Applying Weighting Methodologies to a Commercial Database to Project US Census Demographic Data

THOMAS WASSER, PHD, MED; BINGCAO WU, MS; JOSEPH W. YČAS, PHD; AND OZGUR TUNCELI, PHD

ABSTRACT

OBJECTIVES: The objective was to investigate the viability of projecting demographic information from a large commercial managed care database to the entire US population, and to provide a simple, pertinent weighting scheme.

METHODS: Data from the HealthCore Integrated Research Database (HIRD), a repository of enrollee administrative claims from 14 regionally dispersed US health plans, were compared with US Census data. Census-defined regions, gender, and age groups served as demographic standards. To guard against small differences between these large samples appearing statistically significant, an alternative version of goodness-of-fit statistics was used to assess the overall fit of characteristic group variables.

RESULTS: This study compared 14.8 million HIRD enrollees and the 307.7 million individuals from the 2009 US Census. Gender distribution was similar in the groups: females comprised 49.8% (HIRD) and 49.3% (Census). Relative to the US Census, HIRD enrollees were overrepresented in the midwest, underrepresented in the south, and comparable in the northeast and west, with differences of 1% and 0.6%, respectively. HIRD was overrepresented in the 30-to-59 years category and underrepresented in the <5 years and ≥65 years groups; the groups were similar in the 5-to-30 years age group.

CONCLUSIONS: In the absence of data on disease prevalence, treatment patterns, and outcomes, commercial health plan databases may provide a reasonable representation of the national population when appropriately weighted to reflect differential demographic characteristics. The ability to conduct and rely on the results of such projections could be of value to key stakeholders such as healthcare planners, policy makers, and payers.

Researchers are keenly interested in ascertaining the impact of disease on society. One of the central elements of this determination is knowledge about the number of individual patients with a disease or condition of interest within a specific region, age group, or gender.[1-6] The exact count or even estimates of patients affected by a given disease may not always be available for a variety of reasons, including the absence of reporting requirements or a lack of organized and maintained disease registries or longitudinal patient databases.[7,8] To obtain an understanding of the size of patient populations that are not well quantified and characterized, often the only workable option is to extrapolate from available data in repositories such as registries and health plan databases (among others).

Disease prevalence can be estimated in subpopulations with accessible data,[9-16] but in extrapolating to the general population, systematic differences in demographic composition must be taken into account. In the United States, it is unlikely that data sets in existing commercial health insurance databases will be representative enough by themselves to present an accurate estimate of the national population.[10,11,14-16] As a result, there is considerable interest in census decomposition methodologies or similar approaches that are capable of rendering the data in such nonrepresentative population samples in a form comparable to US Census data.

Cognizant of their role as a vital and reliable source of data on disease prevalence and the size limitations of commercial health plan databases, the objective of this study was to develop a weighting framework for projecting data from commercial databases to a population matching the demographic composition encompassed by the US Census.

Table 1. Weights[a] Against 2009 ACS Estimates Based on Region, Age Group, and Gender

Age Group (years)	US CENSUS BUREAU REGION							
	MIDWEST		NORTHEAST		WEST		SOUTH	
	Male	Female	Male	Female	Male	Female	Male	Female
<5	0.8855	0.8905	1.2679	1.2746	1.4186	1.4118	1.7146	1.7563
5-9	0.7451	0.7419	1.1018	1.1168	1.1734	1.1526	1.4390	1.4487
10-14	0.7097	0.7066	1.1003	1.0791	1.1514	1.1531	1.3670	1.3599
15-19	0.7044	0.7135	1.1029	1.0936	1.0610	1.0310	1.2727	1.3155
20-24	0.6754	0.7012	1.0710	0.9974	0.9656	0.8419	1.1234	1.1725
25-29	0.6552	0.6852	1.1479	0.9461	1.0043	0.8637	1.0839	1.1452
30-34	0.5957	0.6335	1.0103	0.8833	0.9659	0.8806	1.0215	1.1015
35-39	0.5976	0.6234	0.9455	0.8902	0.9130	0.8421	0.9946	1.0857
40-44	0.5975	0.6253	0.9594	0.9051	0.9058	0.8161	0.9895	1.0642
45-49	0.6071	0.6203	0.9690	0.9035	0.8791	0.8297	0.9877	1.0544
50-54	0.6132	0.6193	0.9416	0.8837	0.8852	0.8643	1.0080	1.0668
55-59	0.6166	0.6274	0.9049	0.8661	0.9085	0.9030	1.0538	1.1256
60-64	0.6844	0.6538	0.9182	0.9124	1.0002	0.9778	1.2411	1.2369
65-69	0.8272	0.8822	1.2571	1.3722	1.2349	1.3846	2.0573	2.3048
70-74	1.0572	1.0944	1.7079	1.8468	1.5124	1.9060	3.1928	3.3687
75-79	1.1018	1.1606	1.8038	2.0603	1.8906	2.4179	3.6832	3.4855
80-84	1.1573	1.1835	2.0878	2.4174	1.9982	2.6413	3.4134	3.3911
≥85	1.1026	1.0703	2.1032	2.4972	2.2042	2.4622	3.3312	2.6673

ACS indicates acute coronary syndrome.

[a]Weight = $\frac{\%\text{ of the total US Census population}}{\%\text{ of the overall HIRD population}}$

Table 2. Demographic Comparison Between HIRD and US Census

	HIRD		US Census	
	N (million)	%	N (million)	%
Total Population	14.8[a]	100.0%	307.0	100.0%
Gender				
Male	7.4	49.8%	151.4	49.3%
Female	7.4	50.2%	155.6	50.7%
Region				
Midwest	4.6	31.2%	66.8	21.8%
Northeast	2.5	17.0%	55.3	18.0%
West	3.3	22.7%	71.6	23.3%
South	4.3	29.1%	113.3	36.9%

HIRD indicates HealthCore Integrated Research Database.

[a]Represents approximately 4.82% of the estimated US Census population.

METHODS

Study Design

This study compared data, demographic structures, and characteristics from a large commercial research database, the HealthCore Integrated Research Database (HIRD), which is notable for its size and geographic breadth, with data from the 2009 US Census. To create a basis for the approximation of counts relative to the US Census data, standard statistical procedures incorporating a suitable alternative to the goodness-of-fit method were used to establish weights for the HIRD. The weighting formulation was then tested with a sample of patients from the northeast region of the United States who were diagnosed with acute coronary syndrome (ACS).

Data Source

HIRD

This study utilized a large commercial administrative claims database, the HIRD, which contains a broad spectrum of medical, pharmacy, and laboratory information on more than 46 million enrollees in 14 geographically dispersed managed care plans across the United States. The broad range of service models encompassed by these plans includes health maintenance organizations, point of service, preferred provider organizations, and indemnity plans. The data queried from the HIRD were categorized into geographic regions matching those used by the US Census Bureau.

US Census

The US Census Bureau publishes the American Community Survey results every year. The American Community Survey reports population numbers in categories including age, gender, race, and geographic region. No disease prevalence and other types of healthcare utilization information are collected by the American Community Survey. This study was conducted prior to the official release of the 2010 US Census data; as a result, population estimates from the US Census Bureau's 2009 American Community Survey were used for the total count of individuals residing in the 50 US states.

Researchers had access to limited patient data in this study. Strict measures, in compliance with the 1996 Health Insurance Portability and Accountability Act (HIPAA), were observed to ensure the preservation of patient anonymity and confidentiality throughout. The study did not involve the collection, use, or transmittal of individually identifiable data. It was conducted under the Research Exception provisions of the Privacy Rule, 45 CFR 164.514(e); institutional review board sanction was not indicated.

Inclusion Criteria/Exclusion Criteria

Health plan members within the HIRD who had at least 1 day of health plan enrollment between January 1, and December 31, 2009, were eligible for inclusion in the study. This interval was selected because it represented the most current US Census Bureau's American Community Survey data release available at the time of the study. Patients with ACS were selected to perform the projection demonstration. The disease was identified with *International Classification of Diseases, Ninth Revision, Clinical Modification* (*ICD-9-CM*) codes 410.x1 and 411.1x in the claims database.

Statistical Analysis

Goodness-of-fit statistics was not applicable to match the samples because even small differences would appear to be statistically significant because of the large sample sizes. An alternative interpretation of the fit approach was used to examine the overall fit of the lines—census-defined regions, gender, and age groups—characterizing the HIRD and the US Census data. Statistical analyses were conducted with SAS version 9.2 (SAS Institute Inc, Cary, North Carolina).

Weights

Standard statistical procedures, comprising of an alternative version of goodness-of-fit statistics, were used to establish weights for the HIRD, to facilitate the approximation of counts relative to the US Census data**.** The linear weighting was computed as the percentage of the overall population divided by the percentage within the HIRD. Weighting schemes enable the projection from smaller known samples to larger populations in which the desired prevalence rate and other target information are not known. By using weights in a linear model along with specific variables, it is possible to make projections to the larger population by employing the relevant attributes of smaller population.[17] This equation yielded a multiplication factor that was used to compute the weighted number of patients within a geographic region, age group, and gender for a specific disease type or drug classi-

Table 3. Weight Calculation[a] for the Northeast Region (example)

Age Group (years)	NORTHEAST REGION					
	% of Total HIRD Population		% of the US Census Population		Weight	
	Male	Female	Male	Female	Male	Female
<5	0.44%	0.42%	0.56%	0.54%	1.2679	1.2746
5-9	0.50%	0.48%	0.55%	0.54%	1.1018	1.1168
10-14	0.53%	0.51%	0.59%	0.55%	1.1003	1.0791
15-19	0.59%	0.57%	0.65%	0.62%	1.1029	1.0936
20-24	0.57%	0.59%	0.61%	0.59%	1.0710	0.9974
25-29	0.52%	0.61%	0.60%	0.58%	1.1479	0.9461
30-34	0.56%	0.63%	0.56%	0.55%	1.0103	0.8833
35-39	0.61%	0.66%	0.58%	0.59%	0.9455	0.8902
40-44	0.67%	0.72%	0.64%	0.66%	0.9594	0.9051
45-49	0.72%	0.80%	0.70%	0.72%	0.9690	0.9035
50-54	0.71%	0.78%	0.67%	0.69%	0.9416	0.8837
55-59	0.62%	0.69%	0.56%	0.60%	0.9049	0.8661
60-64	0.51%	0.58%	0.47%	0.53%	0.9182	0.9124
65-69	0.27%	0.29%	0.33%	0.40%	1.2571	1.3722
70-74	0.14%	0.17%	0.25%	0.30%	1.7079	1.8468
75-79	0.11%	0.13%	0.19%	0.27%	1.8038	2.0603
80-84	0.07%	0.10%	0.15%	0.24%	2.0878	2.4174
≥85	0.05%	0.11%	0.11%	0.27%	2.1032	2.4972

HIRD indicates HealthCore Integrated Research Database.

[a] $0.9690 = \frac{0.7019\%}{0.7243\%}$

Figure 1. Age Group Distribution Comparison: HIRD vs US Census (all regions)

HIRD indicates HealthCore Integrated Research Database.

Table 4. Projected Number of ACS Patients in the Northeast From HIRD to US Census Population (example)

Age Group	NORTHEAST REGION					
	HIRD ACS Population		US REPRESENTATIVE MATCHED SAMPLE SIZE TO HIRD Projected ACS Population		US CENSUS Projected ACS Population	
	Male	Female	Male	Female	Male	Female
<5	0	2	0	3	0	53
5-9	1	2	1	2	23	46
10-14	2	3	2	3	46	67
15-19	8	6	9	7	183	136
20-24	16	8	17	8	356	166
25-29	19	17	22	16	453	334
30-34	51	30	52	26	1069	550
35-39	126	49	119	44	2472	905
40-44	239	147	229	133	4758	2761
45-49	452	246	438	222	9089	4612
50-54	715	395	673	349	13,971	7244
55-59	877	411	794	356	16,469	7387
60-64	945	488	868	445	18,005	9240
65-69	596	347	749	476	15,547	9880
70-74	494	359	844	663	17,508	13,758
75-79	474	391	855	806	17,742	16,717
80-84	379	354	791	856	16,420	17,758
≥85	279	444	587	1109	12,177	23,008

ACS indicates acute coronary syndrome; HIRD, HealthCore Integrated Research Database.

fication (**Table 1**). On the basis of these distributions, weights were calculated to adjust for any differences in gender, geographic region, and age distributions observed between the HIRD population and the US census population estimates.

RESULTS

Study Populations and Demographic Comparison

During the 2009 calendar year, the HIRD included 14.8 million enrollees, and the US Census Bureau's 2009 American Community Survey data projected in excess of 307.7 million individuals, within an estimated accuracy of 0.1% (margin of error: 0.001)[18] who were used as a base populations in this study. The HIRD population was similar to the US Census estimates in gender distribution, with females comprising 49.8% and 49.3% of their totals, respectively. Relative to the US Census estimates, the HIRD population appeared overrepresented in the midwest and underrepresented in the south. The HIRD population closely matched US Census estimates for the northeast and west regions, differing by only 1% in the northeast and 0.6% in the west (**Table 2**).

Age Distribution Comparison

The age group distributions of the HIRD and US Census populations are shown in **Figure 1**. The HIRD population had relatively higher representation of age categories between 30 and 59 years; it was underrepresented in the age categories <18 years and ≥65 years relative to the US Census. Although there was close agreement between the 2 populations for ages 5 to 30 years and 55 to 70 years, the overall age group of 18 to 64 years is overrepresented in the HIRD.

Weight Computation Based in the Northeast Region

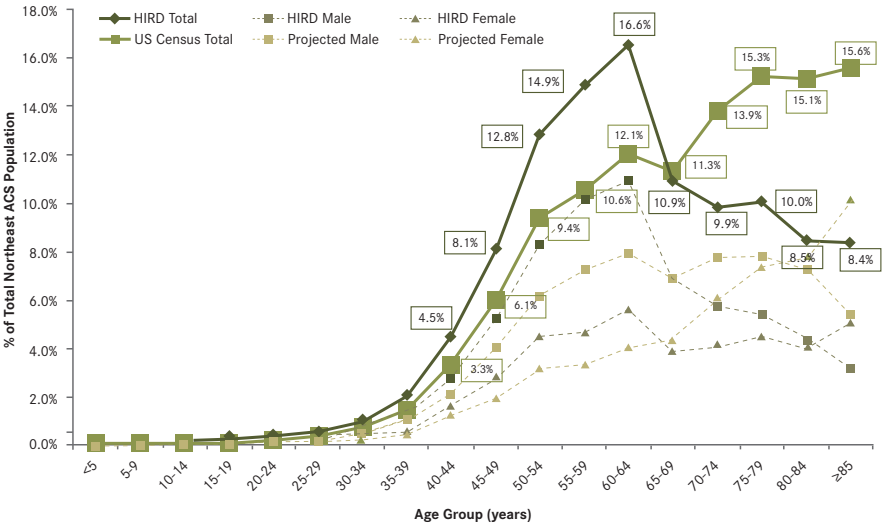
To demonstrate the weight computation model, weight calculations were applied to the northeast region. In 2009, approximately 0.70% of the US Census population was male, aged 45 to 49 years, and lived in the northeast, while around 0.72% of the HIRD population shared the same geographic region, gender, and age characteristics**.** Thus, the weight for the male population aged 45 to 49 years living in the northeast during that time period was 0.9690% (**Table 3**).

Projection of ACS Patients in the Northeast

Table 4 reports the results of weighting the number of HIRD patients with ACS in the northeast region within each age and gender stratum and projecting to the northeast US Census Bureau population. The HIRD had a total of 452 male members

from the northeast region, aged 45 to 49 years, who had at least 1 claim with a diagnosis for ACS from January 1 to December 31, 2009. On the basis of the weight for this population group (0.9690), the projection of ACS diagnosis in a representative sample the same size as the HIRD repository (~4.82% of the US population) would be 438 patients (452 × 0.9690) and 9089 in the overall US Census Bureau population (438 ÷ 4.82%). Application of this weighting scheme results in a greater proportion of ACS patients in the ≥65 years age category and a smaller proportion of patients aged 30 to 64 years relative to the original HIRD estimate (**Figure 2**).

Figure 2. Age Group Distribution of ACS Patients in the Northeast Region: HIRD vs Projected Population (example)



ACS indicates acute coronary syndrome; HIRD, HealthCore Integrated Research Database.

DISCUSSION

While healthcare data are hardly abundantly available for the entire US population, a considerable volume may be found in veritable data silos such as institutional disease registries and the transactional databases of health plans, among other repositories. For healthcare planners and budget directors, access to plausible population estimates is crucial for decision making. One avenue for population level figures for health budget projections and allocations is to extrapolate from smaller data collections. It is essential to have robust and reliable weighting tools, which are capable of achieving the low margin of error requirements, necessary for such projections.[19,20] Driven by this need, this study developed a simple weighting tool to project heath plan data to estimate prevalence rates at the national level.

Although the HIRD represents a population that is slightly less than one-twentieth (~4.82%) of the US population—as represented by the 2009 US Census Bureau count—the data in the HIRD are remarkably representative of the entire US population. HIRD data trended in parallel with the US Census data on gender distribution, regional distribution in the northeast and west regions, but as was expected, it was overweight in the 30 to 59 years age category because the repository consists largely of employer-insured working age people.

Reflecting the source of the majority of people represented in the HIRD repository—enrollees of employer sponsored commercial healthcare insurance—the population aged ≥65 years appears to be relatively underrepresented. Still, the HIRD contains a sizable sample of ≥65-years-old enrollees who may be receiving commercial employer sponsored health benefits, or Medicare advantage, supplement, or Part D benefits. The sample size of this population is substantive enough to allow the application of this weighting methodology to extrapolate the data into the overall US population with statistically acceptable variance.

The weighting methodology developed in this study was tested on the ACS patients from northeast region as an illustration of how the weighting scheme may be applied in practice. While this example specifically addressed ACS patients, it demonstrated how the number of patients in the overall US population for any disease may be estimated from commercially derived healthcare data repositories like the HIRD. This study essentially demonstrated that by using a linear weighting methodology that accounts for differences in geographic regions, age, and gender between an accessible database and the US Census data, it was possible to estimate the prevalence of a number of important healthcare factors. Among areas that may be evaluated using this approach are disease prevalence, healthcare resource utilization, treatment patterns for therapies of interest, and current and potential use of pharmaceutical agents and other treatment modalities.

One of the key objectives of this study was the development of a projection method and a weighting scheme that could be applied to a range of disease conditions and therapeutic categories for which data were available in a repository—such as the HIRD. An important strength of this approach is that it allows for adjustments in the variables or for the updating of estimates of interest with the most current or different data as needed.

Weighed estimations have important planning, resource allocation, and cost management implications for a variety of stakeholders including patients, providers, and payers who have to make decisions based on research results, disease prevalence, treatment availability, and drug utilization, among other factors.

Limitations

The results of the weighting scheme and ACS projection example discussed in this study must be viewed against some important limitations. This study relied on secondary data from commercial health plans across the United States. These data may have some relevance to similar commercial health plans, but only limited external validity

for different patient populations such as the US Medicaid and Medicare programs. In addition, administrative claims lack data on race, ethnicity, and risk factors capable of influencing outcomes. Administrative claims data are prone to over- and underestimations (eg, for patients, disease, medication use, other areas) because of basic assumptions about index events, inability to capture and account for all treatments received by patients, and basic coding and clerical errors. Furthermore, extrapolation was done beyond the point of observable data, contravening a standard requirement of statistical methodology, and likely impacting the robustness of the results. In addition, notable differences existed between the values in the HIRD commercial database and the US Census data. The weights were calculated on the basis of 2009 ACS projections, not official US Census counts.

CONCLUSIONS

Consistent with its commercial employment origins and characteristics, the HIRD repository, while representative of US Census data, was overweighting the 30-to-59 years category. The age groups ≥65 years were underrepresented in the HIRD but still accounted for a substantial sample size. While extrapolations beyond observable data have statistical limitations, in the absence of data on disease prevalence and treatment for the US population as a whole, commercial databases could be viable for projecting patient counts within US Census parameters. This could be invaluable to key stakeholders such as healthcare planners, policy makers, and payers.

Acknowledgments: Bernard B. Tulsi, MSc, provided writing and other editorial support for this manuscript. The authors wish to thank Chaozheng Yang, MS, former research analyst at HealthCore, Inc, for contributions to the study's design and data analysis.

Author Affiliations: HealthCore, Inc (TW, BW, OT), Wilmington, DE; AstraZeneca Pharmaceuticals LP (JWY), Wilmington, DE.

Funding Source: Funding for this research project was provided by AstraZeneca Pharmaceuticals LP.

Author Disclosures: Drs Wasser and Tunceli and Mr Wu are employees of HealthCore, Inc, a wholly owned research and consulting subsidiary of Anthem, a national health insurance company. Dr Yčas was formerly employed by AstraZeneca Pharmaceuticals LP, which provided funding for this study.

Authorship Information: Concept and design (TW, OT, GW, JY); acquisition of data (TW, OT, GW, JY); analysis and interpretation of data (TW, OT, GW, JY); drafting of the manuscript (TW, OT, GW, JY); critical revision of the manuscript for important intellectual content (TW, OT, GW, JY); statistical analysis (TW, OT, GW, JY); provision of study materials or patients (TW, OT, GW, JY); obtaining funding (TW, OT, GW, JY); administrative, technical, or logistic support (TW, OT, GW, JY); and supervision (TW, OT).

Address correspondence to: Thomas Wasser, PhD, MEd, HealthCore, Inc, 123 Justison St, Ste 200, Wilmington, DE 19801. E-mail: twasser@healthcore.com.

REFERENCES

1. Last JM, ed. *A Dictionary of Epidemiology*. 4th ed. New York, NY: Oxford University Press; 2000.

2. Thacker SB. Epidemiology and public health at CDC. *MMWR*. 2006;55(suppl 2):3-4.

3. McKenna MT, Zohrabian A. U.S. burden of disease--past, present and future. *Ann Epidemiol*. 2009;19(3):212-219.

4. Terris M. The Society for Epidemiologic Research (SER) and the future of epidemiology. *Am J Epidemiol*. 1992;136(8):909-915.

5. Terris M. The Society for Epidemiologic Research and the future of epidemiology. *J Public Health Policy*. 1993;14(2):137-148.

6. Thacker SB, Dannenberg AL, Hamilton DH. Epidemic intelligence service of the Centers for Disease Control and Prevention: 50 years of training and service in applied epidemiology. *Am J Epidemiol*. 2001;154(11):985-992.

7. Mehta P, Antao V, Kaye W, et al. Prevalence of amyotrophic lateral sclerosis - United States, 2010-2011. *MMWR*. 2014;63(7):1-13.

8. Adams DA, Jajosky RA, Ajani U, et al. Summary of notifiable diseases. *MMWR*. 2014;61(53):1-121.

9. Chini F, Pezzotti P, Orzella L, Borgia P, Guasticchi G. Can we use the pharmacy data to estimate the prevalence of chronic conditions? a comparison of multiple data sources. *BMC Public Health*. 2011;11:688.

10. Choy M, Switzer P, De Martel C, Parsonnet J. Estimating disease prevalence using census data. *Epidemiol Infect*. 2008;136(9):1253-1260.

11. Costa MA, Huang SS, Moore M, Kulldorff M, Finkelstein JA. New approaches to estimating national rates of invasive pneumococcal disease. *Am J Epidemiol*. 2011;174(2):234-242.

12. Guzmán Herrador BR, Aavitsland P, Feiring B, Riise Bergsaker MA, Borgen K. Usefulness of health registries when estimating vaccine effectiveness during the influenza A(H1N1)pdm09 pandemic in Norway. *BMC Infect Dis*. 2012;12:63.

13. Hanson LA, Zahn EA, Wild SR, Dopfer D, Scott J, Stein C. Estimating global mortality from potentially foodborne diseases: an analysis using vital registration data. *Popul Health Metr*. 2012;10(1):5.

14. Saaddine JB, Honeycutt AA, Narayan KM, Zhang X, Klein R, Boyle JP. Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United States, 2005-2050. *Arch Ophthalmol*. 2008;126(12):1740-1747.

15. Wendt JK, Symanski E, Du XL. Estimation of asthma incidence among low-income children in Texas: a novel approach using Medicaid claims data [published online September 28, 2012]. *Am J Epidemiol*. 2012;176(8):744-750.

16. Zaher C, Goldberg GA, Kadlubek P. Estimating angina prevalence in a managed care population. *Am J Manag Care*. 2004;10(11 suppl):S339-S346.

17. Bethlehem JG, Keller WJ. Linear weighting of sample survey data. *Journal of Official Statistics*. 1987;3(2):141-153.

18. American Community Survey multiyear accuracy of the data (3-year 2008-2010 and 5-year 2006-2010). US Census Bureau website. http://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/MultiyearACSAccuracyofData2010.pdf. Published 2011. Accessed August 13, 2015.

19. Merrill RM, Capocaccia R, Feuer EJ, Mariotto A. Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *Int J Epidemiol*. 2000;29(2):197-207.

20. Nacul LC, Soljak M, Meade T. Model for estimating the population prevalence of chronic obstructive pulmonary disease: cross sectional data from the Health Survey for England. *Popul Health Metr*. 2007;5:8.