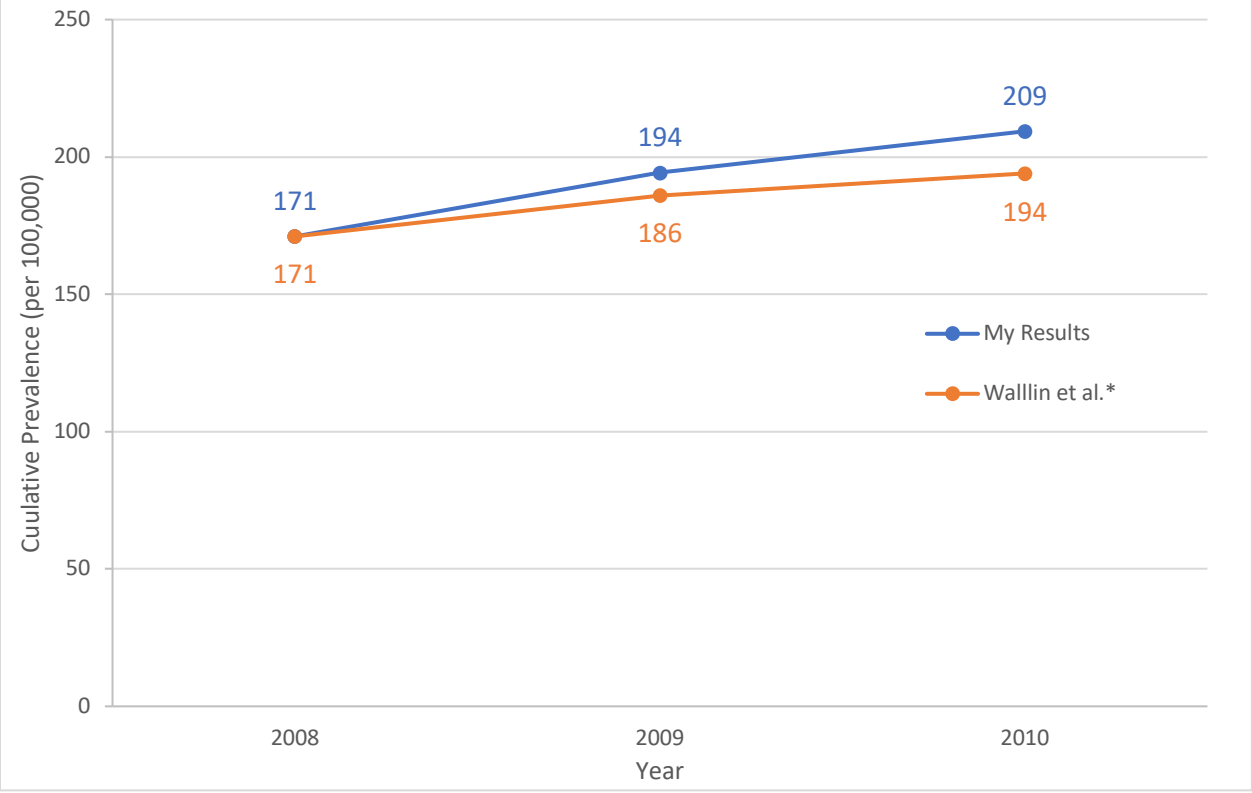

Figure 2. Cumulative Prevalence Estimates During Ascertainment Period 2008 to 2010
*Wallin et al. data was estimated from a figure provided in their publication

Three-year and 10-year Cumulative Prevalence

Cases were identified for each year from 2008 to 2017 to develop a 10-year cumulative prevalence estimate; and for each year from 2015 to 2017 to develop a 3-year cumulative prevalence estimate. The annual cumulative prevalence estimates during the 10-year and 3-year ascertainment periods are displayed in Figure 3. The 2017 10-year cumulative prevalence (250.5/100,000) was 7.6% higher than the 3-year (232.7/100,000).

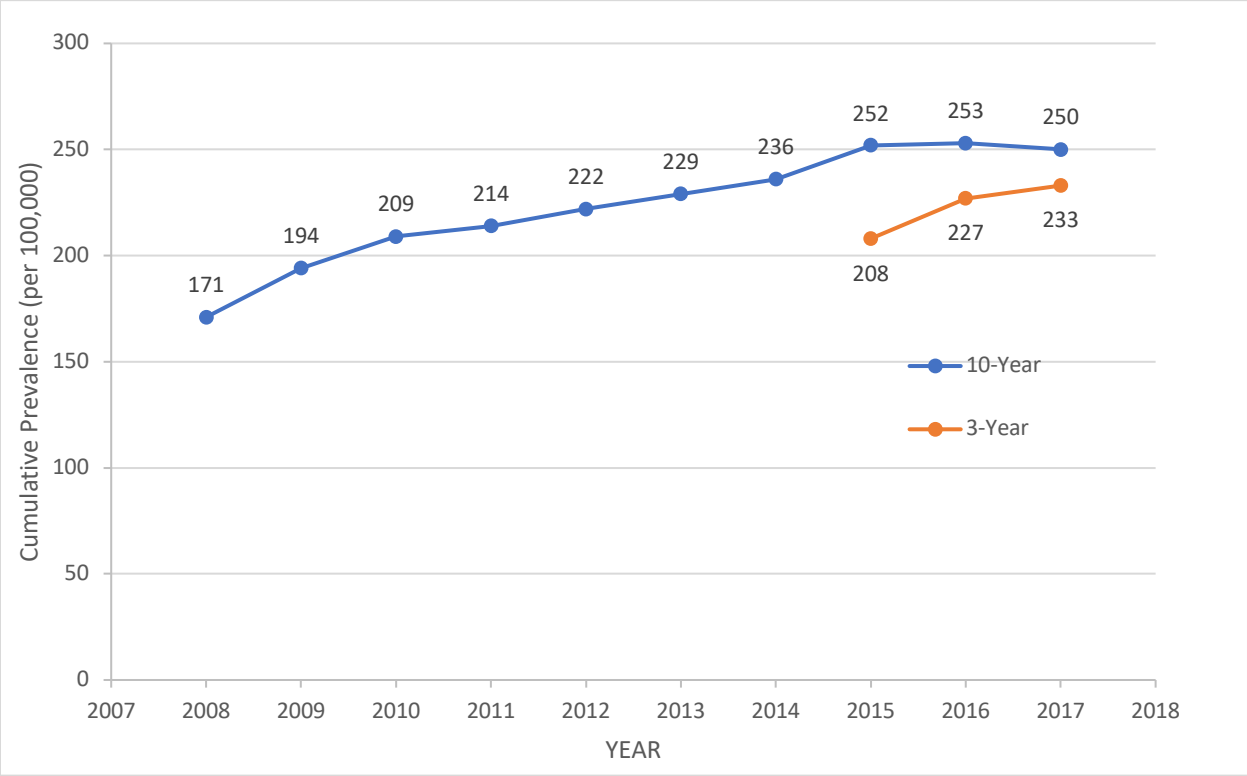

Figure 3. 2017 10-year (2008 – 2017) and 3-year (2015 – 2017) Annual Cumulative Prevalence Estimates from MarketScan® Data (Using DMTs approved through 2017).

When I only included DMTs approved in the US through 2010 (Appendix E) in the algorithm, the 3-year (2008 – 2010) annual cumulative prevalence results were unchanged. However, compared with the inclusion of DMTs approved through 2017, the 2017 10-year and 3-year ascertainment

periods identified 1,902 and 2,701 fewer cases, respectively. This resulted in a 10-year 2017 cumulative prevalence (240.2/100,000) that 10.1% was greater than the 3-year (218.2/100,000) (Figure 4).

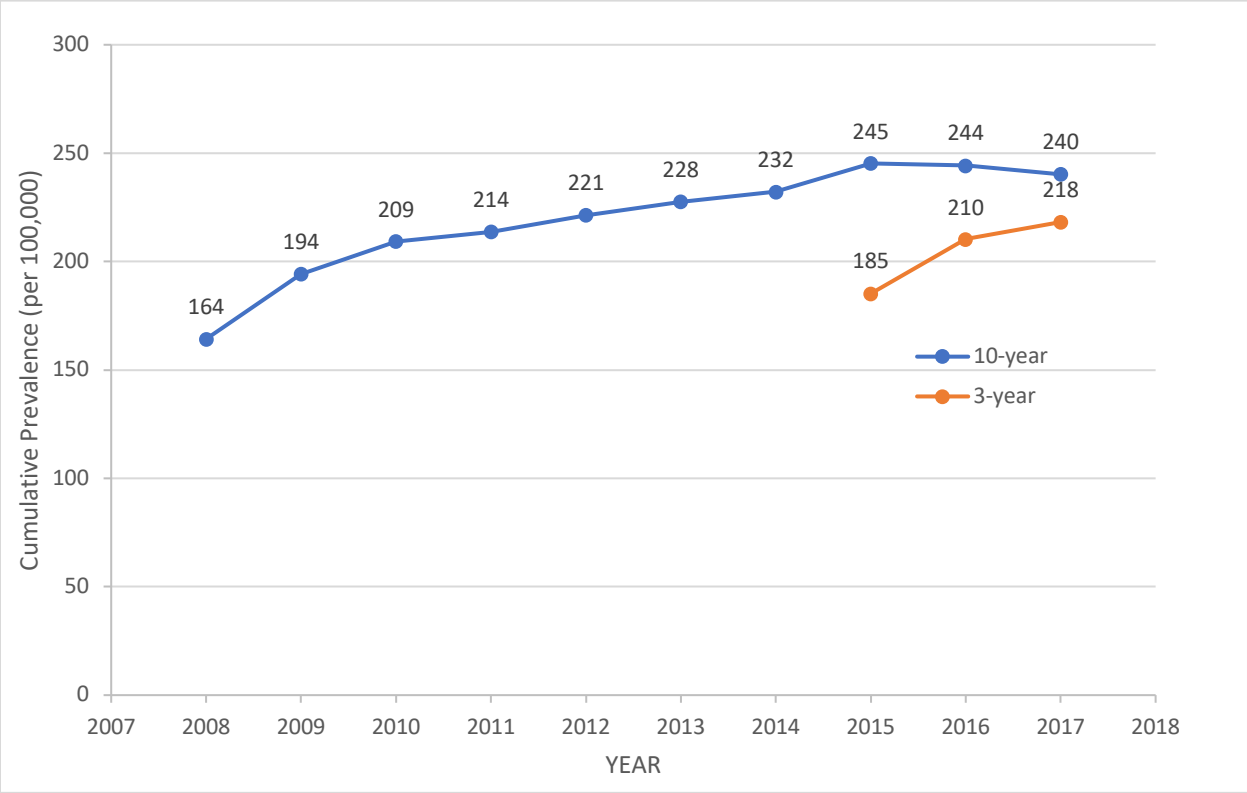

Figure 4. 2017 10-year (2008 – 2017) and 3-year (2015 – 2017) Annual Cumulative Prevalence Estimates from MarketScan® Data (Using DMTs approved through 2010).

Discussion:

I estimated the relative difference in cumulative prevalence of Multiple Sclerosis comparing a 3-year (2015 – 2017) versus a 10-year (2008 – 2017) ascertainment period in a segment of the United States privately insured population from the MarketScan® administrative health claims dataset. By applying a previously validated MS case identification algorithm,[7] I found that the 10-year ascertainment period estimate was 7.6% higher than the 3-year estimate.

My findings are notably different than those reported by Culpepper et al.[7] (37% to 47%), but consistent with the approximately 7% difference identified by another independent researcher using the PharMetrics dataset (E. Yang, personal communication, 2019). Using the Culpepper et al. 3-vs-10-year difference, Wallin et al. estimated 2010 cumulative prevalence to be at least 265/100,000, leading to a 2017 prevalence of at least 338/100,000. However, if the difference is in fact closer to that found independently in both Yang's and in my analysis, these figures are a markedly overestimated. It is also much higher than other recent estimates. Campbell and colleagues reviewed data between 1998 and 2009 from the Medical Expenditure Panel Survey (MEPS); a nationally representative large-scale database linking direct cost, productivity, and other health-related information. Multiple Sclerosis cases were identified via the survey and verified as having a single MS ICD code. They estimated an annual prevalence of 210/100,000.[4] Although survey data has its own limitations, this estimate is not subject to the same shortcomings of limited ascertainment periods in AHC data. In a 2010 global review of Multiple Sclerosis economic impact, the World Health Organization and the Multiple Sclerosis International Federation estimated a US prevalence of 135/100,000; however, this estimate did not use strict epidemiological methods and relied instead on scientific literature review.[5] Another 2016 MS study using PharMetrics 2008 to 2012 AHC data estimated annual point prevalence in the US commercially insured population to be 149/100,000. Cases were identified by 2 inpatient ICD codes, 3 outpatient ICD codes, or 1 DMT claim. However, this point prevalence estimate does not account for the underestimation issue with short ascertainment periods.[3] A summary of these recent estimates for US MS prevalence is available in Appendix G.

There are several potential explanations for the discrepancy in 3-vs-10-year difference results between my analysis and that of Culpepper et al. One possibility is due to the different time periods we assessed. My 3-year and 10-year estimates were compared for 2017, whereas those of Culpepper et al. were for 2010. Between 2010 and 2017, changes in the documentation and management of MS have occurred including the implementation of ICD-10 coding and the advent of oral DMT therapies; however, the transition from ICD-9 to ICD-10 would not be expected to have a large impact because in both systems the coding for MS is a single, 3-character code. Changes in the incidence, survival, and distribution of MS in the US population as well as changes in clinical landscape (i.e. diagnosis, disease-management, etc.) could affect the consistency of the case identification algorithm performance between the two time periods as well.

Considering the potential effects of including DMTs approved since 2010, I also conducted the analysis including only DMTs approved in the US through 2010, as it was validated by Culpepper et al. The 3-year 2010 annual cumulative prevalence results were the same as my original analysis, but the 10-year 2017 cumulative prevalence was greater than the 3-year by 10.1%. In this case, incorporating changes in the treatment landscape of MS since 2010 (i.e. adding more recently approved DMTs) did slightly improve the ability of the 3-year ascertainment period to approximate prevalence using a 10-year ascertainment period, but the difference between the two was not similar to Culpepper's results. Interpretation of this scenario analysis is somewhat limited because patients on newer DMT's who would have otherwise been on older DMTs are not identified using drug criteria.

Another factor that could influence the difference between our results is the different datasets (and thus populations) we evaluated. Culpepper et al. found a 47% difference in the Manitoba dataset, which comes from a different country utilizing a different, single government payer healthcare system. Canada has documented significant increases in prevalence from 1990 to 2010, which may not be comparable to the United States.[14,15] The 37% difference identified in the VA dataset by Culpepper et al. also comes from a distinct population and healthcare system that similarly may not be comparable to my MarketScan® data. Wallin et al. applied these differences found by Culpepper et al. to estimate the US national prevalence from a 3-year ascertainment period using a heterogenous mixture of datasets, but it is not clear how the difference would vary among the datasets used.

An inaccurate picture of the burden of Multiple Sclerosis in the United States could have consequences. Policymakers and health care decision makers could put a disproportionate emphasis on the disease, causing public and private resources to be shifted from other programs in order to conduct research in the disease and its treatment. Clinically, prevalence estimates provide a context for diagnostic decision-making.[16] A perception of higher MS prevalence also could cause physicians to diagnose MS more readily, particularly when patients have atypical presentations or lack specific MRI findings. Patients with similar neurological conditions may be inappropriately initiated on DMT therapy, which can be costly with significant side-effect profiles.

My study has several limitations. First, the specific programming code used by Culpepper et al. or Wallin et al. was not available, so I was not able to ensure I operationalized the case identification algorithm exactly as they had. Similarly, more granular results were not provided by those authors

that would allow me to compare where my cohorts may have deviated from theirs. The only way to compare if my approach was similar to theirs was to compare my 3-year (2008 – 2010) cumulative prevalence results in MarketScan® to those provided by Wallin et al. Although the Chi-square test for heterogeneity indicated the proportion of cases in each age/sex group were not significantly different, Wallin et al. identified 5,115 more cases than I did, indicating an unidentifiable difference between our approaches. Since our 2008 prevalence estimates are identical (171/100,000), our case accumulation methods are possibly dissimilar. Second, I was only able to make a single estimation of the 3-vs-10-year difference, as I was limited to 10 years of data from a single dataset. More data points from my analysis illustrating the difference, and in various years, would provide stronger evidence regarding its nature, particularly because Culpepper's are the only published estimates of this metric in Multiple Sclerosis. Third, the case identification algorithm was not developed or validated for the time period I analyzed, so my 3-vs-10-year difference could be influenced by changes in coding practices, available DMTs, epidemiology, and management of MS.

My objective was to validate the difference in cumulative prevalence comparing 3-year versus 10-year ascertainment periods, as this was a major factor in the calculation of a United States MS prevalence estimate by Wallin et al. I applied the same algorithm used by them in an independent dataset from 2008 through 2017 and found this difference to be 7.6% – considerably smaller than the 37% and 47% estimates found by Culpepper et al. It is not clear what the appropriate 3-vs-10-year difference is, how it changes over time, or why my estimate is starkly dissimilar from the Culpepper et al. estimate However, the discrepancy does not appear to be related to use of newer

DMTs, and the different contexts of our analyses do not likely account for the 30 to 40 percentage point higher difference estimated by Culpepper et al.

The lower estimate of the 3-vs-10-year difference, identified in my analysis and independently by Yang using PharMetrics data, indicates that the contemporary MS prevalence estimated by Wallin et al. may be high, and that further research is needed before their results are considered in influencing policy and clinical decision-making as it relates to MS. Future work should independently investigate this difference in these and alternative datasets and years. Additionally, future studies should validate the algorithm in more recent years with medical records from a population more representative of the United States general population. Using these extrapolation methods to estimate prevalence should be pursued cautiously until stronger evidence is developed.

Appendix A. MarketScan® Research Databases Commercial and Medicare Supplemental Tables Utilized

Table	Contents
Inpatient Admissions Table (I)	Records that summarize information about a hospital admission
Outpatient Services Table (O)	Encounters and claims for services rendered in a doctor's office, hospital outpatient facility, emergency room, or other outpatient facility
Outpatient Pharmaceutical Claims Table (D)	Outpatient pharmaceutical claims
Annual Enrollment Summary Table (A)	Demographic and plan information on users and non-users of services. Single record per-person, per-year

Appendix B. MarketScan® Research Databases Variables Utilized

Variable	Variable Name	Table	Notes
ADMDATE	Date of Admission	I	
DISDATE	Date of Discharge	I	
DOBYR	Patient Birth Year	A	
DX1-DX15	Diagnosis 1 through 15	I, O	15 in I table, 4 in O table (2 in 2008 O table)
ENROLID	Enrollee ID	I, O, A, D	Unique patient identifier
MEMDAYS	Member Days	A	Number of days an enrollee had active coverage
NDCNUM	National Drug Code	D	Full 11 digits of FDA registered number
PDX	Principal Diagnosis	I	The main reason for admission
SEX	Sex of patient	A	
SVCDATE	Date Service Incurred	O, D	Outpatient service or prescription fill

Appendix C. Diagnosis Codes Utilized in Algorithm

	ICD-9	ICD-10
Multiple Sclerosis	340	G35
Inflammatory Bowel Disease	555*, 556*	K50, K51, K52

Appendix D. Multiple Sclerosis Disease-modifying Therapies

Generic name	Brand Name	NDC	ROA
Alemtuzumab	LEMTRADA	58468020001	IV
Daclizumab	ZINBRYTA	74003301	SC
Dimethyl Fumarate	DIMETHYL FUMARATE	51927292900	NA
Dimethyl Fumarate	TECFIDERA	64406000501	PO
Dimethyl Fumarate	TECFIDERA	64406000602	PO
Dimethyl Fumarate	TECFIDERA STARTER PACK	64406000703	PO
Interferon Beta-1b	EXTAVIA	78056999	MR
Fingolimod HCL	GILENYA	78060715	PO
Fingolimod HCL	GILENYA	78060751	PO
Fingolimod HCL	GILENYA	78060789	PO
Fingolimod HCL	GILENYA	78096589	PO
Glatiramer Acetate	COPAXONE	88115003	SC
Glatiramer Acetate	COPAXONE	88115330	MR
Glatiramer Acetate	COPAXONE	68115075030	MR
Glatiramer Acetate	COPAXONE	68546031730	SC
Glatiramer Acetate	COPAXONE	68546032512	SC
Glatiramer Acetate	GLATIRAMER ACETATE	378696032	SC
Glatiramer Acetate	GLATIRAMER ACETATE	378696093	SC
Glatiramer Acetate	GLATIRAMER ACETATE	378696112	SC
Glatiramer Acetate	GLATIRAMER ACETATE	378696132	SC
Glatiramer Acetate	GLATOPA	781323434	SC
Glatiramer Acetate	GLATOPA	781325089	SC
Interferon Beta-1A	AVONEX	54569443300	IM
Interferon Beta-1A	AVONEX	59627000103	IM
Interferon Beta-1A	AVONEX	59627000104	IM
Interferon Beta-1A	AVONEX	59627000205	MR
Interferon Beta-1A	AVONEX	59627000207	MR
Interferon Beta-1A	AVONEX	59627011103	IM
Interferon Beta-1A	AVONEX	59627022205	MR
Interferon Beta-1A	AVONEX PEN	59627000301	MR
Interferon Beta-1A	AVONEX PEN	59627000304	MR
Interferon Beta-1A	AVONEX PEN	59627033304	MR
Interferon Beta-1A	REBIF	44087002203	SC
Interferon Beta-1A	REBIF	44087004403	SC
Interferon Beta-1A	REBIF REBIDOSE	44087332201	SC
Interferon Beta-1A	REBIF REBIDOSE	44087334401	SC

Appendix D (continued). Multiple Sclerosis Disease-modifying Therapies

Generic name	Brand Name	NDC	ROA
Interferon Beta-1a	REBIF	44087882201	SC
Interferon Beta-1a	REBIF REBIDOSE TITRATION PACK	44087018801	SC
Interferon Beta-1B	BETASERON	50419052103	SC
Interferon Beta-1B	BETASERON	50419052115	SC
Interferon Beta-1B	BETASERON	50419052309	SC
Interferon Beta-1B	BETASERON	50419052315	SC
Interferon Beta-1B	BETASERON	50419052335	SC
Interferon Beta-1B	BETASERON	50419052401	SC
Interferon Beta-1B	BETASERON	50419052435	SC
Interferon Beta-1B	EXTAVIA	78056912	SC
Interferon Beta-1B	EXTAVIA	78056961	SC
Interferon Beta-1b	BETASERON	50419052325	MR
Natalizumab	TYSABRI	59075073015	IV
Natalizumab	TYSABRI	64406000801	IV
Ocrelizumab	OCREVUS	50242015001	IV
Peginterferon Beta-1a	PLEGRIDY	64406001501	SC
Peginterferon Beta-1a	PLEGRIDY PEN	64406001101	SC
Peginterferon Beta-1a	PLEGRIDY PEN STARTER PACK	64406001201	SC
Peginterferon Beta-1a	PLEGRIDY STARTER PACK	64406001601	SC
Teriflunomide	AUBAGIO	58468021002	PO
Teriflunomide	AUBAGIO	58468021101	PO

ROA, Route of Administration; IV, Intravenous; SC, subcutaneous; PO, oral; MR, multiple routes

Appendix E. Multiple Sclerosis Disease-modifying Therapies Used by Culpepper et al.

Generic name	Brand Name	NDC
Glatiramer Acetate	COPAXONE	88115330
Glatiramer Acetate	COPAXONE	88115300
Glatiramer Acetate	COPAXONE	669141009*1
Glatiramer Acetate	COPAXONE	10130*12684
Glatiramer Acetate	COPAXONE	480614000
Glatiramer Acetate	COPAXONE	480614056
Interferon Beta-1B	BETASERON	50419052103
Interferon Beta-1B	BETASERON	50419052103
Interferon Beta-1B	BETASERON	50419052115
Interferon Beta-1B	BETASERON	50419*52304
Interferon Beta-1B	BETASERON	50419*52335
Interferon Beta-1B	BETASERON	50419*52309
Interferon Beta-1B	BETASERON	50419*52325
Interferon Beta-1B	BETASERON	50419*52301
Interferon Beta-1B	BETASERON	50419*52315
Interferon Beta-1B	BETASERON	53905*52325
Interferon Beta-1B	BETASERON	53905*52308
Interferon Beta-1B	BETASERON	53905*52301
Interferon Beta-1B	BETASERON	53905*52315
Interferon Beta-1A	AVONEX	59627*00103
Interferon Beta-1A	AVONEX	59627*00205
Interferon Beta-1A	AVONEX	64370*00100
Interferon Beta-1A	AVONEX	62195*601
Interferon Beta-1A	AVONEX	54569443300
Interferon Beta-1A	REBIF	44087882201
Interferon Beta-1A	REBIF	44087002201
Interferon Beta-1A	REBIF	44087002203
Interferon Beta-1A	REBIF	44087004401
Interferon Beta-1A	REBIF	44087004403
Interferon Beta-1A	REBIF	522030022*3
Interferon Beta-1A	REBIF	52203002201
Interferon Beta-1A	REBIF	66303004403
Interferon Beta-1A	REBIF	66303004401
Interferon Beta-1A	REBIF	66303002203
Interferon Beta-1A	REBIF	62195*70503
Interferon Beta-1A	REBIF	62195*70403

Appendix E (continued). Multiple Sclerosis Disease-modifying Therapies Used by Culpepper et al.

Generic name	Brand Name	NDC
Natalizumab	TYSABRI	59075*73015
Natalizumab	TYSABRI	64406*73001
Natalizumab	TYSABRI	64406*73002
Natalizumab	TYSABRI	409980015

Appendix F. Step-by-Step Case Identification and Cumulative Prevalence Calculation

Algorithm Implementation for case identification

I conducted the following steps in each year for both Medicare and Commercial databases
Note: all patients with missing enrollee IDs are removed from dataset prior to analysis

Step 1: Identify enrollees with both Medicare Supplemental and Commercial (dual coverage) to exclude

1. Identify enrollee identification numbers occurring in both Commercial and Medicare Supplemental database enrollment tables for each year

Step 2: Identify total unique MS inpatient (IP) encounters for each patient

1. Identify all IP encounters including a MS ICD code as the principal diagnosis from the inpatient admissions table (I)
    a. Collapse all encounters with overlapping service dates (i.e. admission date through discharge date) or occurring consecutively within 24 hours into a single, unique encounter
        i. Remove duplicate encounters (observations containing the same enrollee ID, admission date and discharge date)
2. Remove non-overnight admission records (i.e. records containing the same enrollee ID, admission date, and discharge date)
3. Sum unique IP encounters for each enrollee

Step 3: Identify total unique MS outpatient (OP) encounters for each patient

1. Identify all OP encounters including with an associated MS ICD diagnosis code (variables DX1 through DX4) from the outpatient services table (O)
    a. Remove duplicate encounters (records containing the same enrollee ID and service date)
2. Remove any OP encounters for each enrollee that occur within an IP encounter (OP service date between IP admission date and discharge date)
3. Sum unique OP encounters for each enrollee

Step 4: Identify total unique MS disease modifying therapy dispensation (DMT) encounters for each patient

1. Identify all DMT records including an NDC number for any MS drug approved DMT through 2017 from the outpatient pharmaceutical services table (D)
    a. Remove duplicate encounters (records containing the same enrollee ID and service date)
2. Remove natalizumab records for enrollees who have an associated diagnosis code for inflammatory bowel disease (IBD) in the inpatient admissions (I) and outpatient services (O) tables
3. Sum unique DMT encounters for each enrollee

Step 5: Identify patients meeting the case definition ([IP+OP+DMT] $\geq$ 3)

1. For each enrollee, sum the total IP, OP, and DMT encounters
2. Remove enrollee records with less than 3 total encounters

Step 6: Remove enrollees with dual coverage identified from Step 1

Appendix F (continued). Step-by-Step Case Identification and Cumulative Prevalence Calculation

Determining Cumulative Prevalence

Obtain cumulative numerators for each year

1. Merge the numerator files from each year up to and including the final year of the ascertainment period
    a. Remove duplicate enrollees
2. Remove enrollees who are not in the denominator file for the ascertainment year

Appendix G. United States Multiple Sclerosis Prevalence Estimates (per 100,000 population) in Recent Studies

Prevalence	Year	Investigator	Data Source	Method
265 – 309	2010	Wallin	Multiple AHC datasets	3-year cumulative prevalence extrapolated to 10-year cumulative prevalence
210	1998 – 2009	Campbell	MEPS	Verified reported MS cases with AHC data; standardized to US population for annual prevalence
135	2010	Trisolini	Published literature	Scientific literature review
149	2010	Dilokthornsakul	PharMetrics	Point prevalence in single AHC dataset

*AHC, Administrative Health Claims; MEPS, Medical Expenditure Panel Survey

REFERENCES:

1. Evans C, Beland SG, Kulaga S, et al. Incidence and prevalence of multiple sclerosis in the americas: A systematic review. *Neuroepidemiology*. 2013;40(3):195-210. doi: 10.1159/000342779 [doi].

2. Wallin MT, Culpepper WJ, Campbell JD, et al. The prevalence of MS in the united states: A population-based estimate using health claims data. *Neurology*. 2019;92(10):e1040. doi: 10.1212/WNL.0000000000007035 [doi].

3. Dilokthornsakul P, Valuck RJ, Nair KV, Corboy JR, Allen RR, Campbell JD. Multiple sclerosis prevalence in the united states commercially insured population. *Neurology*. 2016;86(11):1014-1021. doi: 10.1212/WNL.0000000000002469 [doi].

4. Campbell JD, Ghushchyan V, Brett McQueen R, et al. Burden of multiple sclerosis on direct, indirect costs and quality of life: National US estimates. *Mult Scler Relat Disord*. 2014;3(2):227-236. doi: 10.1016/j.msard.2013.09.004 [doi].

5. Trisolini M, Honeycutt A, Wiener J, Lesesne S. Global economic impact of multiple sclerosis. *Multiple Sclerosis International Federation*. 2010.

6. IBM MarketScan research databases - overview - united states. https://www.ibm.com/us-en/marketplace/marketscan-research-databases. Updated 2019. Accessed May 10, 2019.

7. Culpepper WJ, Marrie RA, Langer-Gould A, et al. Validation of an algorithm for identifying MS cases in administrative health claims datasets. *Neurology*. 2019;92(10):e1028. doi: 10.1212/WNL.0000000000007043 [doi].

8. Nelson LM, Wallin MT, Marrie RA, et al. A new way to estimate neurologic disease prevalence in the united states: Illustrated with MS. *Neurology*. 2019;92(10):469-480. doi: 10.1212/WNL.0000000000007044 [doi].

9. Capocaccia R, Colonna M, Corazziari I, et al. Measuring cancer prevalence in europe: The EUROPREVAL project. *Ann Oncol*. 2002;13(6):831-839.

10. Ng R, Bernatsky S, Rahme E. Observation period effects on estimation of systemic lupus erythematosus incidence and prevalence in quebec. *J Rheumatol*. 2013;40(8):1334-1336. doi: 10.3899/jrheum.121215 [doi].

11. Luxner L. New NMSS study confirms nearly 1 million americans live with MS. . 2019. https://multiplesclerosisnewstoday.com/2019/02/15/newly-published-nmss-study-confirms-nearly-1-million-americans-live-with-ms/. Accessed Apr 13, 2019.

12. MS prevalence. National Multiple Sclerosis Society Web site. http://www.nationalmssociety.org/About-the-Society/MS-Prevalence. Accessed Apr 13, 2019.

13. Koskie B. Multiple sclerosis: Facts, statistics, and you. Healthline Web site. https://www.healthline.com/health/multiple-sclerosis/facts-statistics-infographic. Updated 2015. Accessed Apr 13, 2019.

14. Marrie RA, Fisk J, Stadnyk K, Yu BN. The incidence and prevalence of multiple sclerosis in nova scotia, canada. *Canadian Journal of Neurological Sciences*. 2013;40(6):824-831.

15. Kingwell E, Zhu F, Marrie RA, et al. High incidence and increasing prevalence of multiple sclerosis in british columbia, canada: Findings from over two decades (1991-2010). *J Neurol*. 2015;262(10):2352-2363. doi: 10.1007/s00415-015-7842-0 [doi].

16. Ward MM. Estimating disease prevalence and incidence using administrative data: Some assembly required. *J Rheumatol*. 2013;40(8):1241-1243. doi: 10.3899/jrheum.130675 [doi].