Review of NHIS Public-Design Structures

Van Parsons[1], Chris Moriarity[2]
[1]National Center for Health Statistics, Centers for Disease Control, 3311 Toledo Rd, Rm 3219, Hyattsville, MD, 20782
[2]National Center for Health Statistics, Centers for Disease Control

1. Introduction

Since July 1957, the National Health Interview Survey (NHIS), sponsored by the National Center for Health Statistics (NCHS), has been the main governmental survey fielded to assess the health status of U.S. noninstitutionalized civilian population. The NHIS's complex sampling design has been redesigned after each Decennial Census. Two basic frames have been used over the life of this survey. The original and the two redesigns covering 1957-1984 were based on a subdesign of the Census' Current Population Survey, while the three redesigns covering 1985-2006 were based on independent Area and Permit frame designs.

The NHIS primarily collects and reports self-reported categorical health and demographic information. Historically, the aim has been to produce accurate and reliable statistics at the national level. While statistics at subnational geographical levels, e.g., states and large metropolitan areas, are of great interest, regulations imposed by the Public Health Services Act, Section 308(d) requires that publicly released data avoid identification and disclosure risk. This restriction results in limited direct labeling of any geographic entity on any publicly released microdata sets. Furthermore, the nature of the NHIS design is such that the design levels are geographically clustered, and thus, identification and disclosure risks may increase whenever true sampling clusters are released without some degree of masking. Given the advances in record linkage techniques, computer hardware and availability of auxiliary information files, the NCHS public data release policy covering geographical and design structures has become more restrictive in recent times. Table 1 contrasts year 2006 levels of design disclosure with those from 1996 and earlier.

The public data restrictions may have changed somewhat over time, but there are three basic goals in producing an NHIS public-use file:

i.) Allow efficient design-based analyses
ii.) Selectively release design information:
Weights, Strata, Primary Sample Units (PSU)
iii.) Avoid identification and disclosure risks

In this paper we discuss the past and present public-use design structures.

2. Changes in Public-Use Design Structures

Before the late 1980's the public-use data system used by the NHIS was set up for limited usage. This was due to the highly restrictive mainframe computational environment at that time, and limited choices of analytical software. Starting in the late 1980's, a transition began to personal computing and Internet data dissemination. Now there is wide dissemination of the NHIS public data via the Internet and its usage is enhanced by an abundance of analytical software. As a result, limitations on public release have increased.

Prior to the summer of 2007 analysts of historical data were faced with several challanges. For some sample design periods there was more than one public-use design structure available for variance estimation. The historical focus of the NHIS had been on analyses of single years of data, and thus the design structures for combining years of data, along with guidance for software implementation, were not clearly defined.

To remedy this situation NCHS initiated the creation of "standardized" public-use design structures for the NHIS that would:

1. Provide simple structures for maximum usage.

2. Define consistent structures over each "~10-year" sample design period that allow for analysis of pooled data across survey years. Through auxiliary files and/or new documentation, these structures are being implemented to work with all NHIS microdata from fiscal year 1963 to the present. (More details follow below.)

3. Provide standard error estimates "close" to the "best" in-house design structure standard error estimates.

NCHS is providing documentation for these structures, along with specific guidance for using the following complex survey variance estimation software: SUDAAN, SAS Survey Procedures, Stata, SPSS, VPLX, and R (including R's add-on "survey" package). References and detailed discussion for these and several other such complex survey software packages are at the web site *http://www.hcp.med.harvard.edu/statistics/survey-soft/*.

Briefly, the standardized structures are based on previously-developed structures. In some time periods, modifications and/or extensions were implemented. All of the structures consist of Pseudo-Strata, containing at least two Pseudo-PSUs per Pseudo-Stratum. The presence of two or more Pseudo-PSUs per Pseudo-Stratum and the use of a "with replacement" sampling assumption assure that all of the software mentioned above is capable of producing standard error estimates from NHIS data, including subgroup analyses.

Within a sample design period, pooled analyses should treat the data years being pooled as dependent, as the annual samples were drawn from the same geographic areas. Sample cases from a given geographic area, sampled in different years, should be assigned to the same Pseudo-Stratum for variance estimation. The standardized structures assure this. When a pooled analysis crosses a sample design change (e.g., 2005-2006), the samples from the different designs should be treated as independent. Some modification of the Pseudo-Strata variables may be needed. Currently NCHS has provided some guidance for this situation, and plans to provide more guidance in the future. For a pooled analysis that is a combination of "within" and "across" sample designs (e.g., 2004-2006), the years within a sample design (e.g., 2004-2005) should be treated as dependent, and then the chunks of "within" data (e.g., 2004-2005, 2006) should be treated as independent from each other.

From 1997 to the present, the NHIS public use files contain Pseudo-Stratum and Pseudo-PSU codes that can be used directly in variance estimation software. From 1980 to 1996, the NHIS public use files contain sufficient information to create Pseudo-Stratum and Pseudo-PSU codes. For public use files prior to 1985, auxiliary data files will be made available at the NCHS website, following the schedule below, that a data user can download and link to the NHIS public use files to obtain Pseudo-Stratum and Pseudo-PSU codes. (For the 1980-1984 public use files, the variance estimation information on the files is equivalent to the content of the auxiliary files. However, the auxiliary files will contain a Pseudo-Stratum variable, which the user would have to create if working only with the information in the public use files.) As NCHS releases the oldest NHIS microdata files (fiscal year 1963 - calendar year 1969) on the Internet for the first time, auxiliary data files for variance estimation will be released simultaneously. The NHIS data files and documentation can be accessed at the NCHS web site: *http://www.cdc.gov/nchs/nhis.htm*.

Internet Release dates

Data Years	Status
1958[a]-1962[a]	No microdata
1963[a]-1968	Late 2007 or early 2008
1969	Fall 2007
1970-1986	August 2007
1987-2006	Now

[a] Fiscal year (e.g., fiscal year 1958 was July 1, 1957 to June 30, 1958)

3. Public-Use: Masking of the 2006 Design

Duncan and Pearson (1991) give an excellent discussion of approaches to the masking of microdata. For design structures we chose a method that "blurs" original labels by grouping. An original cluster of the design, say $U_0(i) =\{u_{1,i}, u_{2,i},\ldots,u_{k(i),i}\}$, is a collective set of units where each unit in the cluster has a common label, e.g., PSU, county, or Census block. Knowledge of the characteristics within a cluster, e.g., sampled demographic frequencies, may create geographical disclosure risk for those clusters whose public-use geographical structure is consistent with a population standard geographical structure, e.g. cluster defined as county.

We assume that if $U_0(1)$, $U_0(2)$ are two original structural geographical clusters of the design, the location of each possibly at risk of disclosure because of its characteristics, then a combined larger cluster, say, $U_c(1,2)=\{u_{1,1}, u_{2,1},\ldots,u_{k(1),1}, u_{1,2}, u_{2,2},\ldots,u_{k(2),2}\}$, the union of element units, has less geographical disclosure risk than the original components. This masking method is the foundation of for the current public-use NHIS design. Our masking rules and goals for the public data follow.

3.1 General Masking Rules

If $\mathbf{U_0} = U_0(1), U_0(2),\ldots, U_0(k)$ is a set of original sampling clusters, some of which need masking, then the task is to create a reduced set of masked clusters,
$\mathbf{U_c} = U_c(g_1), U_c(g_2), \ldots , U_c(g_m)$, $m < k$, with some or all g_j representing combined clusters. The collapsing should be defined so that for a linear statistic, T, defined equivalently over $\mathbf{U_0}$ and $\mathbf{U_c}$,
$E(\mathit{Var}_c (T \mid \mathbf{U_0})) \approx E(\mathit{Var}_0 (T \mid \mathbf{U_c}))$ where Var_0 and Var_c represent the variance estimators for the original and collapsed clusters, respectively, and E is the true design expectation operator. We seek a collapsing that results in a collapsed-unit-based variance estimator having little bias over that of the original variance estimator.

While unbiasedness of the variance estimator is a desirable quality, it is also important that the variance estimator exhibit stability for a broad range of statistics.

The design-based variance estimators considered are based on squared deviations of cluster totals within strata, and associated with the estimator is a measure of stability, the "degrees of freedom", often coarsely estimated as number of clusters minus the number of strata. Along with the unbiased criterion, we also target a collapsing method that allows for a "large" number of degrees of freedom when considering estimation over the larger domains.

For variance estimation we frequently assume that, at some sampling level, the true complex-design sampling structures can be approximated by a stratified independent sampling of clusters. In such a setting a collapsing strategy can be analytically expressed.

Let $U_0(1)$, $U_0(2)$, ... , $U_0(n)$ be a collection of n (assumed to be even) independently sampled clusters organized by their sampling strata, possibly just one stratum or the concatenation of several strata. Consider a collapsing of these n units into 2 distinct units, one with label "+" and the other with label "-''. Let $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ be a random assignment process independent of the original sampling scheme that collapses by assigning +1 or -1 to each $U_0(i)$, with $P(b_i = 1) = P(b_i = -1) = 1/2$, but with the b_i's dependent by constraining $\sum b_i = 0$. Clusters are collapsed based on the sign of the b_i's;

$U_+ = \{ u_{j,i}$: for index i, $b_i = +1$, $u_{j,i}$ is a unit in $U_0(i)$ $\}$ and

$U_- = \{ u_{j,i}$: for index i, $b_i = -1$, $u_{j,i}$ is a unit in $U_0(i)$ $\}$

Now for a linear estimator, $T = \sum x_i$, summed over the original n units, with each x_i defined over the original $U_0(i)$, and assumed to be an unbiased estimator for some characteristic μ_i, we can define $T = T_+ + T_-$, the sum over the two collapsed units, U_+ and U_- , and a collapsed variance estimator for T is $Var_c\,(T) = (T_+ - T_-)^2$;

$E\,(T_+ - T_-)^2 = E(\sum b_i x_i)^2 = Var(T)$ + collapsing bias, with the collapsing bias = $Var_{\mathbf{b}}(\sum b_i\, \mu_i)$.

The choice of organization of the strata for the original n units and the choice of assignment rule **b** will affect the magnitude of the collapsing bias. In a simple random sample environment the usual single stratum variance estimator has (n-1) degrees of freedom, but the collapsed version has 1 degree of freedom. However, this simple procedure can be applied repeatedly over a partition of a large sample to form a larger set of Pseudo-Strata each with two Pseudo-PSUs, i.e., the two collapsed clusters. The overall degrees of freedom can still be large.

3.2 Applications to NHIS

While not intended to be comprehensive and avoiding the special situations that frequently arise, the following discussion provides the generating ideas to the NHIS design masking methods.

3.2.1 Collapsing within Non-Selfrepresenting(NSR) Strata

For the most part, two PSUs, counties or aggregates of counties, are selected from each NSR stratum using Durbin's sampling method. The joint probabilities of PSU selection are never publicly released, and the public variance estimators treat the sampling as with replacement. In two NSR strata, say A and B, containing two PSUs each, say 1 and 2, the original clusters can be ordered A.1, A.2, B.1, and B.2. The assignment process **b** is defined (b_A, $-b_A$, b_B, $-b_B$) where b_A and b_B are independent. Such a process collapses PSUs across strata with no collapsing bias if the PSUs are truly independent with the same expectation within the true strata.

3.2.2 Collapsing within Selfrepresenting(SR) Strata

The sampling clusters within the SR strata, typically metropolitan areas with large populations, are block-clusters. The NHIS sample size within an SR block-cluster is noticeably smaller than the corresponding sample within an NSR PSU. If the SR sample clusters were directly released, the public user may be able to distinguish between metro and non-metro areas to some extent. To avoid any potential disclosure risks, the SR clusters were targeted for collapsing to sizes that were less distinguishable from that of the NSR PSUs.

The true SR cluster sampling was systematic in nature. Following the method described in Wolter (1985), section (7.2.1), the sample was regarded as a stratified sample with two units drawn independently from successive strata. The nature of the true sampling leads to hypothetical strata consisting of population block-clusters based upon similar geography and minority status; such hypothetical strata were joined by similar regional geography and/or minority status to an aggregate size somewhat consistent with the size distribution of the true NSR strata. For such an aggregated stratum hypothesized to have k substrata with two independent samples each, the assignment process **b** is defined (b_1, $-b_1$, b_2, $-b_2$,…, b_k, $-b_k$), where b_j , j=1,2,…,k, are independent.

The collapsing bias = $\sum (\mu_{j.1} - \mu_{j.2})^2$, the sum over the k substrata. If the units within these hypothetical systematic sampling strata are somewhat homogeneous, then the collapsing bias should be small.

3.2.3 NHIS Public-Use Design Summary

The SR and NSR sample cluster collapsing yields 300 Pseudo-Strata each with a sample of two Pseudo-PSUs, where the sample is treated as sampled with replacement. About every 10 original SR block-clusters get collapsed to 1 Pseudo PSU, and only small sample or apparently at risk NSR PSUs get collapsed. There is much less variation in size of the masked PSUs than that exhibited in the original sampled clusters. The main advantage is the "blurring" of specific geography, but the coarsening of design features also makes the public-use design somewhat robust over time. A drawback may be the loss of *degrees of freedom*, but for broadly dispersed domains that should not be a problem.

4.0 Variance Estimation: Masked Public Use Design versus In-house design

For Public-Use data analyses the design-based options are somewhat limited to imposing the "2 PSUs per STRATUM With Replacement" (referred to as *2PSWR*) structure on the data with an option of Poststratification. For this paper we considered the SUDAAN software (2004) and the open source R software (2007), including R's add-on "survey" package, which does complex design variance estimation.

To reflect the in-house design information, a finer structured Yates-Grundy-Sen 2-stage variance estimator form that included joint probabilities of PSU selection and block-cluster sampling within minority density strata was considered. The SUDAAN software was chosen for analysis. At this time only SUDAAN seems capable of handling such a structure and only for linearization, but SUDAAN has the option of poststratification for means and totals.

Our study should be considered preliminary and limited in scope. We considered person-level means and totals for four variables: activity limitation, no insurance, fair/poor health, and number of doctor visits. These variables were considered over aggregations and intersections of the domains:

Gender :	male, female
Race/Ethnicity:	Hispanic, Black, Asian, Other
Age:	0 to 17 (y), 18 to 44 (1), 45 to 64 (2), 65+ (e)

Census Region: NE, MW, S, W.

(The underlined red highlighted letters are used to label domains in some of our figures.)

Many survey software packages have options on methodology, e.g., linearization, balanced repeated replication (BRR), and on the use of poststratification. Our experience is that most public-use data users treat the supplied final weights as pure sampling weights and use a linearization approach along with the 2PSWR design.

Using the SUDAAN and R survey packages along with their options, we made a limited comparison of the coefficient of variation (CV) for estimated means using the four variables and 79 domains mentioned earlier. Figure 1 treats the SUDAAN 2PSWR design as a standard and plots selective alternative public-use strategy CV's versus the corresponding CV from the SUDAAN 2PSWR design. First, we observed that the R survey package using 2PSWR option gave identical results. The poststratification options for both SUDAAN and R gave results of the same order of magnitude, and in general did not differ too much from those of SUDAAN with 2PSWR design. The R survey package has the ability to generate poststratified replicate weights from a 2PSWR design. For this paper we used the poststratified BRR method in R, and as of this writing, SUDAAN and most other software packages cannot do this. Given that most NHIS statistics are non-linear in nature, a poststratified BRR method may be a best overall choice.

In Figures 2, 3 and 4 we compare some of the public-use structures with a "best" in-house structure for the "no insurance" variable. In Figure 2 we see that both the public-use poststratified and 2PSWR method CV's for the estimated mean follows the in-house CV fairly closely on all domains, with the CV for 2PSWR design, perhaps somewhat more deviant on the larger domains (left side of x-axis). The patterns for CV's for estimated totals are quite different from those of the CV's for means. In Figure 3 one sees that the CV's based on the 2PSWR design tend to be quite a bit larger than the "best" in-house. A poststratified 2PSWR design tracks the in-house CV somewhat more closely.

Figure 4 shows the confidence intervals for the in-house versus the 2PSWR designs. One sees no practical difference on the y-scale presented here.

References

Duncan, G.T., Pearson, R.W., (1991). Enhancing Access to Microdata While Protecting Confidentiality: Prospects for the Future, *Statistical Science*, Vol.6, No.3, pp.219-239.

Wolter, K. M., (1985), *Introduction to Variance Estimation,* Springer: Berlin: New York.

R Development Core Team (2007), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
ISBN 3-900051-07-0, URL http://www.R-project.org.

Research Triangle Institute (2004), *SUDAAN Language Manual, Release 9.0,* Research Triangle Park, NC: Research Triangle Institute.

Table 1. NHIS Design Disclosure on Public-Use Files Past and Present

	Pre-1997	2006
True Strata and Sampled Clusters	Mostly intact, but with random labels	Major masking, Many pseudo-clusters, Random labels
Weights	Base weight, Non-response First-stage adjust, Post-stratification	Pre-poststratification Poststratification
Selection Probabilities	No	No
(Non-) Certainty Regions	Yes	No
Fine-level Geography	Census Region NE, S, MW, W, MSA status	Census Region NE, S, MW, W

Fig. 1 2006 NHIS Public−Use CV=(Stderr/Mean)
Computations for Estimated Means over 4 variables and 79 domains
SUDAAN and R(survey−Lumley) software
Using a 2−PSU per Stratum With−Replacement Design

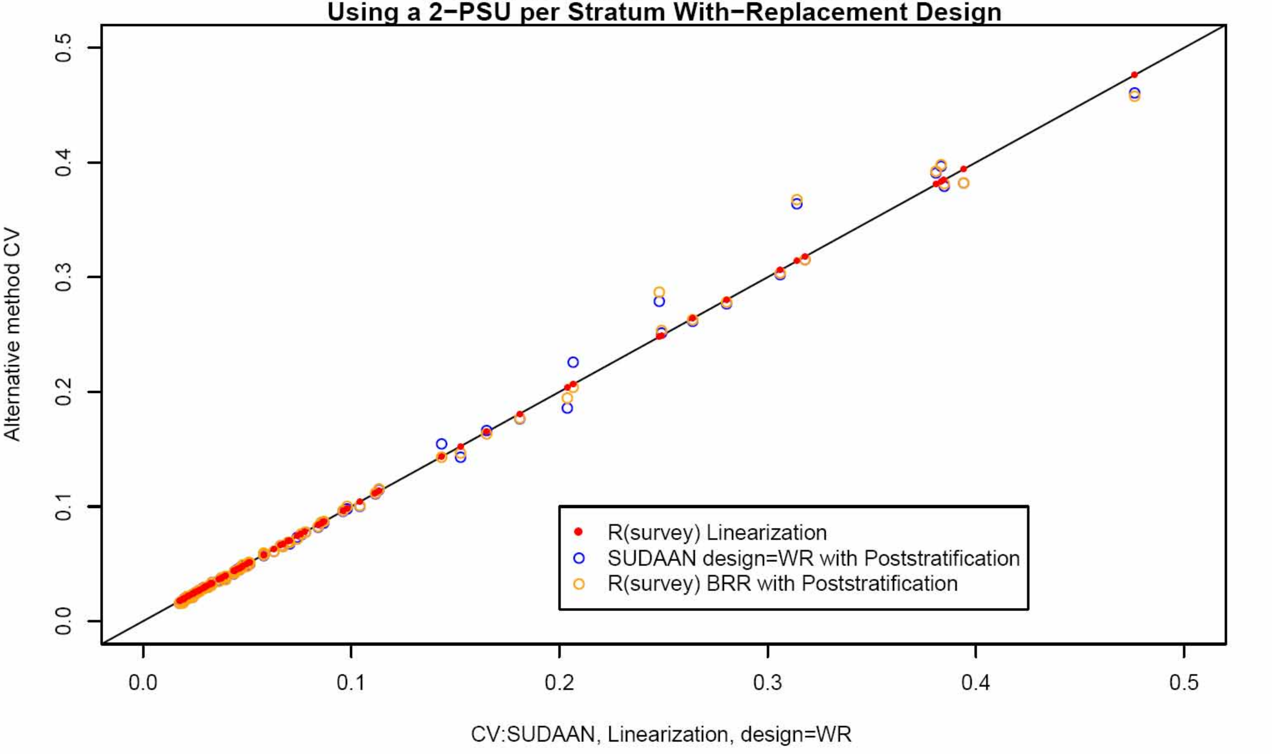


Fig. 2 Estimated CV's for Proportion: Persons with No Health Insurance
for different variance−design structures

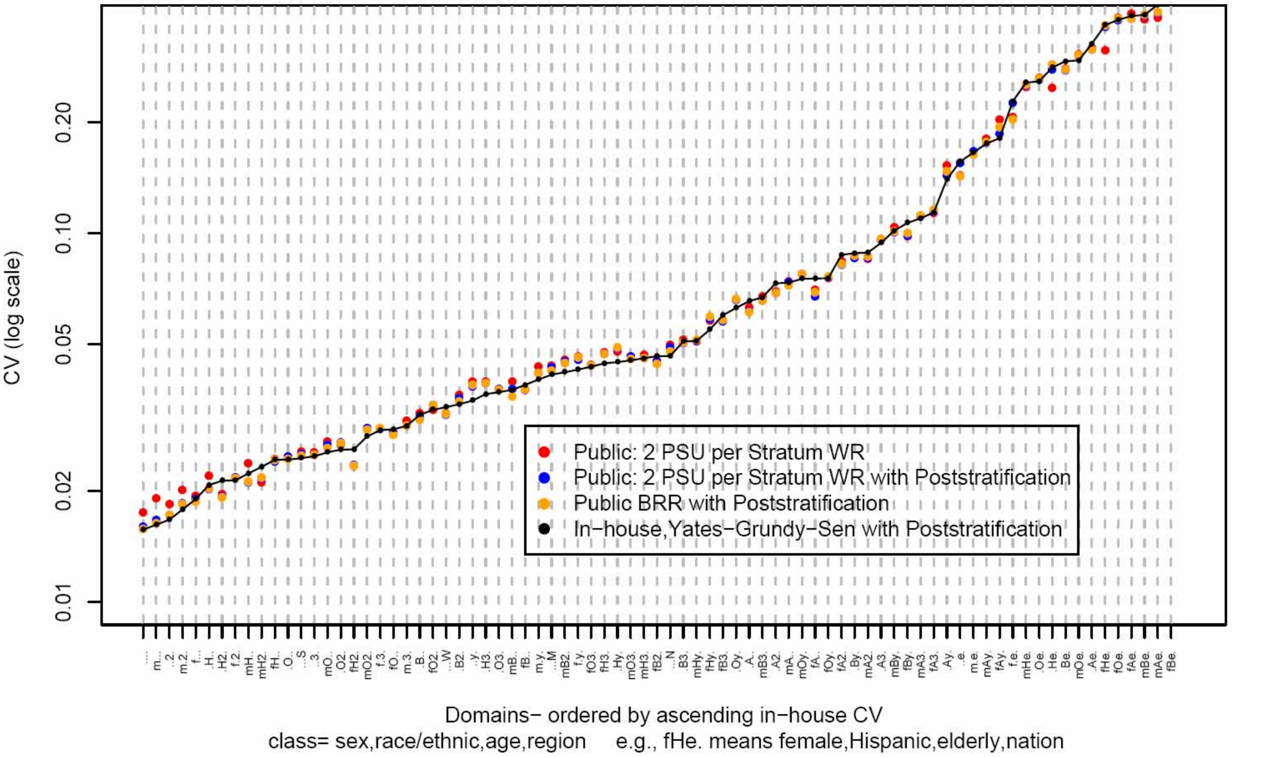


Fig. 3 Estimated CV's for Total: Persons with No Health Insurance for different variance−design structures

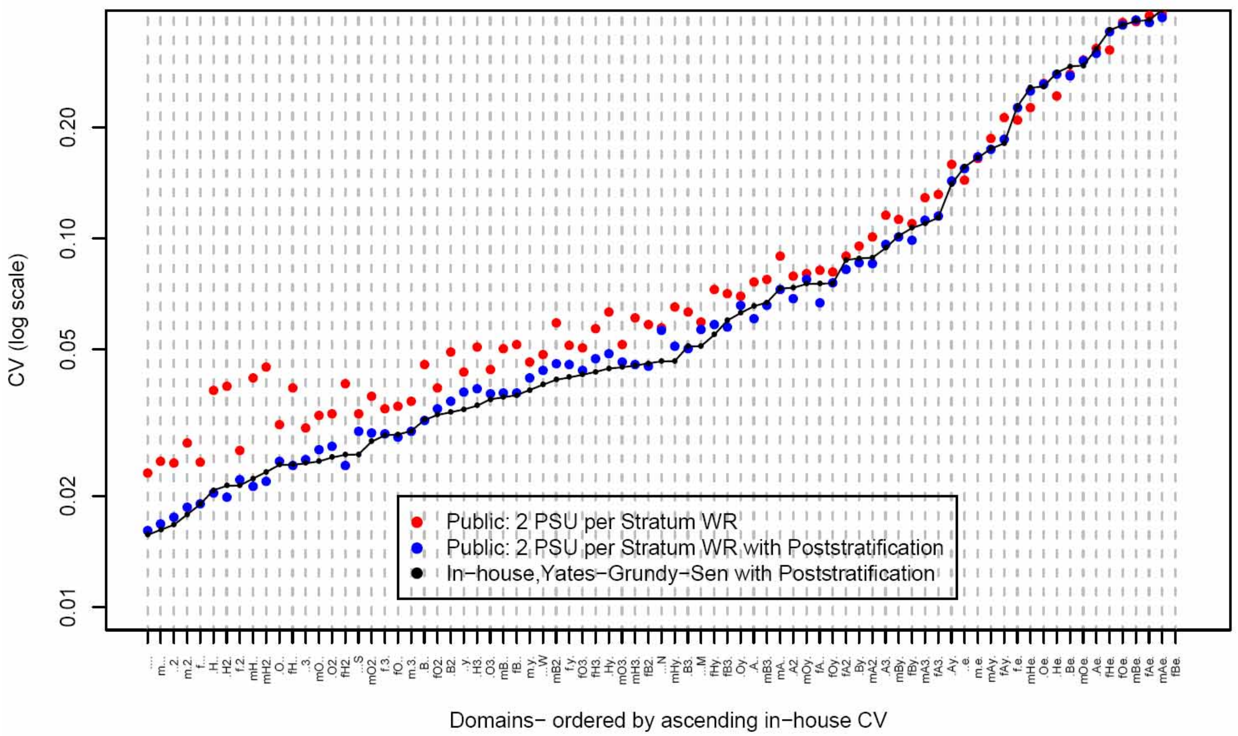
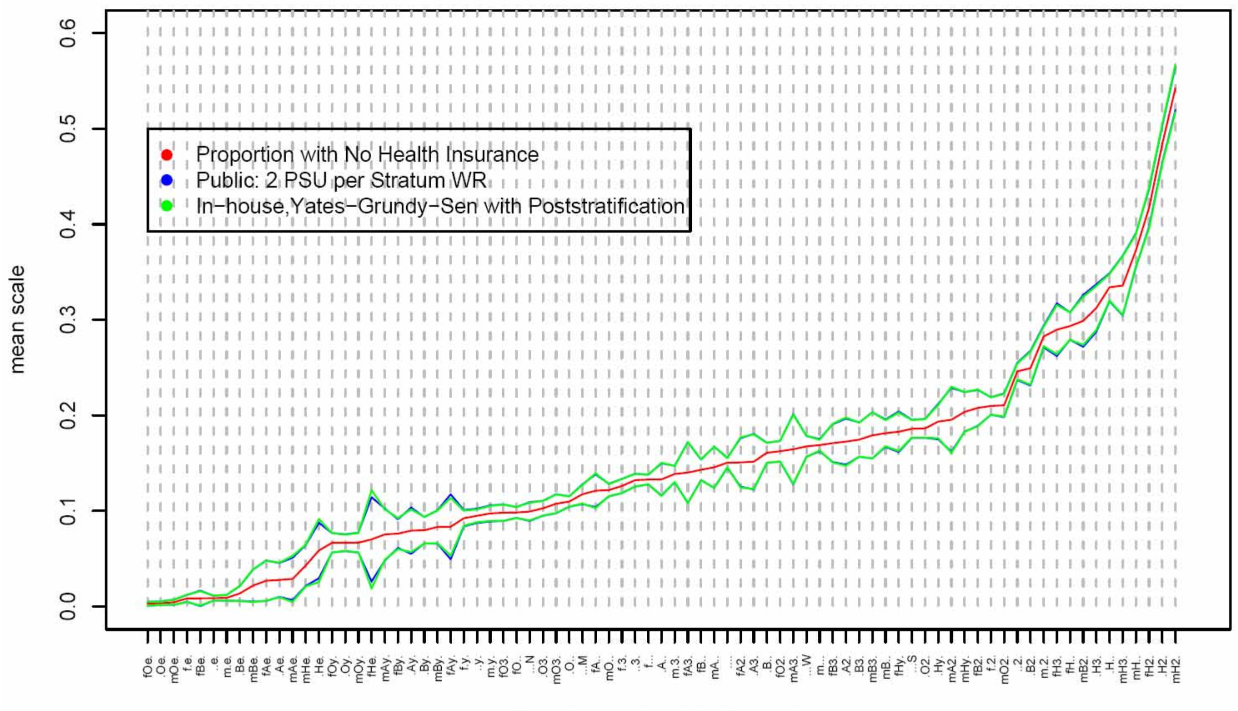


Fig. 4 Confidence Intervals for Proportions of Persons with No Health Insurance for Public−Use design and Inhouse design



class= sex,race/ethnic,age,region e.g., fHe. means female,Hispanic,elderly,nation