Combining Information from Multiple Data Sources : Challenges and opportunities

Trivellore Raghunathan (Raghu)

Survey Research Center

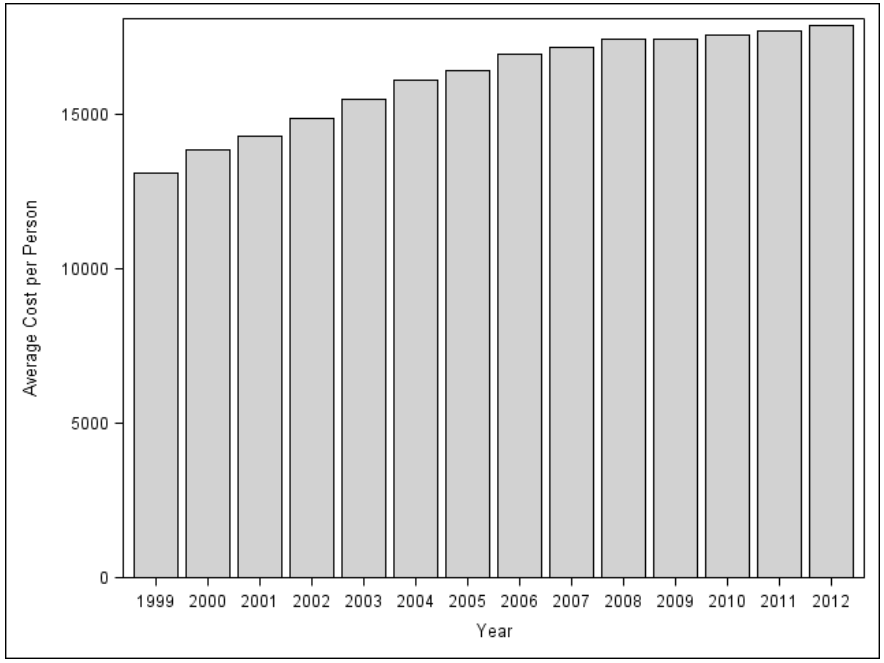
University of Michigan

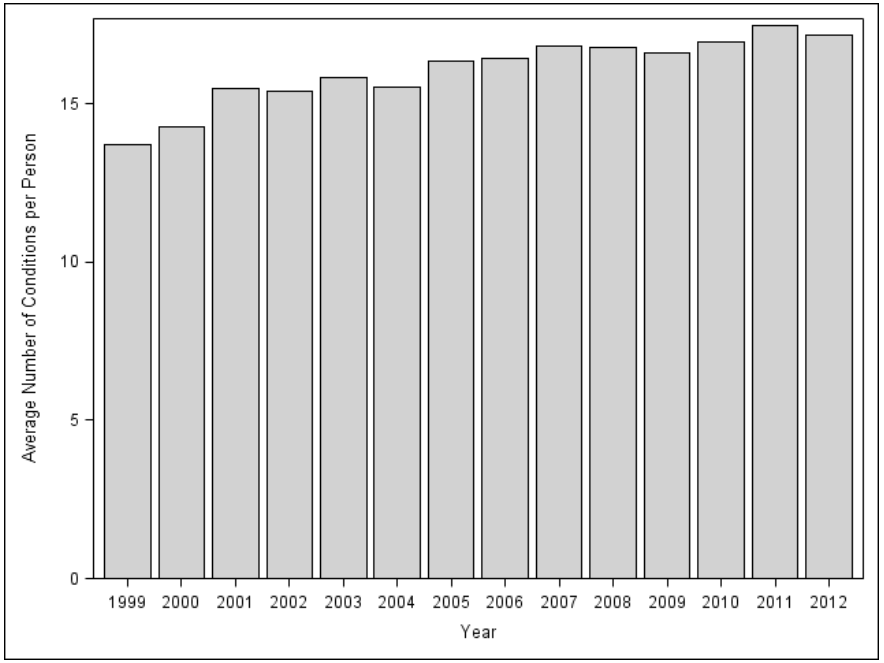
Joint work with David Cutler (Harvard), Susan Stewart (NBER), Kaushik Ghosh (NBER), Kassandra Messer (UM), Irina Bondarenko (UM), Pat Berglund (UM), Paul Imbriano (UM)…

Opportunities

- Digital revolution, though old, has become an important source due to computational ability and cheap storage
- Social media, credit card transactions, purchasing, electronic health records, banking data, real estate, etc. are becoming accessible non-survey data sources
- Survey data based on probability samples for policy research is facing challenges
  - Declining response rates
  - Increasing costs
- Not able to collect all the information needed
- Leverage data from multiple sources to address important problems

Trends in Average Cost and Number of Health Conditions (65 years of age or older)





Three Objectives

1. Estimate prevalence rates and assess trends for various diseases/screening
2. Estimate costs attributable to each disease and assess trends in these costs
3. Dissect the change in the total cost
    1. Attribute to the change in prevalence rate
    2. Attribute to the change in cost of treating the health conditions

Population and Data Sources

- Four age groups: >=65, 45-64, 18-44 and <=17
- Survey Data: MCBS, MEPS, NHIS, NHANES, HRS, PSID, NCS
- Non Survey Data: Medicare Claims, Provider data, IMS, HMO, Prescription prices
- Information from Clinical Studies
- Identified about 120 disease/screening conditions (Health Conditions)

Primary Data Source (Age 65 and older)

- Medicare Current Beneficiary Survey (MCBS)
  - Age 65 and older
  - Years 1999-2010 (2012)
- 107 diseases and screening dummy variables
- Community dwelling and Institutionalized (nursing home, assisted living) populations
- Purely covered on Medicare
- Adjustments
  - Propensity score weighting to compensate for excluding HMO enrollees
  - Multiplier to cost so that weighted estimated population total agrees with published national health expenditure
  - All costs are in 2010 dollars

Objective 1: Estimation of Prevalence

- 107 health conditions: Ever having this condition; some during the specific time period
  - Option 1: Use the Medicare claims (any claim) indicating particular ICD-9 codes
    - The prevalence rates based on this definition:
    - Reasonable for some chronic diseases
    - Low rate for acute conditions and some chronic diseases
  - Option 2: Calibrate the claims using benchmark data
    - Self-report from the National Health and Nutritional Examination Survey (NHANES)
    - All calibrated claims can be thought of as "Ever Having Disease"

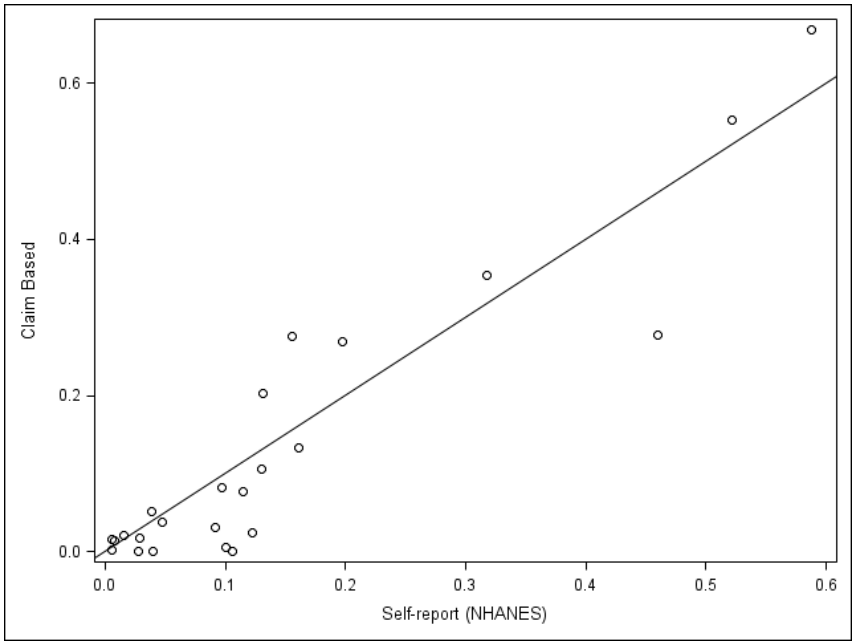
Prevalence of Health Conditions

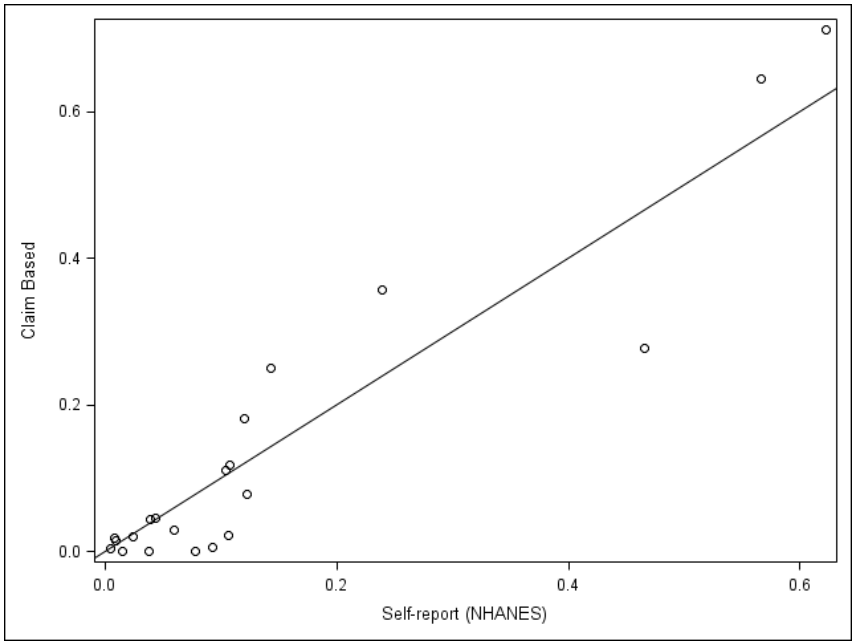
Age group=65+, Year=2001

Disease	SR NHANES	SR MCBS	Claims MCBS
Hyperlipidemia	43.81 (2.15)		35.97 (0.68)
Hip Fracture	3.51 (0.82)	3.71 (0.19)	1.12 (0.12)
Asthma	9.29 (1.27)		4.3 (0.2)
Diabetes	17.9 (1.1)	18.9 (0.5)	18.6 (0.6)
Hypertension	55.9 (1.6)	59.6 (0.6)	47.9 (0.8)
Thyroid Disorders			13.90 (0.5)
Depression			4.69 (0.3)
Dermatologic Diseases			26.66 (0.63)

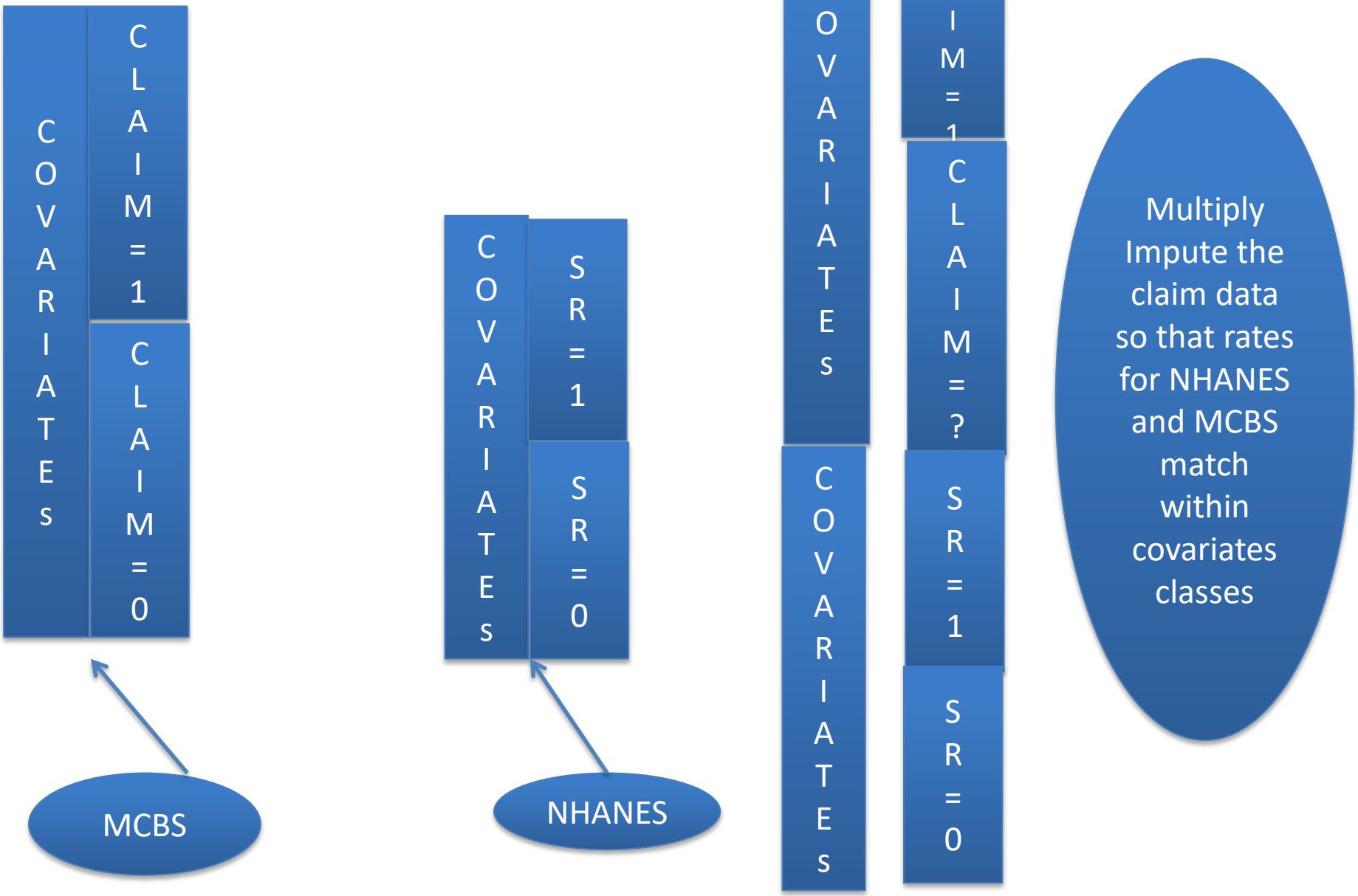
Claim-based disease definitions utilized AHRQ, CCS, and ICD-9-CM

A Scatter plot of self-report and Claim-based prevalence rates for 2005 and 2012





Schematic Display



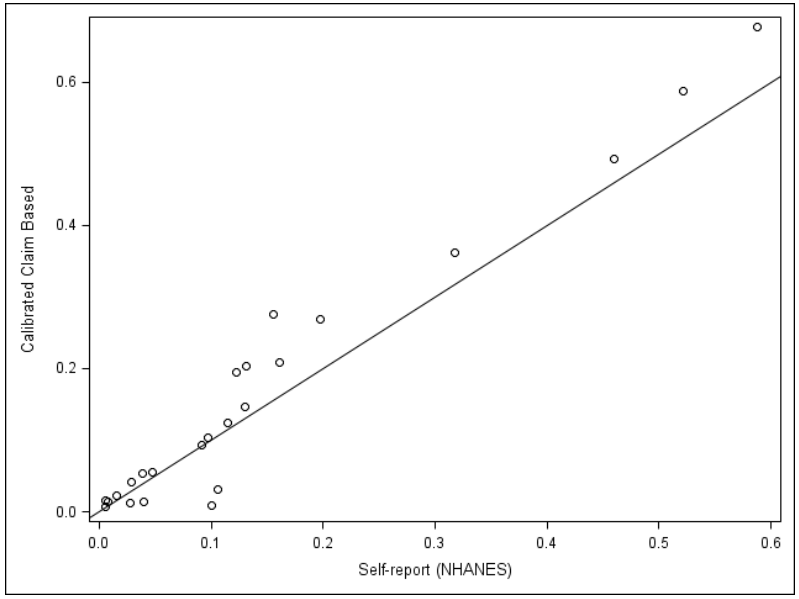
Calibration and Analysis

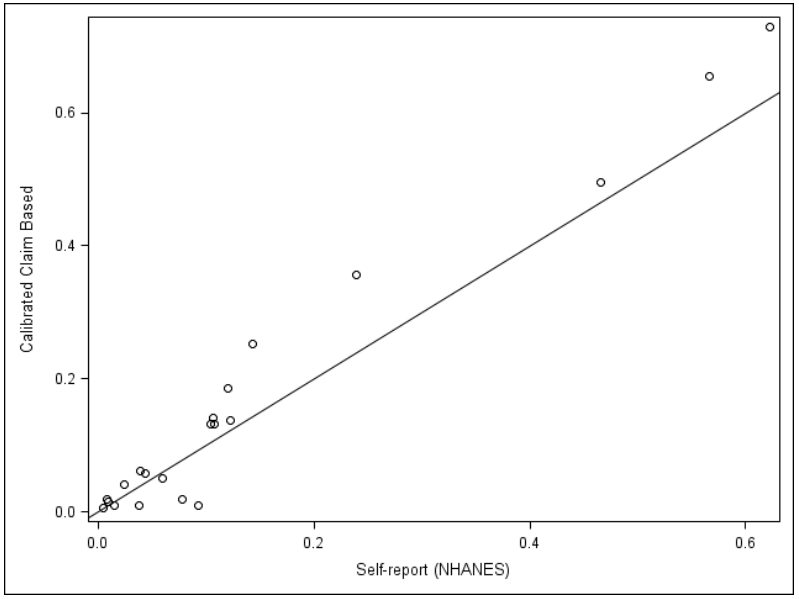
- For diseases with no self-report
  - Construct a measurement error model relating claim and calibrated claim
  - Impute calibrated claims based on this model
- Calibration carried out for each year, separately for Community and Institutionalized populations
- Five imputed data sets with calibrated claims
- All other missing covariates were also imputed
- Obtained the prevalence rates for each disease and year
- Performed a trend analysis using a hierarchical model (random intercepts and slope)
- Performed numerous model diagnostics

Estimated Prevalence of Select Cardiovascular Diseases and Risk Factors for Participants 65 Years and Older: NHANES 2009-2010 and MCBS 2009

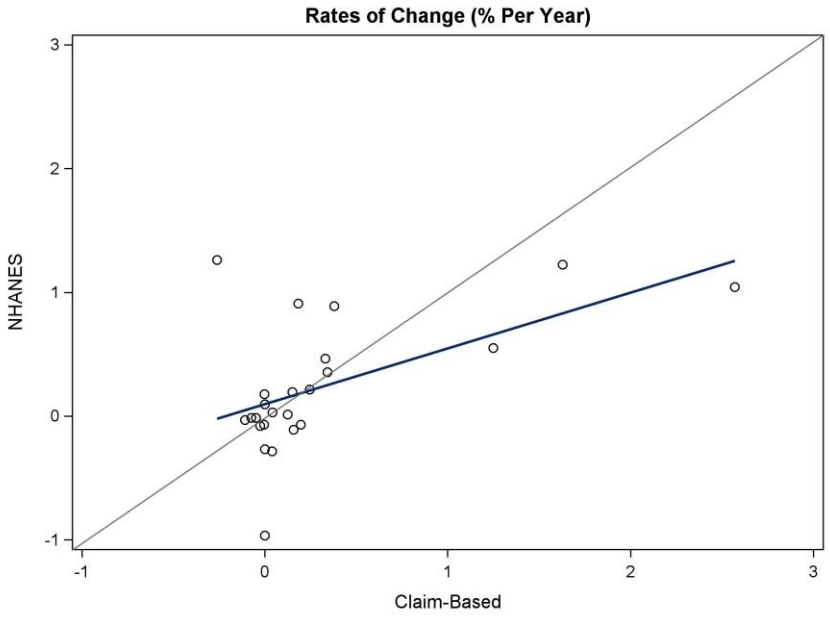
Medical Condition	SR NHANES	SR MCBS: Community (Not Used in the analysis)	MCBS: Claims	Calibrated Claims
Diabetes Mellitus	23.72 (1.38)	23.90 (0.64)	32.15 (0.67)	32.15 (0.67)
Undiagnosed Diabetes Mellitus	2.34 (0.58)			2.04 (0.27)
Hyperlipidemia	51.35 (2.33)	52.31 (1.15)	61.36 (0.89)	62.43 (1.44)
Undiagnosed Hyperlipidemia	1.71 (0.84)			1.64 (0.35)
Hypertension	63.59 (1.75)	69.21 (0.85)	68.50 (0.87)	71.41 (1.49)
Undiagnosed Hypertension	3.17 (1.35)			2.55 (0.57)
Acute myocardial infarction (AMI)	8.58 (0.73)	13.58 (0.55)	2.30 (0.19)	11.19 (1.08)
Acute hemorrhagic stroke			0.64 (0.09)	1.17 (0.21)
Ischemic stroke			5.11 (0.35)	8.12 (0.54)
Any stroke	8.18 (1.03)	11.40 (0.51)	5.40 (0.36)	8.62 (0.58)

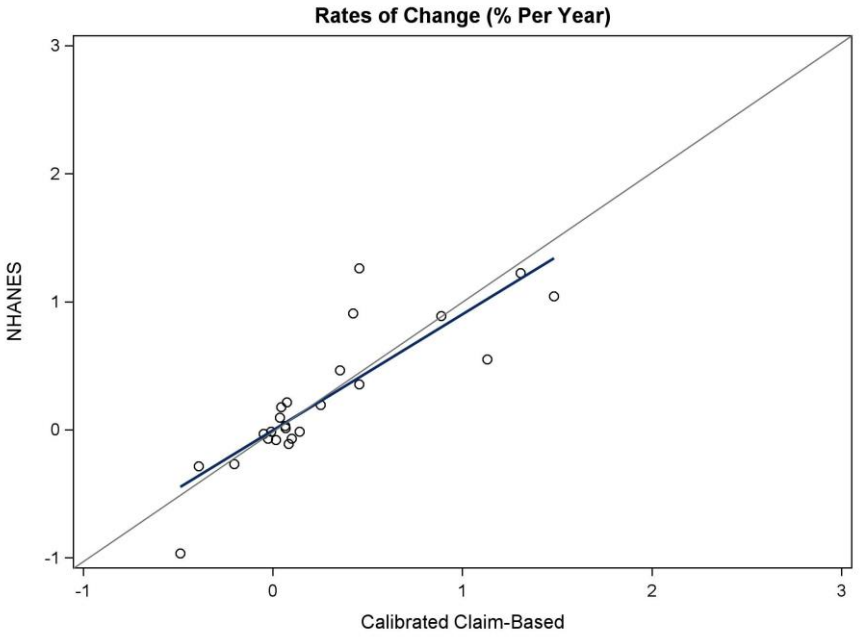
A Scatter plot of self-report and Calibrated Claim-based prevalence rates for 2005 and 2012





Scatter plots of Trend Estimates from Self-report, Claim-Based and Calibrated Claim Based Prevalence Rates





Objective 2: Cost Attribution

- Attributable cost estimated as the difference between those with and without a particular disease other things (covariates and all other diseases) being equal

$$D_j = Disease$$

$$D_{(-\,j)} = Other\ Diseases$$

$$X = Covariates$$

$$Y = Total\ Cost$$

$$A_j = E(Y \mid D = 1, X, D_{(-\,j)}) - E(Y \mid D = 0, X, D_{(-\,j)})$$

Outline of Methods

- A logistic regression model to predict disease dummy variable with covariates and other disease dummy variables as predictors
- Propensity score used to create strata
- Mean difference in the cost for those with and without the disease was computed in each strata
- The weighted average of these differences was defined as the attributable cost for the disease
- Computed attributable cost for all 80 diseases and for all 12 years 1999-2010

Cost Model

- Aggregated individual level cost computed by adding attributable costs for individual level diseases

$$A_j = \textit{Attributed cost for Disease } j$$

$$D_{ij} = 1 \textit{ if subject } i \textit{ has disease } j \textit{ and } 0 \textit{ otherwise}$$

$$Ag.C_i = \sum_{j=1}^{80} A_j D_{ij}$$

- The Aggregated costs and the actual cost may not agree as the cost depends upon several other factors such as hospital stays, number of conditions etc

Cost Model (Contd.)

- Regression model adjustment to predict actual cost

$A_j = \textit{Attributed cost for Disease } j$

$D_{ij} = 1 \textit{ if subject } i \textit{ has disease } j \textit{ and } 0 \textit{ otherwise}$

$$Ag.C_i = \sum_{j=1}^{80} A_j D_{ij} \text{ (Aggregated Cost)}$$

$Ac.C_i = \textit{Actual Cost}$

$$Ac.C_i = Ag.C_i \left[\beta_0 + \sum_k \beta_k X_{ik} \right] + \varepsilon_i$$

$X_1 = \textit{Number of Health Conditions}$

$X_2 = \textit{Number of Health Conditions squared}$

$X_3 = \textit{Dummy variable for no inpatient stays}$

$X_4 = \textit{Number of inpatient stays}$

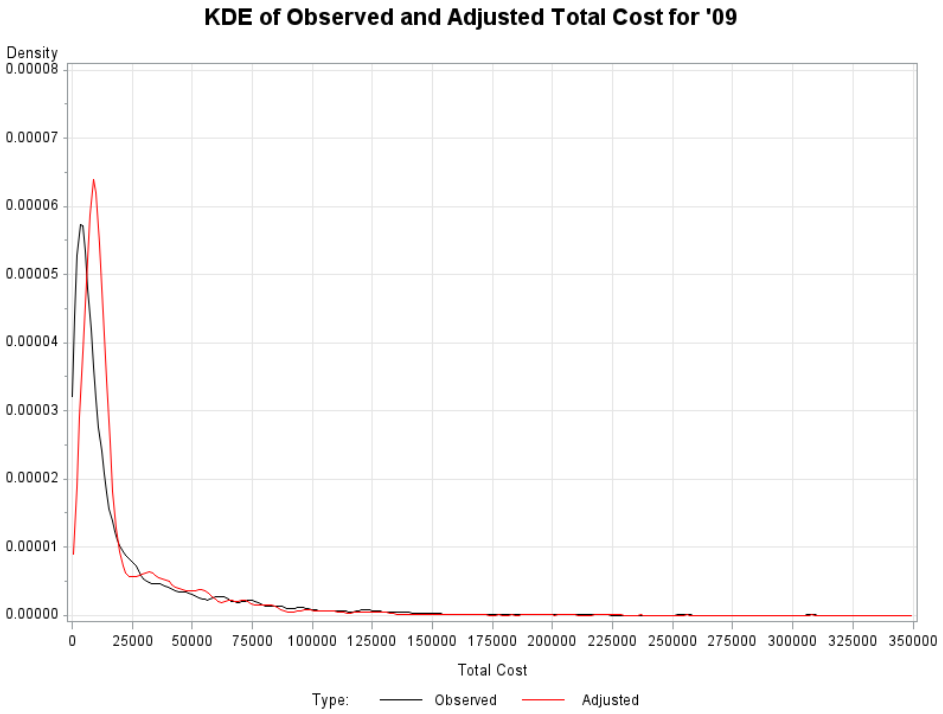
$X_5 = \textit{Number of inpatient nights}$

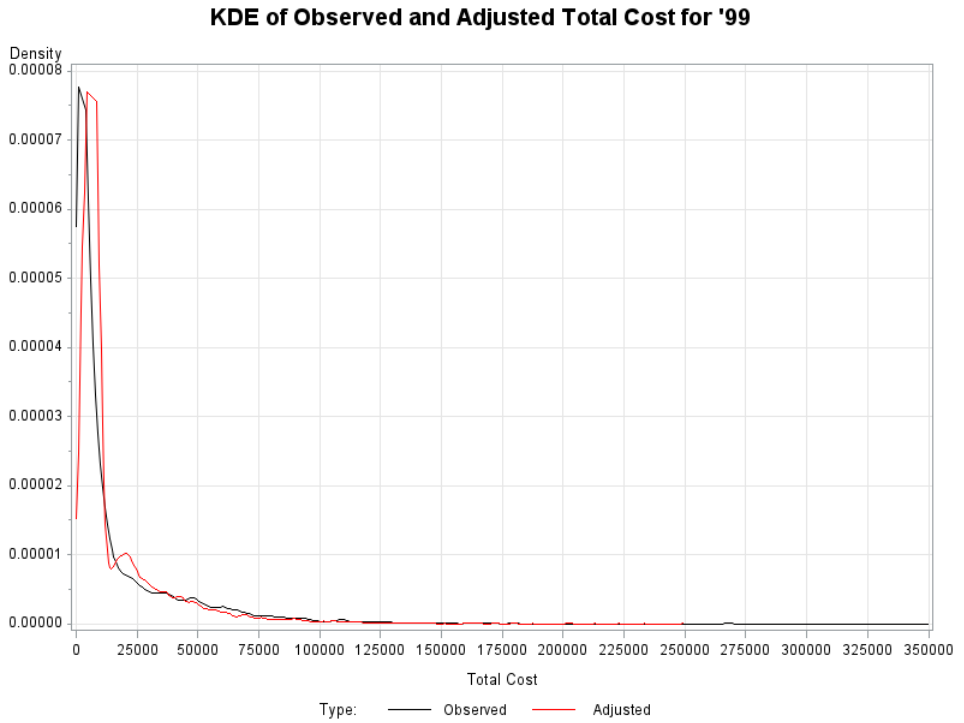
$X_6 = \textit{Dummy for Death during the year}$

$X_7 = \textit{Number of months alive during the year}$

$X_8 = \textit{Number of days institutionalized}$

Comparison of actual and adjusted cost

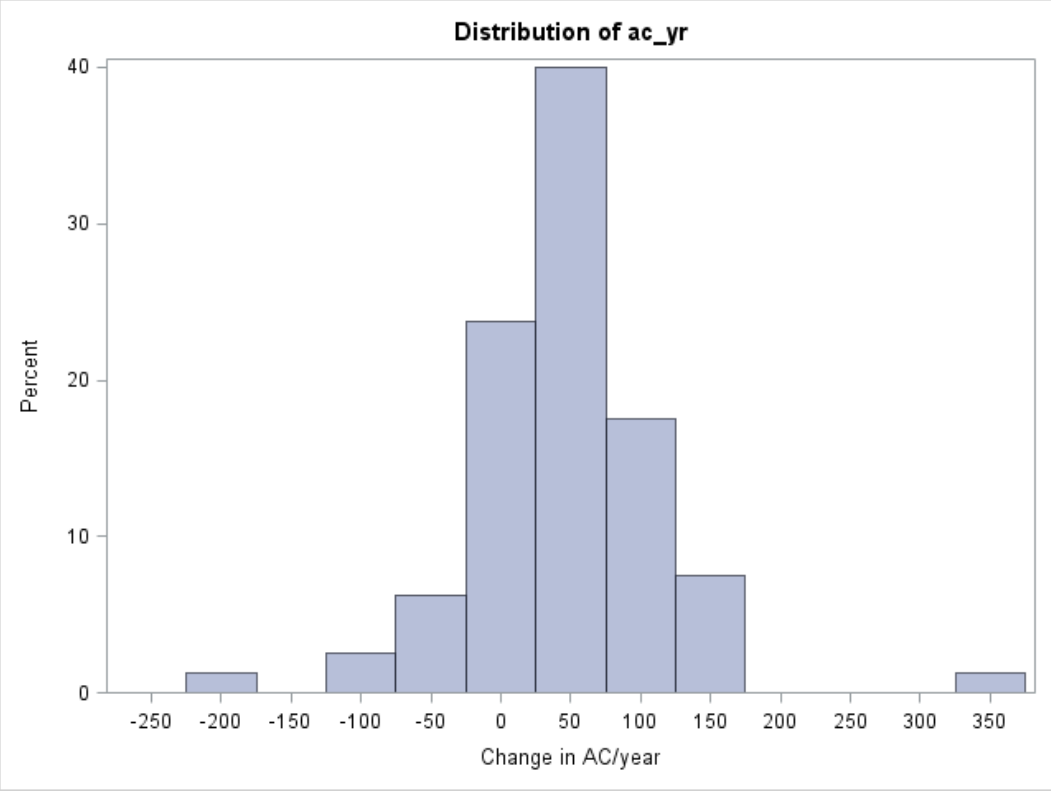




Attributable costs, Prevalence rate of diseases, and 6 covariates are the building blocks for predicting cost using the regression model

Changes in Attributable costs over the 11 year period

- Fitted a hierarchical random effect models for the attributable cost with random intercepts and slopes across the 80 diseases (some diseases were combined due to low prevalence rates)



Objective 3: Cost-Disease Prevalence Dynamics

- Counter factual Cost per person were computed by applying the attributable cost for Year t to the Prevalence rate for Year s with all other covariates remaining the same.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1999	13103	13560	14462	14473	14206	14016	14608	14445	14645	14312	13847	13959	13710	13439
2000	13405	13885	14890	14924	14705	14546	15222	15113	15370	15064	14621	14783	14730	14519
2001	12884	13346	14296	14325	14179	14042	14734	14599	14873	14584	14212	14370	14365	14165
2002	13413	13895	14826	14901	14779	14616	15252	15166	15450	15184	14765	14935	14948	14800
2003	13953	14530	15555	15624	15520	15375	16129	16051	16383	16060	15622	15824	15931	15727
2004	14752	15348	16523	16567	16371	16134	16959	16837	17077	16746	16194	16379	16349	16009
2005	14469	15006	16054	16091	15881	15739	16432	16301	16604	16319	15837	16006	16026	15789
2006	14887	15468	16659	16703	16521	16324	17076	16944	17320	17051	16568	16716	16776	16493
2007	14600	15196	16348	16396	16328	16186	16946	16853	17188	16860	16475	16666	16673	16389
2008	15321	15870	17111	17086	16980	16714	17580	17371	17753	17460	16967	17122	17097	16778
2009	15430	16005	17208	17301	17262	17061	17886	17820	18168	17920	17471	17620	17754	17533
2010	15205	15856	17114	17208	17141	16968	17836	17761	18094	17820	17336	17574	17647	17381
2011	15425	16047	17386	17412	17311	17102	17934	17791	18187	17898	17498	17725	17720	17513
2012	15760	16371	17687	17813	17716	17459	18407	18231	18546	18284	17907	18170	18155	17896

Analysis

Cost Year	Prevalence year				
	1999	2002	2005	2008	2011
1999	$13,103	$14,473	$14,608	$14,312	$13,710
2002	$13,413	$14,901	$15,252	$15,184	$14,948
2005	$14,469	$16,091	$16,432	$16,319	$16,026
2008	$15,321	$17,086	$17,580	$17,460	$17,097
2011	$15,425	$17,412	$17,934	$17,898	$17,720

Average yearly Change	Dollar (SE)	Percent (SE)	
Due to Prevalence	$83 ($12)	0.5% (0.05%)	
Due to Cost	$287 ($10)	1.8% (0.04%)	
Total	$370.00	2.3%	

Average Cost, Change due to Prevalence and Change due to Cost for 7 broad categories of diseases

category	Mean Cost	Change($)/Year due to prevalence	Change($)/Year due to Cost	Percent Change/Year due to Prevalence	Percent Change/Year due to Cost
Cancer	588	5.38 (0.38)	0.67 (1.38)	0.90 (0.06)	0.16 (0.26)
Chronic and Disabling Conditions	1702	1.81 (1.34)	31.65 (1.38)	0.15 (0.08)	1.89 (0.08)
Recoverable Acute Conditions	3975	15.45 (2.60)	57.90 (2.64)	0.39 (0.07)	1.47 (0.07)
Non-Fatal Chronic Conditions	2736	29.36 (1.26)	76.32 (2.88)	1.12 (0.05)	3.03 (0.13)
Non-Fatal Acute Conditions	3626	-3.78 (1.72)	76.95 (2.78)	-0.10 (0.05)	2.17 (0.08)
Other Ill-Defined Conditions	3249	35.33 (1.57)	42.02 (1.90)	1.10 (0.05)	1.26 (0.05)
Screening	304	-0.28 (0.49)	-5.36 (2.31)	-0.17 (0.16)	-1.45 (0.65)

Issues

- Differences in the type of respondents and source of responses.
  - (1) Face-to-face interview of respondents reporting on health conditions and (2) Physician reporting about the patients based on medical records.
- Differing modes of data collection: Mail, Telephone, face-to-face or a mix.
- Survey context: Response error properties might differ in the two surveys.
  - (1) Survey may be conducted by a well known Federal Agency
  - (2) Reputed institution, but not that well known.
- Differences in the survey design.
  - (1) NHIS is a face-to-face survey
  - (2) NHANES involves a face-to-face survey as well as measurement/Lab
  - Respondent recalling abilities may differ under these two survey-design settings.
- Differences in the question wording or the placement of the questions with the same wording may provide different stimuli to respondents and, hence, different error properties.
- Combining from a survey (where every respondent receives the same stimuli) and an administrative data source (absence of or unknown nature of stimuli)

Conclusion

- With these challenges, *combining data from a mix of probability and non-probability sources* provides exciting opportunities for the increasing world of "big data," where large quantities of poor or unknown quality data in terms of representativeness and measurement error can be improved with the use of high quality probability sample data
- It is dangerous to think that we do not need high quality probability surveys anymore