Method Research Using NACC UDS Database

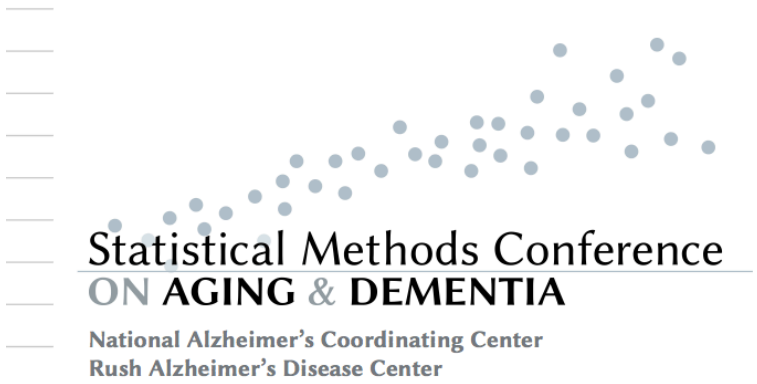
Xiao-Hua Andrew Zhou

Professor, Department of Biostatistics
Associate Director of National Alzheimer's Coordinating Center, University of Washington, Director of Biostatistics Unit, VA Puget Sound Health Care System

November 27, 2015; NACC ADC Meeting

Update on a Recent Conference

Statistical Methods Conference on Aging & Dementia, sponsored by NACC and hosted by Rush Alzheimer's Disease Center on Sept 24, 2015.



Working Groups

- Challenges in randomized controlled trials of pre-clinical AD, MCI, and dementia, including better outcomes in the form of single and multiple composite measures for clinical trials (Chair: LAUREL BECKETT; CO-CHAIR: STEVE EDLAND)
- Challenges in the study of natural history of aging and dementia and risk score composites and trajectories: Towards dynamic risk profiles (CHAIR: SUE LEURGANS; CO-CHAIRS: GRACIELA MUNIZ, SHARON XIE)
- Modeling disease progression: Analysis of change before dementia and modeling attrition, mortality, and left truncation (CHAIR: RICHARD KRYSCIO; CO-CHAIR: ANDREW ZHOU)

Analytic Challenges in UDS Databases

- Verification Bias and Imperfect Gold Standard
- Modeling transition probabilities among normal, MCI, and dementia
- Missing data in longitudinal studies
- Truncation by death
- Causal inference in observational studies

Verification Bias in Evaluation of Diagnostic Tests

- When one is interested in assessing the accuracy of neurological tests in detecting Alzheimer's disease (AD), we need to know the true AD status of a subject.
- The current gold standard for diagnosing the AD is by neuropathological brain autopsy.
- For many subjects, we may not be able to obtain their true disease statuses either due to the fact that they were alive or that autopsy was not performed even when they were dead.

Definition of an ROC Curve

- The ROC curve for Y is a plot of $S_1(t)$ versus $S_0(t)$ for $-\infty < t < \infty$.
- If we define the inverse function of $S_d(t)$ by

$$S_d^{-1}(p) = \sup\{t : S_d(t) \geq p\},$$

we can write the ROC curve as

$$ROC_x(p) = S_1(S_0^{-1}(p)), \tag{1}$$

where $p = S_0(t)$, the false positive rate (FPR) corresponding to a cutoff point t in the domain of the survival distribution function, S_0.

- That is, the ROC curve is a plot of $ROC(p)$ versus p for $0 \leq p \leq 1$.

Verification Bias

- To estimate sensitivity/specificity, predictive values, and ROC curves, we assume that we know the disease status of each patient under the study.
- In clinical practice, however, some of the patients with test results may not have verified disease status.
- When one is interested in assessing the accuracy of neurological tests in detecting Alzheimer's disease (AD), we need to know the true AD status of a subject.
- The current gold standard for diagnosing the AD is by pathology.

Verification Bias, continued

- For many subjects, we may not be able to obtain their true disease statuses either due to the fact that they were alive or that autopsy was not performed when if they were dead.
- When only verified subjects are used in estimation of the accuracy of diagnostic tests, the estimated accuracy of the tests may be biased. This type of bias is called verification bias.

How to Correct for Verification Bias

- A patient without disease verification = missing the value of the true disease status.
- The problem of verification bias is a special type of missing-data problems.
- Methods for data with missing values can be used for dealing with verification bias.

Missing at Random (MAR) Verification Process

- MAR Verification Process: the probability of disease verification depends only on test results and other observed covariates, not on the disease status itself.
- Non-MAR: the probability of disease verification depends on the true disease status, itself.

Limitations of Existing Methods

- Most of the methods reply the assumption that either the model for the verification process or the model for the probability of having the disease is correct.

New Non-parametric Method

- We are proposing an improved bias correction procedure for the non-parametric estimation of the ROC curve and AUC in the presence of verification bias.
- Our estimator is based on a unified empirical-likelihood-based estimation for all pairs of the true and false positive rates.
- The proposed estimator has multiple robust properties.
    - When either one of several models for the verification process or one of several models for the disease process is correct, the proposed estimator is consistent.

Example

- We illustrate our method by applying it to the uniform data set developed and maintained by the National Alzheimer's Coordinating Center (NACC).
- The test under evaluation is the mini-mental state examination (MMSE), a 30-point questionnaire used to screen for cognitive impairment, with lower values indicating more severe impairment. In accordance with convention, we used 30 minus the original test score as our response, so that larger response values were more indicative of disease status.
- Our interest is focused on how well the MMSE test can predict progression to dementia among subjects who have at least four visits to one participating disease center during the period January 2005 to May 2014

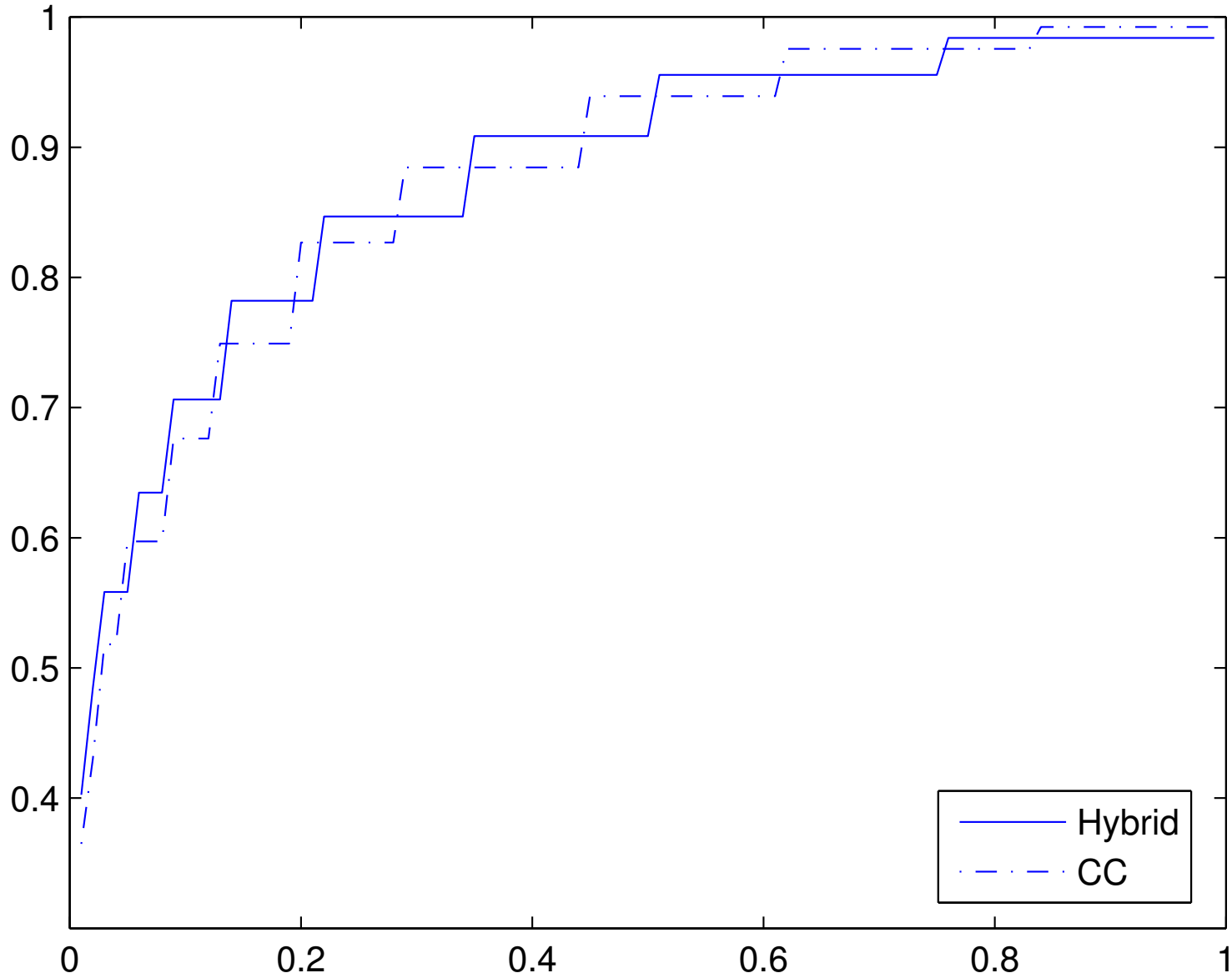
Example

- The true dementia status D was determined based on the consensus clinical diagnosis result at a patient's last visit to his/her participating center, about one fourth missing.
- The covariates considered were: X_1, indicating a patient's age at the last test; X_2, a binary variable indicating male; X_3, a binary variable indicating white people; X_4, indicating a subject's years of education; X_5, the subject's living situation, 1 if lives alone and 0 otherwise; X_6, subject's current martial status, 1 for married and 0 otherwise; X_7, the subject's health history related to heart attack, 1 if presence of a history of the condition and 0 otherwise; X_8, binary variable indicating having stroke before; X_9, a binary variable indicating presence of Parkinson's disease; X_{10}, a binary variable indicating presence of seizures.
- $X = (X_1, \ldots, X_10)^\top$.

Working Models

- For the conditional disease model, $P(D = 1 \mid X)$: (1) probit model, (2) logistic regression model, and (3) the complement loglog model
- For the verification process model, $P(V = 1 \mid X)$: a probit model, and two logistic regression models with different functional forms of X.

Estimated ROC Curves for NACC Database



Imperfect Gold Bias

- Even the diagnosis on AD using neuropathological brain autopsy may be wrong, the estimated accuracy of diagnostic tests using imperfect gold standard may be biased, called imperfect gold standard bias.
- Wang and Zhou (2015, submitted) proposed a new method for estimating the accuracy of promosing biomarkers in the cerebral spinal fluid (CSF) in the presence of gold standard.
- In this paper, we propose a latent profile approach to simultaneously handling the issues that accompany the absence of a gold standard, unknown biomarker distributions, and differential covariate effects.
- Our approach further provides an optimal combination of multiple biomarkers for risk prediction in the presence of a gold standard.

Modeling transition among normal, MCI, dementia, and death

- AD-related pathological changes are believed to begin 10 years or more before the appearance of any detectable clinical symptom, and decades before sufficient cognitive impairment accrues to warrant a clinical diagnosis of AD
- It is important research to use biomarkers and other features of patients to detect transition probabilities among normal, MCI, dementia, and death states.
- Chen and Zhou (2015, Journal of Multivariate Analysis) proposed a new non-homogeneous Markov process for modeling transition probabilities among normal, MCI, and dementia, and death with incomplete clustered life history data.

NACC UDS Data Set

Parameter	Proposed Method			Naive Analysis		
	HR	95%LCL	95%UCL	HR	95%LCL	95%UCL
Normal → MCI:						
SEX(F)	0.972	0.795	1.200	1.022	0.827	1.292
fhdem	1.275	1.032	1.570	1.372	1.087	1.720
MMSE	1.020	0.981	1.030	1.011	1.006	1.021
AGE	1.026	1.013	1.040	1.033	1.014	1.041
MCI → Dementia:						
SEX(F)	0.708	0.580	0.865	0.712	0.572	0.877
fhdem	1.246	1.012	1.534	1.191	0.955	1.380
MMSE	0.997	0.984	1.011	0.999	0.988	1.024
AGE	1.023	1.011	1.035	1.020	1.007	1.033
Dementia → Death:						
SEX	0.660	0.549	0.794	0.614	0.515	0.732
fhdem	1.077	0.887	1.307	1.117	0.928	1.345
MMSE	0.875	0.864	0.886	0.986	0.877	0.996
AGE	1.045	1.033	1.056	1.046	1.035	1.058

Generalized Partially Linear Models for Incomplete Longitudinal Data

- In observational longitudinal studies, like NACC UDS, interest often lies in estimation of the population-level relationship between the explanatory variables and dependent variables
- Inclusion of population-level information in longitudinal data can increase estimation efficiency.
- Chen and Zhou (2014, JRSS C) considered a generalized partially linear model for incomplete longitudinal data in the presence of the population-level information.
- A pseudo-empirical likelihood-based method is introduced to incorporate population-level information, and non-random drop-out bias is corrected by using a weighted generalized estimating equations method.
- A three-step estimation procedure is proposed, which makes the computation easier.

Truncation by Death in Longitudinal Data

- It is common that in observational studies, the outcome of interest is truncated by death, meaning that a subject had died before the outcome could be measured.
- In this case, restricted analysis among survivors may be subject to selection bias. It is hence of interest to estimate the survivor average causal effect (SACE), defined as the average causal effect among subjects who would survive under either exposure.
- Wang and Zhou (2015, submitted) consider the identification and estimation problems of SACE. We propose to identify a substitution variable for the latent membership of the always-survivor group. The identifiability conditions required for a substitution variable are similar in idea to conditions for an instrumental variable.
- We show that SACE is non-parametrically identifiable with use of a substitution variable, and propose a

Truncation by Death in Longitudinal Data

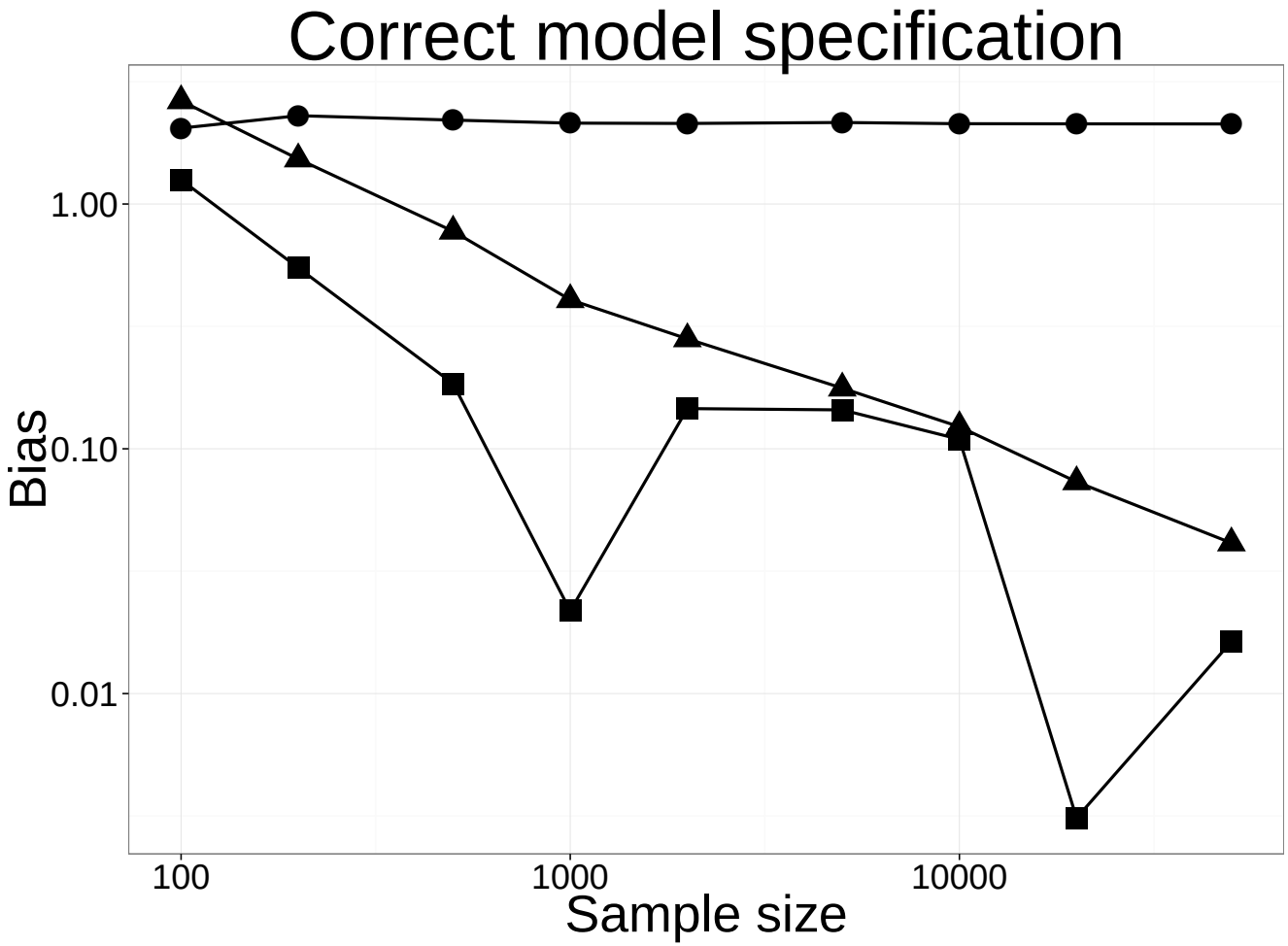
- It is common that in observational studies, the outcome of interest is truncated by death, meaning that a subject had died before the outcome could be measured.
- In this case, restricted analysis among survivors may be subject to selection bias. It is hence of interest to estimate the survivor average causal effect (SACE), defined as the average causal effect among subjects who would survive under either exposure.

Causal Inference in Observational Data

-
- Wang and Zhou (2015, in revision) consider increasing the number of subclasses with sample size and show that this enables consistently estimating the average causal effect via subclassification at a square root convergence rate.

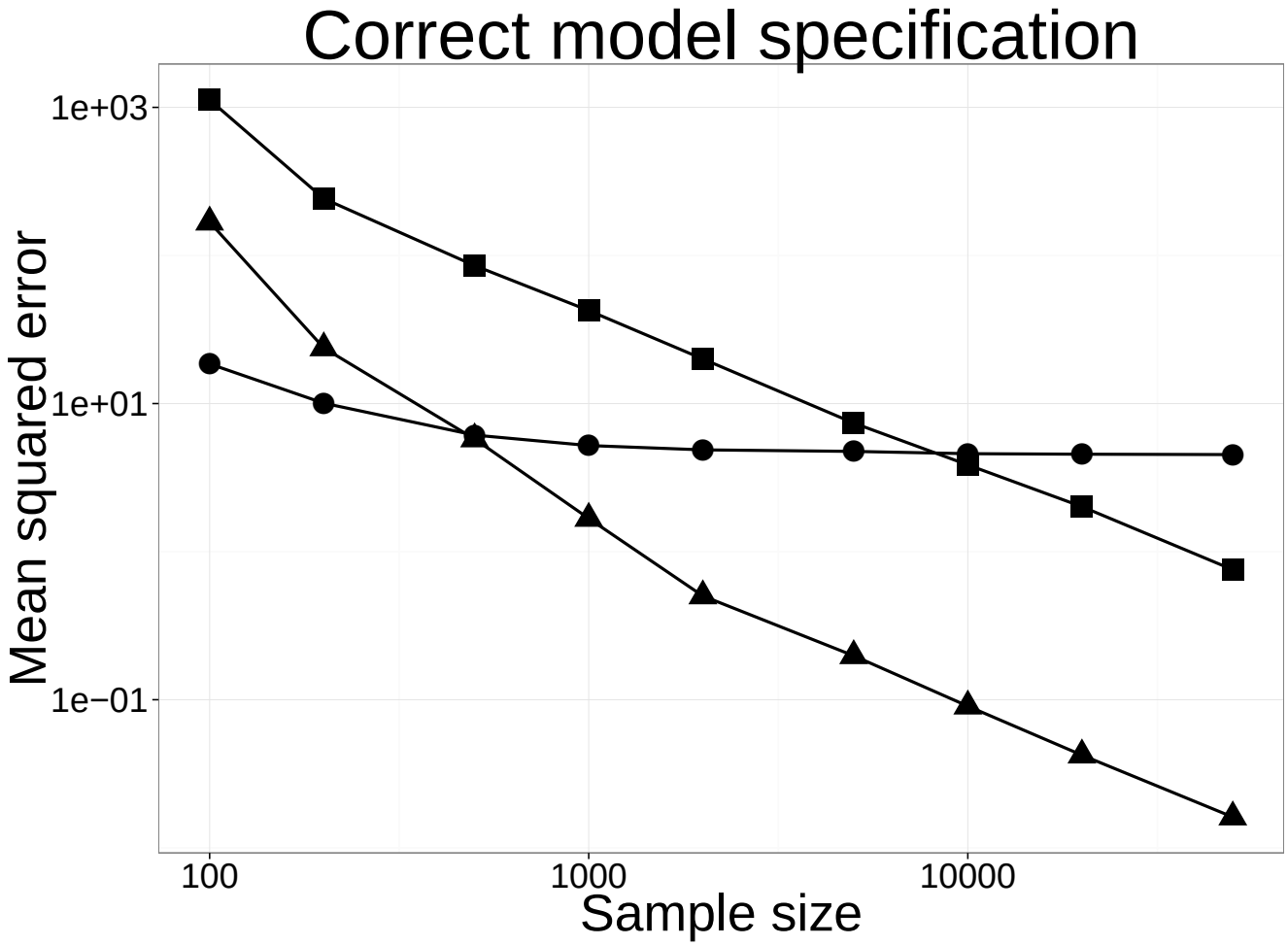
Correct PS model: bias

Stratification, IPW and the proposed estimator:



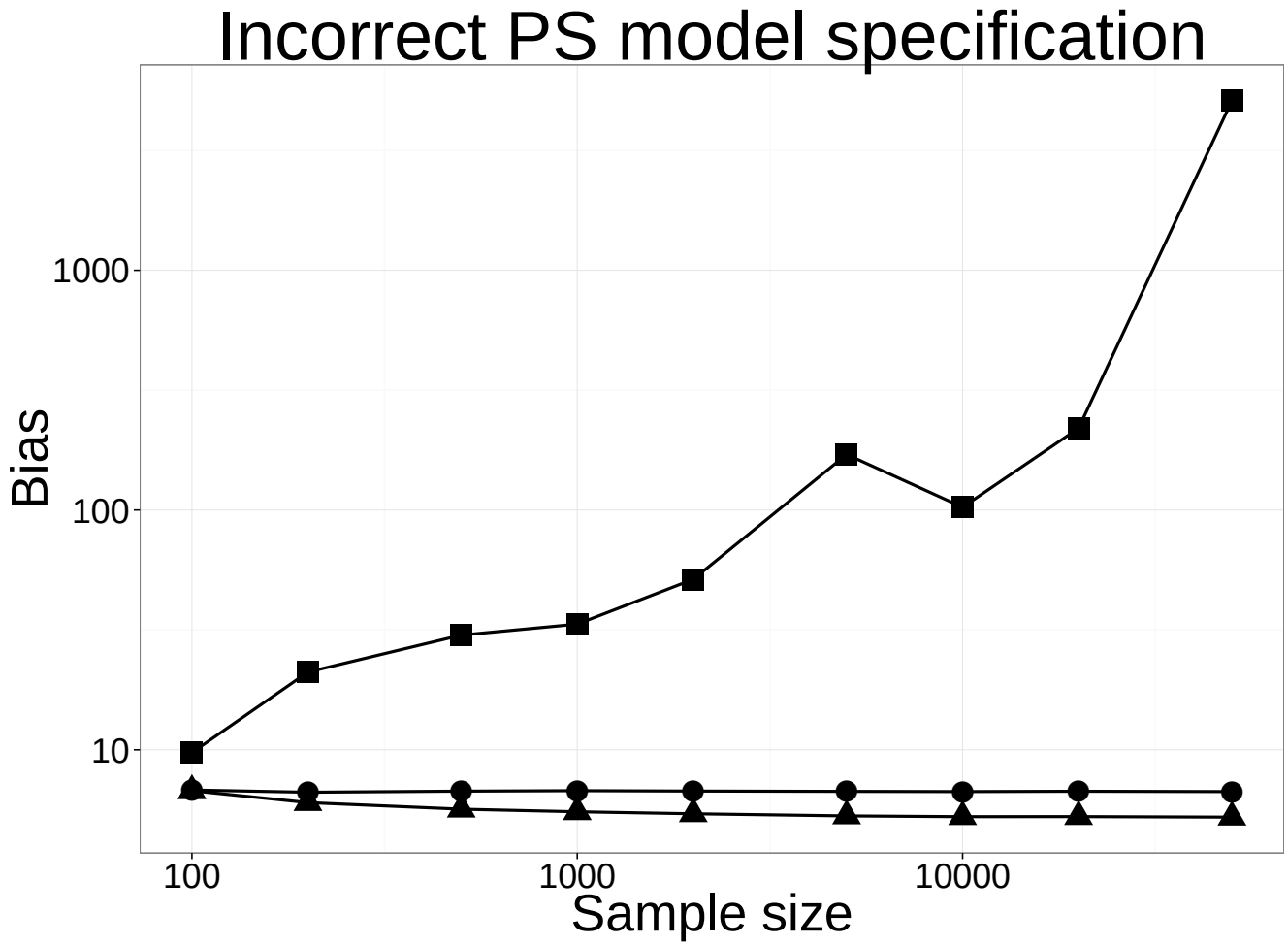
Correct PS model: mean squared error

Stratification, IPW and the proposed estimator:



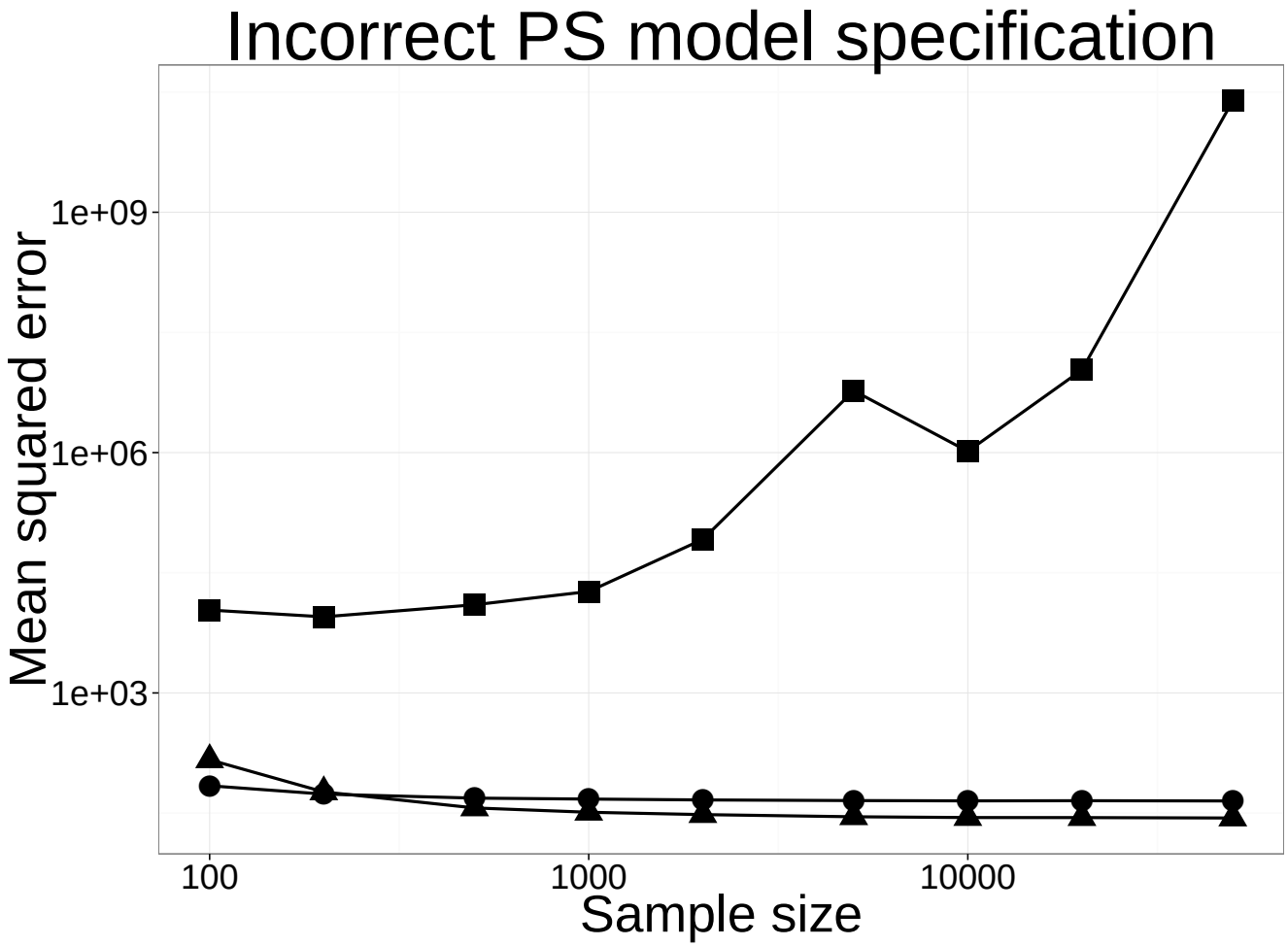
Misspecified PS model: bias

Stratification, IPW and the proposed estimator:



Misspecified PS model: mean squared error

Stratification, IPW and the proposed estimator:



Other Method Research Using NACC Databases

- Dr. Xiong and his colleagues at Washington University in St Louis have been conducted research in combine multiple cognitive tests to better power clinical trials, using NACC database
- Dr. Gao and her colleagues have been working on a multivariate finite mixture latent trajectory model using UDS data.
- Dr. Tripodis has been working on a dynamic factor for multivariate time series using UDS data set
- Dr. Steenland has tried to use the marginal structure model to adjust for attrition in sudying the relationship between depression and cognitive decline using UDS database.