Published in final edited form as:

Med Care. 2017 August ; 55(8): 782–788. doi:10.1097/MLR.0000000000000753.

Using Self-Reports or Claims to Assess Disease Prevalence: It's Complicated

Patricia St.Clair, Sc.B., Étienne Gaudette, Ph.D., Henu Zhao, Ph.D., Bryan Tysinger, M.Sc., Roxanna Seyedin, M.Ph., and Dana P. Goldman, Ph.D.

Schaeffer Center for Health Policy and Economics, University of Southern California Price School and School of Pharmacy, 635 Downey Way, Suite 210, Los Angeles, CA 90089

Abstract

Background—Two common ways of measuring disease prevalence include 1) using self-reported disease diagnosis from survey responses; 2) using disease-specific diagnosis codes found in administrative data. Because they don't suffer from self-report biases, claims are often assumed to be more objective. However, it is not clear that claims always produce better prevalence estimates.

Objective—Conduct an assessment of discrepancies between self-report and claims-based measures for two diseases in the US elderly to investigate definition, selection, and measurement error issues which may help explain divergence between claims and self-report estimates of prevalence.

Data—Self-reported data from three sources are included: the Health and Retirement Study (HRS), the Medicare Current Beneficiary Survey (MCBS), and the National Health and Nutrition Examination Survey. Claims-based disease measurements are provided from Medicare claims linked to HRS and MCBS participants, comprehensive claims data from a 20% random sample of Medicare enrollees, and private health insurance claims from Humana Inc.

Methods—Prevalence of diagnosed disease in the US elderly are computed and compared across sources. Two medical conditions are considered: diabetes and heart attack.

Results—Comparisons of diagnosed diabetes and heart attack prevalence show similar trends by source, but claims differ from self-reports with regard to levels. Selection into insurance plans, disease definitions, and the reference period used by algorithms are identified as sources contributing to differences.

Conclusions—Claims and self-reports both have strengths and weaknesses, which researchers need to consider when interpreting estimates of prevalence from these two sources.

Corresponding Author: Étienne Gaudette, Ph.D. Phone : 213-821-6179; fax : 213-740-3460; egaudett@usc.edu.

Patricia St.Clair, Sc.B. : Phone : 213-821-6479; fax : 213-740-3460

Étienne Gaudette, Ph.D. : Phone : 213-821-6179; fax : 213-740-3460

Henu Zhao, Ph.D. : Phone : 213-821-8876; fax : 213-740-3460

Bryan Tysinger, M.Sc. : Phone : 213-821-7964; fax : 213-740-3460

Roxanna Seyedin, M.Ph. : Phone : 310-612-5768; fax : 213-740-3460

Dana P. Goldman, Ph.D. : Phone : 213-821-7948; fax : 213-740-3460

Declaration of Interests: We declare that there are no conflicts of interests.

A preliminary version of this research was presented at the 50th Annual Conference of the Canadian Economics Association on June 3, 2016.

1 Introduction

Accurate tracking of disease dynamics across populations is essential to understand population health and efficiently allocate medical resources. Estimates of disease prevalence may be drawn from many sources. Two relatively convenient secondary data sources for these estimates include nationally representative survey data with questions asking “Has a doctor ever told you...” and administrative data with diagnosis codes, such as Medicare claims. Previous comparisons between disease measures in claims and self-reports have shown differences, and authors have described individual characteristics that are associated with discordance, but causes of variations in overall prevalence estimates have not been widely explored. Because they don't suffer from self-report biases, claims are often considered “the truth” in concordance studies, i.e., discrepancies between self-reports and claims are interpreted as misreporting by self-reports. [1-4] However, it is not clear that claims produce better prevalence estimates.

This paper provides an assessment of discrepancies between self-reports and claims-based prevalence estimates for two diseases in the US elderly, one chronic (diabetes) and one acute (heart attack). We use self-report data from three sources, namely the Health and Retirement Study (HRS), the Medicare Current Beneficiary Survey (MCBS), and the National Health and Nutrition Examination Survey (NHANES). Claims-based estimates are provided from Medicare claims linked with HRS and MCBS participants, comprehensive claims data from a 20% random sample of Medicare beneficiaries, and private health insurance claims from Humana Inc.

Our work expands upon the literature that compares self-report of diseases and claims-based data [1, 4-13]. An important development for that literature occurred in 2003, when the Center for Medicare and Medicaid Services introduced a set of algorithms to identify disease through the Chronic Condition Data Warehouse (CCW) in an effort to more readily and accurately identify beneficiaries with particular conditions. The validated CCW algorithms are based primarily on International Classification of Diseases (ICD-9 or ICD-10) diagnoses codes within a given reference period.

Several studies compared individual concordance between self-reports of various conditions with linked Medicare claims-based diagnosis, and found considerable disagreement between claims and self-reports [5-7]. Other studies have documented disease prevalence, in particular diabetes. A finding common to the HRS, NHANES, the National Health Interview Survey, and the Medicaid Behavioral Risk Factor Survey is that they all display lower diabetes prevalence than those found in claims data [5, 7, 14, 15]. However, it is yet unclear whether this discrepancy is a product of too many individuals identified by claims as having diabetes (false positives in claims), or too few individuals self-reporting as having been diagnosed with diabetes (false negatives in surveys).

This paper compares disease prevalence across data sources and their subsamples to investigate multiple potential sources of divergence between claims and self-report. These may be grouped in three broad topics: definition issues, selection issues, and measurement error issues.

2 Methods

2.1 Data sources

We use five data sources to produce estimates of diabetes and heart attack prevalence: 1) HRS; 2) MCBS; 3) NHANES; 4) a 20% random sample of Medicare beneficiaries and their claims; and 5) claims and enrollment from a private health insurer (Humana). Because the algorithm we use to identify diabetes in claims requires individuals to be continuously observed for two years, we limit our analysis to those 67 or older and enrolled in Medicare in all the data sources. Because Medicare claims are incomplete for beneficiaries enrolled in Medicare Advantage (MA), we categorize survey respondent enrollment as traditional fee-for-service Medicare (FFS) or MA to compare to estimates derived from FFS or MA claims. MA plans are a type of Medicare health plan offered by private companies such as Humana. [16].

Table 1 shows descriptive statistics in 2008, the last year for which we have data across all sources. The years used in analysis are presented in column headers for each source. Besides differences across enrollment type, which will be examined later, sources are similar with regard to their demographics and disease prevalence profiles.

The nationally representative HRS, MCBS, and NHANES surveys include questions on diagnosed diabetes and heart attacks, providing the self-reported measures. The HRS and MCBS samples have been linked to Medicare claims, allowing claims-based estimates of prevalence for those enrolled in FFS Medicare. We note the potential bias in prevalence estimates from using HRS-linked claims which originates from the decision of respondents to allow researchers to link their responses with Medicare claims. Because of this non-random selection, HRS-linked claims may not be representative of the fee-for-service population. Please see Supplemental Digital Content 3 for further discussion. MCBS has Medicare records for all respondents so for that data source this is not an issue. We show weighted results in all analysis of survey and linked claims data, unless otherwise stated. The 20% Medicare sample is used to estimate prevalence in a large sample of traditional Medicare beneficiaries. Humana claims are used to estimate prevalence in the MA population, with the caveat that they are only representative of the population covered by this insurer.

2.2 Survey Questions for Diabetes and Heart Attack

To measure self-reported prevalence of diagnosed diabetes and heart attack, we use measures constructed from responses to questions that ask whether a doctor has told the individual that they have diabetes or had a heart attack. Question wording and methodological considerations specific to each surveys are described in the Supplemental Digital Content 3.

2.3 Methods to Develop Measures using Claims Data

Sample selection criteria in the claims data include continuous enrollment during the measurement period, so that we observe a complete set of their claims. We apply CCW algorithms to claims data for diagnosed diabetes and acute myocardial infarction (AMI).

CCW algorithms scan particular claim types for a list of diagnosis codes within a reference period of one to three years. We also identify acute coronary syndrome (ACS), a broader diagnostic category which includes AMI and other conditions with diminished cardiac blood supply, from claims data. We do so following Yasaitis et al. [6], who indicated that clinical presentations for ACS “may be undistinguishable from AMI”, so that ACS patients may reasonably report a heart attack, and found that this interpretation improved agreement between self-reported and claims-derived measures. Diagnosis codes, reference periods and methodological considerations specific to each dataset are described in the Supplemental Digital Content 3.

2.4 Analyses

We compare disease prevalence across data sources and their subsamples to investigate potential sources of divergence between claims and self-reports. We first compute high-level overviews of diabetes and heart disease prevalence across all sources. We then investigate sources of divergence, which may be grouped in three broad topics: definition issues, selection issues, and measurement error issues.

Regarding definition issues, we first study how self-reports of heart attacks align with AMI and ACS detected in claims. This aims to replicate in population prevalence a result from Yasaitis et al. [6]. Their study documented a low agreement between individual self-reports and AMI claims, but found it improved when using the broader ACS definition to compare with self-reports. Second, we ensure that the claims algorithm reference period, i.e. the amount of time over which the algorithm scans claims to identify disease diagnoses, is also applied to survey data by using dates of self-reported heart attacks.

The selection issue investigated in this paper concerns the population covered by claims data. In sources such as Medicare claims, for instance, claims are only complete for the population that was continuously enrolled in FFS Medicare. These sources cannot provide reliable estimates for the population that was enrolled in MA plans. We investigate whether levels and trends of diabetes and heart attacks prevalence differ by enrollment.

Finally, we turn our attention to forgetfulness, a plausible source of measurement error in self-reports [17, 18]. Especially among older individuals who are more likely to suffer from multiple conditions and be cognitively impaired and who also show more discordance at the individual level, forgetfulness may contribute to lower disease prevalence in self-reports than in claims. We compare gaps between claims and self-reports by age group to ask whether this factor may play an important role in explaining the differences.

3 Results

3.1 Prevalence overview

An overview of disease prevalence among the population aged 67 and over across sources is shown in Figure 1. With the exception of NHANES, all sources present an upward trend in diabetes prevalence, with their 2012 values about 7 percentage points higher than their 2002 values. We find important differences in level by source, with the highest prevalence about 5 percentage points above the lowest for most years. We also find systematically higher

diabetes prevalence in claims data (Medicare 20% sample, HRS- and MCBS-linked claims, and Humana) than in self-reports (NHANES, HRS and MCBS). As was previously mentioned, this discrepancy has been previously documented [5, 7, 14, 15], but sources of differences are not yet well identified.

In contrast with the upward trend found in diabetes, Panel B of Figure 1 presents a pattern of declining rates of heart attack across sources. Divergences between claims and self-reports are mostly absent for heart attack. With the exception of HRS, which diverges somewhat after 2006, prevalence rates across sources are all contained within 0.5 percentage points of each other.

3.2 Definition Issue: Concordance between survey wording and claims

A definition issue that may cause disagreement between disease prevalence in self-reports and claims data concerns the concept of the disease studied. Heart attacks are a good case study for this potential issue. The MCBS and HRS ask their respondents whether they had an “acute myocardial infarction or heart attack” and a “heart attack or myocardial infarction” since the last interview. AMI is often considered a synonym for heart attack and it would seem natural to compare those identified by the CCW algorithm in claims to self-reports of heart attacks. However, doing so produces significantly lower rates in claims than in self-reports, as shown in Figure 2.

A plausible explanation for this discrepancy is that survey respondents may respond “yes” after a diagnosis of events and conditions other than AMI. For this reason, Yasaitis et al. [6] indicate that a better proxy for self-reports of heart diseases in claims is ACS, an umbrella term that includes AMI as well as unstable angina. As seen in Figure 2, rates of ACS in claims are remarkably closer to heart attack self-reports in HRS. Similar patterns are also found in MCBS data (Figure A1, Supplemental Digital Content 2).

These findings suggest that data users should be careful about disease definitions and concepts when measuring prevalence. For instance, using claims may be preferable to self-reports to correctly assess the number of AMIs in a given year.

3.3 Definition Issue: Reference period

A second definition issue we investigate concerns implicit imbalance in the reference periods of survey data and claims-based algorithms. For instance, the HRS asks respondents if a doctor *ever* told them they had diabetes or high blood sugar, which implies an unlimited reference period. In contrast, the CCW algorithm flags individuals as having diabetes if they met diabetes-associated diagnosis criteria in claims with a two-year reference period. Perhaps because of that imbalance, a previous study found that the CCW identified only 69% of NHANES I Epidemiologic Follow-up Study patients with diabetes within the 2-year reference period.[19]

In figure 4, we use the Medicare 20% sample data to investigate the potential implications of increasing the reference period for diabetes in claims. The solid line is the diabetes prevalence found with the CCW algorithm and its 2-year reference period. When the reference window is allowed to increase, as expected, the algorithm identifies more

individuals. Compared to the 2-year reference, a 5-year reference period, still much shorter than the “ever” concept found in the HRS questionnaire, produces a 3.5 percentage points higher prevalence.

A potential issue with this exercise is that using a 5-year reference period with the Medicare 20% sample may distort the sample by imposing that individuals were continuously enrolled in FFS for five years. For that reason, we show in Figure 4 the prevalence with a 2-year reference period when applying a restriction that the sample was enrolled in FFS for 5 years. The resulting prevalence is less than 0.01 percentage point removed from the baseline prevalence. This indicates that the higher prevalence found with a 5-year reference is indeed due to the longer reference period and not to sample distortion.

These findings are troubling, since considering a longer reference period consistent with HRS survey wording would further increase the gap between self-reports and claims. In any case, given the sensitivity of the CCW algorithm to the reference period, researchers should systematically consider alternative reference periods to best match their research objective.

We note that there is no such issue with heart attacks, given that the MCBS and HRS surveys provide sufficient information to infer whether they occurred in the past year, a reference period consistent with both the CCW and Yasaitis et al. [6] algorithms for AMI and ACS.

3.4 Selection Issue: Selection into Medicare Advantage plans

A selection issue specific to claims data pertains to the population enrolled in the plans they cover. For the 20% Medicare sample, the MCBS-linked claims and the HRS-linked claims, prevalence estimated from claims is limited to the population that selected into traditional FFS. A notable feature found in Table 1 is the difference between characteristics of the FFS and MA samples. MA enrollees are more likely to be non-white, diabetic and somewhat younger than FFS enrollees, confirming results from previous studies [7, 20-22]. To complicate matters further, selection into MA is not stable over time. Between 2004 and 2007, for instance, a new risk-adjustment system was introduced to incentivize insurers to enroll less healthy patients in private plans [23-26], which contributed to larger numbers or elderly selecting into MA. Because of changing incentives, the composition of the population has also shifted.

In this context, nationally representative surveys which ask respondents about their enrollment status, such as HRS, are useful. Figure 3 reveals a shift in diabetes trend by enrollment after 2004. From 2004 to 2012, prevalence in the FFS population grew at a steady rate of 0.7 percentage point per year, while that of the MA population grew by 1.3 percentage points per year. By 2012, diabetes prevalence in the MA population was significantly higher than that of the FFS population. The 2012 point estimate is 26.7% (s.d. of 0.9%) in the MA population and 24.7% (s.d. of 0.6%) in the FFS population. Confidence intervals are omitted from the figure for clarity. We find similar, but noisier, patterns in MCBS data (Figure A2, Supplemental Digital Content 2).

Interestingly, we found no similar patterns in heart attacks (Figure A3, Supplemental Digital Content 2). This appears consistent with analyses of the risk-adjustment reform, which

found that the new adjustment factors particularly incentivized the enrollment of diabetes patients in MA plans (Newhouse et al, 2013). These results suggest that researchers may want to compare survey prevalence for divergence by enrollment status before assuming FFS Medicare claims data to be representative of the elderly population.

3.5 Measurement Error Issue: Forgetfulness

It is not clear whether higher diabetes rates in claims mean that there are too few diabetes self-reports (false negatives) in surveys, too many patients identified as having diabetes by the CCW algorithms (false positives) in claims, or a combination of both. In the Supplemental Digital Content 1, we use HRS self-reports and linked Medicare claims to look at a potential source of false negatives in surveys: measurement error due to forgetfulness in old age. Our findings are consistent with forgetfulness contributing to measurement error at older ages, but other factors may also affect older respondents.

4 Discussion

This study estimates recent prevalence of diagnosed diabetes and heart attack events across several self-report and claims sources with the objective of drawing useful recommendations for researchers. To ensure comparability across sources, we restricted our analyses to the population aged 67 and older which was continuously enrolled in FFS Medicare and MA plans.

A key finding is that selection into different enrollment plans can matter greatly. In particular, relying on traditional Medicare ignores an increasingly important and distinct proportion of the elderly population, which opts to enroll in MA plans. With diabetes, we saw that doing so likely underestimates the true disease prevalence, since diabetes rates have become significantly higher in MA than in FFS patients. Such issues with claims could potentially be identified and addressed by relying on private MA claims data. However, as with the Humana claims we used, insurers may be geographically concentrated and unlikely to be nationally representative. Nationally representative surveys do not face the same issue.

A second finding is that, in some cases, much of the variation between claims data and surveys can be explained by differences in concepts. The consistently higher prevalence estimates using the broader ACS definition of heart attack as opposed to that used for AMI serves as an example. Survey rates could be replicated in claims when considering the broader disease category that includes other events likely to be described as heart attacks by respondents. The finding that it also improved agreement [6] suggests that ACS aligns with the interpretation of survey participants. This suggests that claims may be best suited for research on specific diseases likely to be misreported or less well-defined in surveys. Surveys may be preferred because they are nationally representative, but researchers need to be aware of potential interpretation issues from respondents. A validation study of the ACS algorithm could be the basis for an additional CCW condition for those researchers working with both surveys and claims data.

Variation in algorithms for identifying disease in claims explains part of the resulting differences in prevalence estimates and concordance in the literature. Comparison of results

across studies would benefit from grounding in a standard algorithm and lead to better understanding measures of disease taken from claims, which is essential as we attempt to apply large administrative data to research questions. CCW provides an excellent starting point in this pursuit by defining standard algorithms that are relatively simple to implement. However, the algorithms do not all perform equally well for all conditions[1, 19], and should be scrutinized on a condition-by-condition basis before considering claims as “truth”. The still unresolved example of diabetes rates being higher in claims despite the shorter reference period (two years versus ever)suggests that the CCW algorithm could include substantial false positives, especially given the documented poor capacity of the algorithm to detect known cases of diabetes[19, 27]. The CCW algorithm requires two diabetes diagnoses at least a day apart within the reference period for outpatient claims, to address the practice of coding a “rule-out” diagnosis when ordering tests for diabetes. Perhaps this requirement is not stringent enough to eliminate false positives, particularly if the reference period is extended.

A limitation of this article is that we did not look into agreement at the individual level. In future work, we will focus on fewer data sources and investigate how data issues impacted entification of disease in individuals. In particular, we will further investigate reasons for the larger diabetes prevalence in claims. As part of our explorations of claims data, we observed a non-trivial proportion of individuals with CCW-identified diabetes in a single year, or more frequently pairs of years because of the 2-year reference periods, followed by multiple years without a diabetes diagnosis. Analysis of this group should reveal potential sources of false positives. For example, is it possible that individuals had two lab tests, associated with an annual physical, i.e. that they were both “rule-out” diagnoses? Bringing provider types into the algorithm may help address this possibility and its potential effects on prevalence estimates. Should we find a method that produces estimates closer to those found in self-reports, it could be the basis for a validation study that could inform the CCW algorithm.

Future work should also consider additional data sources to investigate which of survey data or claims are more accurate. In particular, data from electronic health records (EHR) may provide a more reliable “gold standard” to compare to claims and self-reports, although these too have limitations when individuals see multiple providers. Prescription data is another relevant source of information for conditions that typically require medication, such as diabetes and hypertension.

In conclusion, the comparisons of estimated prevalence by claims and self-report reveal that there is more to learn in effective application of both administrative and survey data to research. Surveys are critical for providing context, including rich information about the individuals, to the observations regarding population health. They help us validate, improve, and better understand prevalence measures we derive from claims data, and provide less biased samples from which to produce estimates. Claims data provide a different view, with detailed standard coding for diagnoses and treatments. But those may be subject to trends and regional differences in coding practice, and to shifts in regulations and policy. They can help us understand what respondents mean when they answer questions about their health. Together these different data sources can be applied to better interpret and track disease dynamics across populations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding disclosure: The authors are grateful to the National Institute on Aging for its support through the Roybal Center for Health Policy Simulation (grant no. P30AG024968). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Wolinsky FD, et al. The concordance of survey reports and Medicare claims in a nationally representative longitudinal cohort of older adults. *Med Care*. 2014; 52(5):462–8. [PubMed: 24714584]
2. Zuvekas SH, Olin GL. Accuracy of Medicare expenditures in the medical expenditure panel survey. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*. 2009; 46(1):92–108.
3. Cleary PD, Jette AM. The validity of self-reported physician utilization measures. *Medical care*. 1984;796–803. [PubMed: 6492908]
4. Wolinsky FD, et al. Hospital episodes and physician visits - The concordance between self-reports and Medicare claims. *Medical Care*. 2007; 45(4):300–307. [PubMed: 17496713]
5. Heiss, F., et al. Measuring Disease Prevalence in Surveys: A Comparison of Diabetes Self-reports, Biomarkers, and Linked Insurance Claims, in *Insights in the Economics of Aging*. University of Chicago Press; 2015.
6. Yasaitis LC, Berkman LF, Chandra A. Comparison of self-reported and Medicare claims-identified acute myocardial infarction. *Circulation*. 2015; 131(17):1477–85. discussion 1485. [PubMed: 25747935]
7. Sakshaug JW, Weir DR, Nicholas LH. Identifying diabetics in Medicare claims and survey data: implications for health services research. *BMC Health Serv Res*. 2014; 14(1):150. [PubMed: 24693862]
8. Taylor DH Jr, Fillenbaum GG, Ezell ME. The accuracy of medicare claims data in identifying Alzheimer's disease. *J Clin Epidemiol*. 2002; 55(9):929–37. [PubMed: 12393082]
9. Østbye T, et al. Identification of dementia: agreement among national survey data, medicare claims, and death certificates. *Health services research*. 2008; 43(1p1):313–326. [PubMed: 18211532]
10. Rector TS, et al. Specificity and Sensitivity of Claims-Based Algorithms for Identifying Members of Medicare+ Choice Health Plans That Have Chronic Medical Conditions. *Health services research*. 2004; 39(6p1):1839–1858. [PubMed: 15533190]
11. Kern EF, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. *Health Serv Res*. 2006; 41(2):564–80. [PubMed: 16584465]
12. Klabunde CN, Harlan LC, Warren JL. Data sources for measuring comorbidity: a comparison of hospital records and medicare claims for cancer patients. *Med Care*. 2006; 44(10):921–8. [PubMed: 17001263]
13. Harrold LR, et al. Validity of gout diagnoses in administrative data. *Arthritis Rheum*. 2007; 57(1): 103–8. [PubMed: 17266097]
14. Day HR, Parker JD. Self-report of diabetes and claims-based identification of diabetes among Medicare beneficiaries. *National health statistics reports*. 2013:1.
15. Ngo DL, et al. Agreement between Self-Reported Information and Medical Claims Data on Diagnosed Diabetes in Oregon's Medicaid Population. *Journal of Public Health Management and Practice*. 2003; 9(6):542–544. [PubMed: 14606195]
16. Centers for Medicare & Medicaid Services. Medicare Advantage Plans. [cited 2016 July 31, 2016]; Available from: <https://www.medicare.gov/sign-up-change-plans/medicare-health-plans/medicare-advantage-plans/medicare-advantage-plans.html>

17. Kent, RA. Analysing quantitative data: Variable-based and case-based approaches to non-experimental datasets. Sage; 2015.
18. McDonald-Miszczak L, Neupert SD, Gutman G. Does cognitive ability explain inaccuracy in older adults' self-reported medication adherence? *Journal of Applied Gerontology*. 2009; 28(5):560–581.
19. Gorina Y, Kramarow EA. Identifying chronic conditions in Medicare claims data: evaluating the Chronic Condition Data Warehouse algorithm. *Health Serv Res*. 2011; 46(5):1610–27. [PubMed: 21649659]
20. Sakshaug JW, Kreuter F. Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*. 2012
21. Al Baghal T, Knies G, Burton J. Linking administrative records to surveys: differences in the correlates to consent decisions. *Understanding Society at the Institute for Social and Economic Research*. 2014
22. Sala E, Knies G, Burton J. Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*. 2014; 17(5):455–473.
23. McWilliams JM, Hsu J, Newhouse JP. New Risk-Adjustment System Was Associated With Reduced Favorable Selection In Medicare Advantage. *Health Affairs*. 2012; 31(12):2630–2640. [PubMed: 23213147]
24. Newhouse JP, et al. Do Medicare Advantage plans select enrollees in higher margin clinical categories? *Journal of health economics*. 2013; 32(6):1278–1288. [PubMed: 24308879]
25. Brown J, et al. How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program. *American Economic Review*. 2014; 104(10):3335–64.
26. Medicare Payment Advisory Commission. Report to the Congress: Medicare and the health care delivery system. 2013 MedPAC.
27. Richesson RL, et al. A comparison of phenotype definitions for diabetes mellitus. *J Am Med Inform Assoc*. 2013; 20(e2):e319–26. [PubMed: 24026307]

Figure 1. Prevalence of diabetes and heart attack by source, population aged 67 and over

A. Diabetes.

B. Heart Attack. NHANES is omitted from the figure since it asks respondents if they “... ever had a heart attack”, which differs from HRS and MCBS questions (see Supplemental Digital Content 3 for more information)

Figure 2. Prevalence of heart attack, AMI and ACS in HRS, population aged 67 and over enrolled in fee-for-service Medicare

Figure 3. Prevalence of diabetes in the 20% sample of Medicare Claims by time in fee-for-service Medicare plan and algorithm reference period, population aged 70 and over

Figure 4. Prevalence of diabetes in HRS self-reports and linked Medicare claims by plan enrollment, population aged 67 and over

Table 1
Descriptive Statistics by Data Source and Enrollment Status, 2008

A. Self-reports

Source (period used)	HRS (2002-2012)		MCBS (2002-2012)			NHANES* (2002-2011)		
	All	FFS	All	Mixed	MA	All	Mixed	MA
Average age	76.8	76.9	76.4	75.5	76.4	76.6	73.7	74.7
Male (%)	41.5%	41.1%	43.2%	41.7%	42.8%	41.9%	53.9%	42.1%
Non-Hispanic white (%)	85.4%	86.9%	81.4%	73.9%	84.8%	88.9%	71.5%	81.8%
Black (%)	8.0%	7.8%	8.2%	11.5%	7.9%	6.4%	13.9%	8.2%
Hispanic (%)	6.6%	5.3%	10.4%	14.6%	7.4%	4.7%	14.6%	6.7%
Disease prevalence (%)								
Diabetes	22.5%	22.4%	23.2%	20.8%	24.2%	23.1%	25.5%	21.7%
Heart attack	2.4%	2.5%	2.1%	2.3%	1.8%	1.6%	1.5%	
Sample size	9,371	7,173	1,818	380	8,586	6,033	530	1,392

B. Claims

Source (period used)	HRS-linked claims (2002-2008)		MCBS-linked claims (2002-2012)		Medicare 20% Sample (2002-2013)			Humana (2007-2012)
Enrollment	FFS		FFS		FFS	MA	Mixed	MA
Average age	77.6		77.0		76.6	76.0	75.6	73.7
Male (%)	40.5%		41.9%		41.3%	41.6%	40.3%	45.5%
Non-Hispanic white (%)	91.2%		88.9%		88.1%	79.3%	68.9%	
Black (%)	7.1%		6.4%		7.0%	9.2%	17.7%	
Hispanic (%)	1.7%		4.7%		4.9%	11.5%	13.3%	
Disease prevalence (%)**								
Diabetes	27.1%		27.8%		27.4%			27.4%
Acute myocardial infarction	0.9%		0.8%		0.8%			0.8%
Acute coronary syndrome	2.3%		1.9%		1.9%			1.8%
Sample size	6,224		6,033		4,252,744	1,391,827	126,948	763,844

* NHANES column corresponds to year 2007-2008;

** Disease prevalence in claims derived using Chronic Conditions Data Warehouse algorithms. FFS – population continuously enrolled in fee-for-service Medicare; MA – population continuously enrolled in Medicare Advantage plans; Mixed – population with a mixed fee-for-service and Medicare Advantage enrollment.