A Method for Constructing a Longitudinal Sample of Medicare Patients with Application to Diabetes Outcomes Research

Timothy L. McMurry[1], Jennifer M. Lobo[1], Soyoun Kim[1], Hyojung Kang[2], Min-Woong Sohn[3*]

[1]University of Virginia, Department of Public Health Sciences, Charlottesville, VA 22908

[2]University of Illinois, Department of Kinesiology and Community Health, Champaign, IL 61820

[3]University of Kentucky, College of Public Health, Lexington, KY 40506

*Correspondence to Min-Woong Sohn, min-woong.sohn@uky.edu

Abstract

Background We present a method of randomly drawing a longitudinal sample of Medicare patients to 1) conduct longitudinal analysis of care use and outcomes over a ten-year follow-up; 2) provide a representative cross-sectional sample in each year during this time period; and 3) provide adequate precision for estimates in comparisons that involve minority patients at the county level. This method was applied to patients with diabetes in the Diabetes Belt (a region in the Appalachian and southern US with higher rates of diabetes) and surrounding counties.

Methods We used the Medicare Master Beneficiary Summary Files (A/B/C/D and Chronic Conditions segments) to identify eligible patients for each year. We targeted a sample of just under 900,000 patients per year. The 2006 sample is stratified by county and white/minority status, and targeted at least 250 patients in each stratum with the remaining sample allocated proportional to county size with oversampling of the minority population. Patients who were alive, did not move between counties, and stayed enrolled in Medicare fee-for-service (FFS) were retained in the sample for subsequent years. Non-retained patients were replaced with a sample of patients in their first year of Medicare FFS (e.g., new enrollees) or patients who moved into a sampled county during the previous year.

Results The resulting sample contains an average of 899,266 patients each year and closely matches population demographics and chronic conditions. For all years in the sample, the weighted sample average age and the population age differ by < 0.01 years; the proportion white is within 0.01%; and the proportion female is within 0.08%. No difference was statistically significant at the $\alpha = 0.05$ level.

Conclusions This carefully constructed survey sample will allow us to to perform longitudinal and cross-sectional analysis on healthcare utilization and outcomes. This sampling strategy can be easily adapted to other projects that require random samples of Medicare beneficiaries for longitudinal follow-up with possible oversampling of some sub-populations.

Keywords Diabetes; Medicare claims; sample; longitudinal

Background

Medicare claims data capture national data on health care utilization for Americans aged 65 years old or older, disabled, or with end-stage renal disease (ESRD). Medicare is the only source of national data on healthcare utilization, thus its importance for epidemiological and health services research cannot be overemphasized. Due to costs of acquiring Medicare claims data from the Centers for Medicare and Medicaid Services (CMS), full data on the Medicare population with longitudinal follow-up over several years may not be feasible or practical. For this reason, researchers often need to work with a representative sample of the Medicare population. In this paper we present a method of drawing a random sample that is representative of the Medicare population for both longitudinal follow-up and cross-sectional analysis.

We will illustrate this method with an example of a sampling design for a longitudinal study of Medicare patients with diabetes living in the Diabetes Belt and surrounding counties. The study is designed to assess the care received by and outcomes for these patients. The Diabetes Belt (Barker et al., 2011) is a collection of 644 counties in the Appalachian and southeastern United States that had diabetes prevalence of at least 11% in 2007. This area continues to have high prevalence of diabetes (CDC, 2020), which motivates the study of care delivery and outcomes in this population. Our survey additionally includes diabetic patients in geographically surrounding counties; these counties are expected to be as culturally similar as possible while providing a comparison population with a lower burden of diabetes and diabetes care.

The end goals of the study informed the sampling design we will describe. We plan to use this data to track changes in patient care, practice patterns, and outcomes over time. The sample we describe was designed to provide valid inference around these goals for patients with diabetes living in the Diabetes Belt and surrounding counties during the years from 2006 to 2015. In addition to providing longitudinal assessment of patients, the study is also designed to produce representative cross-sectional samples in each year, and to provide comparisons within counties.

From a sampling design perspective, the goals we have outlined are somewhat in conflict. For example, if the goal is to provide similar precision within each county, then the optimal sampling design would, as well as possible, sample approximately the same number of people in each county. In contrast, if the goal is to provide the best population level descriptive data, then sampling from each county in proportion to its size is approximately optimal (Lohr, 1999, Section 4.4, 104–108). The desire to compare white and minority populations in our study suggested oversampling of whichever group is smaller in each county. While surveys designed for a specific primary analysis can be further optimized, our survey needs to provide reasonable analytic power for multiple aims. This sampling design will provide good precision for a wide range of analyses.

Our goal was to enable estimates with good precision for a wide range of outcomes while keeping the total sample to less than 1,000,000 beneficiaries, the lowest tier of Medicare data pricing (RESDAC, 2016). With 100,000 set apart for a third comparison group beyond the scope of this report, our target sample size was just under 900,000. In the remainder of this paper we describe the methods used to construct this sample and we present an analysis that demonstrates the resulting sample was representative of the target population and that demographic estimates based on this sample have high precision and accuracy.

Methods

Population

We used the Medicare Master Beneficiary Summary Files (A/B/C/D and Chronic Conditions segments) to identify Medicare patients meeting inclusion criteria each year from 2006 to 2015. To be eligible for inclusion, Medicare patients needed to have been previously diagnosed with diabetes (identified in the Chronic Conditions segment), be living in the Diabetes Belt or surrounding counties, and be enrolled in Medicare Fee-for-Service for 12 months each year. Patients enrolled in Medicare HMOs were excluded because their claims data are not available.

Diabetes Belt and Surrounding Counties

The Center for Disease Prevention and Control (CDC) identified 644 counties across 15 states in the Appalachian region and southeastern US as the Diabetes Belt (Barker et al., 2011). Some or all counties in Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Ohio, Pennsylvania, South Carolina, Tennessee, Texas, Virginia, and West Virginia comprise the Belt. We used the CDC's definition based on 2008 data in this study. We identified 310 counties that are closest but not contiguous to the Belt counties as surrounding counties. They were chosen to serve as a basis for comparisons with the Belt counties. Counties that are immediately adjacent to the Belt were not included among the surrounding counties because some patients may cross county boundaries to seek care and may confound our estimates on preventive care utilization and complication rates.

Construction of the 2006 Sample

The sample for 2006 was a random sample of eligible patients stratified by county and race. We divided race/ethnicity into two groups, non-Hispanic white and all minorities. We did not further sub-divide the sample by race/ethnicity because there were relatively few individuals of Hispanic ethnicity and other races of Medicare age residing in the Diabetes Belt and surrounding areas during this time period. A very small number of patients with missing race/ethnicity were included in the sampling frame along with the white population.

In order to balance the competing needs for county and regional level inference, we first allocated a sample of 500 persons to each county (or the county eligible population if less than 500). We considered several alternatives between 500 to 1000. We found that 500 allowed a complete enumeration for the smallest 18% of counties and at least 50% sampling for 70% of counties while still allowing significant sampling in the most populous regions. We then allocated the remaining available sample to each county proportional to the size of its un-sampled population, with the constant of proportionality chosen to produce a sample size as close as possible to the 900,000 person target; the resulting sampling rate was approximately 30% of the remaining population.

Within counties we then initially tried to allocate a sample of size 250 (or the population size if less than 250) to the white population and 250 for the minority population. Remaining samples allocated to the county were then divided between the white and minority populations according to the proportion

$$p^s_{im} = 2(p^r_{im} - 1/2)^3 + (p^r_{im} - 1/2)/2 + 1/2,$$

where p^s_{im} represents the non-white proportion in the remaining sample for county i, p^r_{im} represents the non-white proportion of the unsampled population of county i (Figure 1). In counties where the non-white population is in the minority, this formula oversamples the non-white population by

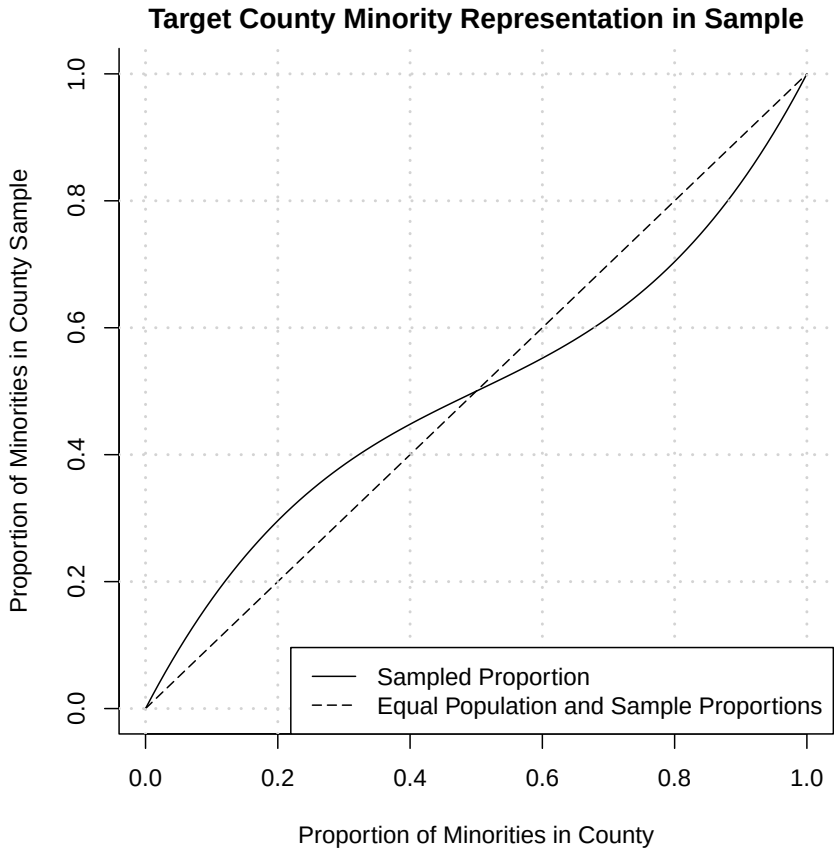

Figure 1: Targeted sampling proportion of minorities in each county.

a rate of approximately two-to-one when the non-white population is proportionally small, while dropping to equal sampling as the white and non-white populations become equal. In counties where the white population is in the minority, the white population was oversampled. Our goal was to oversample whichever group (white/non-white) was in the minority in each county in order to improve within county comparisons while still providing significant coverage of the white population, which encompasses approximately 80% of the population living with diabetes in this region.

Once we had defined the sample size by stratum (county and white/non-white), we then selected patients using simple random sampling within strata. Survey weights were defined to be the stratum population size divided by the stratum sample size.

Construction of Subsequent Years' Samples

Sampling in subsequent years was slightly complicated by the demands of retaining patients for longitudinal follow-up and ensuring a cross-sectionally representative sample in each year. However, the guiding principle is straightforward.

The guiding principle is that the 2006 sample was representative of all Medicare patients who had diabetes and met inclusion criteria. If for 2007 we retained all patients from the 2006 sample who had remained alive and eligible, then these patients were representative of the population who had been eligible for at least 1 year (and they were therefore approximately 1 year older than the overall population). In order to replace patients in the 2006 data set who had died, enrolled in a Medicare HMO, or moved, we replaced them with an appropriately weighted sample of patients

who were not eligible for inclusion in 2006.

We constructed the 2007 replacement sample to first allocate at least 10 patients to each stratum to ensure that we add new beneficiaries in every county every year. Additional patients were then allocated to each stratum to target the overall sample size as would have been calculated using the 2006 sampling procedure on the 2007 county populations. All replacement patients were sampled from the population who would have been ineligible in 2006 (not enrolled in Medicare, in a Medicare HMO, lived elsewhere, or were first diagnosed with Diabetes in 2007). Sampling weights were calculated as the number of first year eligible white/minority population in each county divided by the corresponding fill-in sample size. We similarly constructed the 2008–2015 replacement samples.

Comparison of Sample to Population

In order to ensure the sample demographics reflected the underlying population for each year, we compared the randomly selected sample to the population. This analysis was performed using weighted survey sample analysis procedures (e.g. Stata "survey" suite of programs) with weights as described above and sampling strata defined by county, white/minority, and year the patient was added to the sample.

Variance Estimation

Estimating variances (for standard errors and p-values) is a long-standing challenge in survey analysis with many possible approaches. For standard statistics (means, proportions, totals, regression coefficients), Taylor series based methods are built into all statistical survey analysis packages (e.g. SAS Survey procedures, the R Survey package, or the Stata Survey suite). We used this approach for the cross-sectional analyses presented below. For more complicated statistics, such as longitudinal models, resampling or jackknife methods are typically used; see Wolter (2007) for a good and very readable overview.

Data cleaning was performed in SAS (v9.4, Cary, NC) and Stata SE (v15.1, College Station, TX); the random sample was generated using an R (v3.6.1, Vienna, Austria) program which is available on request; and descriptive statistics and comparison to the reference population were calculated using Stata survey programs. This study was approved by the University of Virginia institutional review board (IRB #21127).

Results

Our goal was to create a sample in each year from 2006 to 2015 that satisfied our research need for longitudinal follow-up and cross-sectional analysis. We targeted a stratified random sample of about 900,000 from the Diabetes Belt and surrounding counties. Our sample design yielded an average sample size of 899,266 (standard deviation 408) over the 10-year study period. A total of 28% of the 2006 sample was retained for the full 10 year study period; these 28% included more than 200,000 non-Hispanic white and 70,000 minority patients. Table 1 shows year-by-year retention based on year of initial sampling. Although Hispanic and other race/ethnicity groups represented less than 1% of our total sample, the study retained a substantial number ($\sim$ 2000 or more) for area-wide subgroup comparison and for longitudinal follow-up.

	Initial Inclusion Year										
Survey Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
2006	899,846 (100%)										899,846
2007	758,145 (84%)	140,283 (100%)									898,428
2008	650,673 (72%)	118,533 (84%)	130,027 (100%)								899,233
2009	571,730 (64%)	103,100 (73%)	111,862 (86%)	112,963 (100%)							899,655
2010	498,506 (55%)	90,072 (64%)	96,198 (74%)	95,862 (85%)	118,543 (100%)						899,181
2011	442,871 (49%)	80,911 (58%)	85,636 (66%)	83,960 (74%)	102,075 (86%)	104,002 (100%)					899,455
2012	389,932 (43%)	72,357 (52%)	76,100 (59%)	74,207 (66%)	88,559 (75%)	88,406 (85%)	109,928 (100%)				899,489
2013	338,896 (38%)	63,811 (45%)	67,016 (52%)	65,078 (58%)	77,228 (65%)	75,609 (73%)	92,888 (84%)	118,605 (100%)			899,131
2014	291,437 (32%)	56,067 (40%)	58,805 (45%)	57,016 (50%)	67,392 (57%)	65,440 (63%)	79,394 (72%)	99,827 (84%)	123,466 (100%)		898,844
2015	252,870 (28%)	49,980 (36%)	52,255 (40%)	50,724 (45%)	59,695 (50%)	57,696 (55%)	69,796 (63%)	86,294 (73%)	104,600 (85%)	115,491 (100%)	899,401

Table 1: Longitudinal retention (sample size and percent) in the Medicare data by year of initial inclusion. Rows are the sampled year of Medicare records and columns indicate the year subjects were initially sampled.

In order to assess the resulting sample, for each year of the survey we made cross-sectional comparisons of the weighted sample to the population defined from the Medicare Master Beneficiary Summary Files (described in Table 1). Population size, race, and previous year sample eligibility were the factors we used in determining the sample. We therefore focused our descriptive statistics on race, age, and population totals. Age is a particularly important variable for assessment because if the fill-in samples were incorrectly constructed, we would expect to see drift from the underlying population as the retained samples from previous years aged. We additionally included sex because it is an important factor in most health outcomes and it provides a good additional point of comparison that was not incorporated in the sampling design.

Results are shown in Table 2. For all years in the sample, the weighted sample average age and the population age differ by less than 0.01 years; the proportion white is within 0.01%; and the proportion female is within 0.08%. No difference was statistically significant at the $\alpha = 0.05$ level. Figure 2 shows that in the last year of follow-up (2015) the weighted age distribution closely matches the population age distribution. This comparison provides a visual check that the fill-in samples from years 2007–2015 were appropriately weighted to allow for valid cross-sectional comparisons of age.

	Age			Sex (% Male)			% Non-Hispanic White (or Missing)		
Year	Pop.	Samp.	p-value	Population	Sample	p-value	Population	Sample	p-value
2006	75.09	75.09	0.927	818,368 (42.9%)	812,114 (42.8%)	0.193	1,494,096 (78.8%)	1,494,096 (78.8%)	1.000
2007	75.19	75.19	0.896	828,638 (43.2%)	826,912 (43.2%)	0.126	1,524,182 (79.5%)	1,523,807 (79.5%)	0.930
2008	75.24	75.23	0.949	844,511 (43.4%)	842,987 (43.4%)	0.249	1,543,822 (79.4%)	1,543,209 (79.4%)	0.888
2009	75.29	75.29	0.718	867,876 (43.6%)	866,994 (43.5%)	0.497	1,573,679 (79.0%)	1,573,056 (79.0%)	0.764
2010	75.37	75.37	0.871	890,438 (43.8%)	890,163 (43.8%)	0.872	1,599,612 (78.6%)	1,598,752 (78.6%)	0.927
2011	75.38	75.38	0.641	929,910 (44.0%)	930,815 (44.0%)	0.361	1,658,040 (78.4%)	1,657,943 (78.4%)	0.956
2012	75.32	75.33	0.317	955,839 (44.2%)	956,853 (44.2%)	0.261	1,692,521 (78.2%)	1,692,285 (78.2%)	0.979
2013	75.30	75.31	0.267	970,658 (44.5%)	971,854 (44.5%)	0.255	1,702,352 (78.0%)	1,702,401 (78.0%)	0.975
2014	75.27	75.27	0.860	978503 (44.8%)	980,321 (44.9%)	0.117	1,700,496 (77.9%)	1,700,924 (77.9%)	0.909
2015	75.27	75.27	0.958	997,913 (45.1%)	999,868 (45.2%)	0.170	1,717,930 (77.6%)	1,719,008 (77.6%)	0.834

Table 2: Cross-sectional comparison of population and weighted sample by age, sex, and race/ethnicity by year.

Discussion

In this report we described a sampling method which produced a representative sample of Medicare patients with good cross-sectional properties while still retaining patients for longitudinal analysis. We will use this data to assess trends in preventive care utilization, long term outcomes, disparities, and associations between preventive care and diabetic complications in patients with diabetes. These goals require the sample to be valid both longitudinally and cross-sectionally.

A primary goal in this work was to document these methods for future researchers who might be interested in obtaining representative samples of Medicare claims data. In preparing for this project, we found only limited literature describing longitudinal sampling designs that could serve as a reference. Smith et al. (2009) offers a very high level overview and describes the principles of sampling design for longitudinal surveys. Other articles address subsets of our challenge. For example, Wolinsky et al. (2014) discusses matching Medicare claims to a longitudinally followed cohort without need for cross-sectional inference, while Thompson (2015) and Carrillo and Karr (2013) focus primarily on analytic approaches rather than design. While longitudinal surveys are not rare (for example the Population Assessment of Tobacco and Health (PATH) study, Hyland et al., 2017), they are largely the purview of governments or large survey research organizations

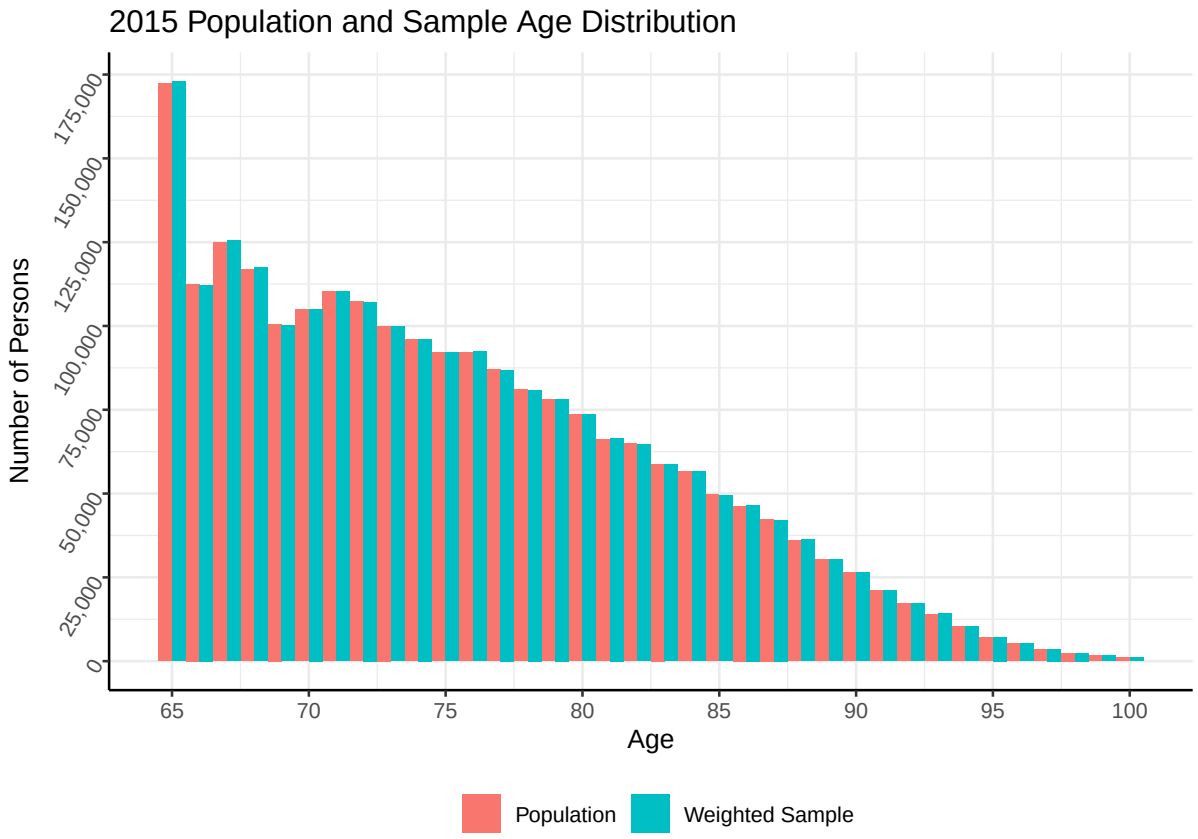

Figure 2: Comparison of population and weighted sample age distribution for 2015.

because it is hard for a small team to longitudinally track and follow-up with a large number of patients. Nonetheless, there are cases, such as this one, where a longitudinal survey is feasible as part of a smaller project. We hope this report will be useful to researchers interested in designing their own studies of this type.

In constructing this sample, we found that it was relatively easy to produce a representative sample for the baseline year (2006). Significantly more care needed to be taken to identify the sampling frame for subsequent years. An advantage of working with samples this large is that there is abundant power to identify potential problems before collecting data.

This study will be limited by the scope of Medicare claims data. In particular, this cohort only includes patients ages 65 and up, and it excludes those patients enrolled in a Medicare HMO. In addition, Medicare only captures claims data, so we will not have access to full clinical records.

Conclusions

We demonstrated that a representative sample of Medicare beneficiaries can be carefully constructed to be used in cross-sectional as well as longitudinal analyses. This sampling method will make the data request much more affordable. The computer algorithms we created can be used by future researchers in drawing random representative samples from Medicare claims data.

List of abbreviations

CDC Center for Disease Prevention and Control

CMS Centers for Medicare and Medicaid services

ESRD End-stage renal disease

Declarations

Ethics approval and consent to participate

This study was approved by the University of Virginia Internal Review Board (IRB #21127). Consent was not required because it is a retrospective analysis of existing data.

Consent for publication

This study does not include any individual data.

Availability of data and materials

The data that support the findings of this study are available from the Research Data Assistance Center (ResDAC, `https://www.resdac.org/`); they are used under agreement and cannot be released publicly. Interested researchers who would like to work with this or similar data should contact ResDAC. R code used to construct the sample is available from the authors on request.

Competing interests

The authors have no competing interests.

Funding

This work was funded by the NIH/NIDDK grant R01DK113295.

Authors' contributions

TM developed the sampling algorithm with MS, implemented it, and wrote the majority of the manuscript. JL helped frame the research questions and sampling approach, and carefully reviewed and edited this manuscript. SK defined the original population from which this cohort was drawn, and was responsible for checking and evaluating the resulting sample. HK helped frame the research question and carefully reviewed the manuscript. MS developed the sampling algorithm with TM, implemented the algorithm independently, helped frame the research questions.

Acknowledgements

Not applicable.

References

Barker LE, Kirtland KA, Gregg EW, Geiss LS, Thompson TJ (2011) Geographic distribution of diagnosed diabetes in the us: a diabetes belt. American Journal of Preventive Medicine 40(4):434–439

Carrillo IA, Karr AF (2013) Combining cohorts in longitudinal surveys. Survey Methodology 39(1):149–182

CDC (2020) National Diabetes Statistics Report. URL, URL `https://www.cdc.gov/diabetes/data/statistics/statistics-report.html`, accessed March 26, 2020

Hyland A, Ambrose BK, Conway KP, Borek N, Lambert E, Carusi C, Taylor K, Crosse S, Fong GT, Cummings KM, Abrams D, Pierce JP, Sargent J, Messer K, Bansal-Travers M, Niaura R, Vallone D, Hammond D, Hilmi N, Kwan J, Piesse A, Kalton G, Lohr S, Pharris-Ciurej N, Castleman V, Green VR, Tessman G, Kaufman A, Lawrence C, van Bemmel DM, Kimmel HL, Blount B, Yang L, O'Brien B, Tworek C, Alberding D, Hull LC, Cheng YC, Maklan D, Backinger CL, Compton WM (2017) Design and methods of the population assessment of tobacco and health (path) study. Tobacco Control 26(4):371–378, DOI 10.1136/tobaccocontrol-2016-052934, URL `https://tobaccocontrol.bmj.com/content/26/4/371`

Lohr SL (1999) Sampling: Design and Analysis, 1st edn. Duxbury, Pacific Grove

RESDAC (2016) CMS Price List for Research Files. URL, URL `https://www.resdac.org/sites/resdac.umn.edu/files/CMS%20Price%20List%20for%20Research%20Files_23.pdf`, accessed February 5, 2020

Smith P, Lynn P, Elliot D (2009) Sample design for longitudinal surveys. Methodology of Longitudinal Surveys pp 21–33

Thompson ME (2015) Using longitudinal complex survey data. Annual Review of Statistics and Its Application 2:305–320

Wolinsky FD, Jones MP, Ullrich F, Lou Y, Wehby GL (2014) The concordance of survey reports and medicare claims in a nationally representative longitudinal cohort of older adults. Medical Care pp 462–468

Wolter K (2007) Introduction to Variance Estimation, 2nd edn. Springer Science & Business Media, New York