

Getting More from Analyses of Data Sampled from a Defined Cohort

Norman Breslow

Department of Biostatistics

University of Washington, Seattle

Plenary Lecture

World Congress of Epidemiology

Porto Alegre, Brazil

September 23, 2008

Acknowledgements

- Jon Wellner
- Thomas Lumley
University of
Washington, Seattle
- Lloyd Chambless
University of North
Carolina, Chapel Hill
- Michal Kulich
Charles University,
Prague
- Christie Ballantyne
Baylor College of
Medicine, Houston

Outline

- Problem: sampling from defined cohort
 - Nested case-control vs case-cohort sampling
 - Current published analyses ignore good data
- Epidemiology meets survey sampling
 - Two-phase stratified sampling designs
 - Adjusted sampling weights use more data
 - Post-stratification, calibration, estimation
- Re-analysis of data from Atherosclerosis Risk in Communities (ARIC) cohort
- Simulated case-cohort studies from National Wilms Tumor Study (NWTs) cohort

Sampling from Defined Cohort

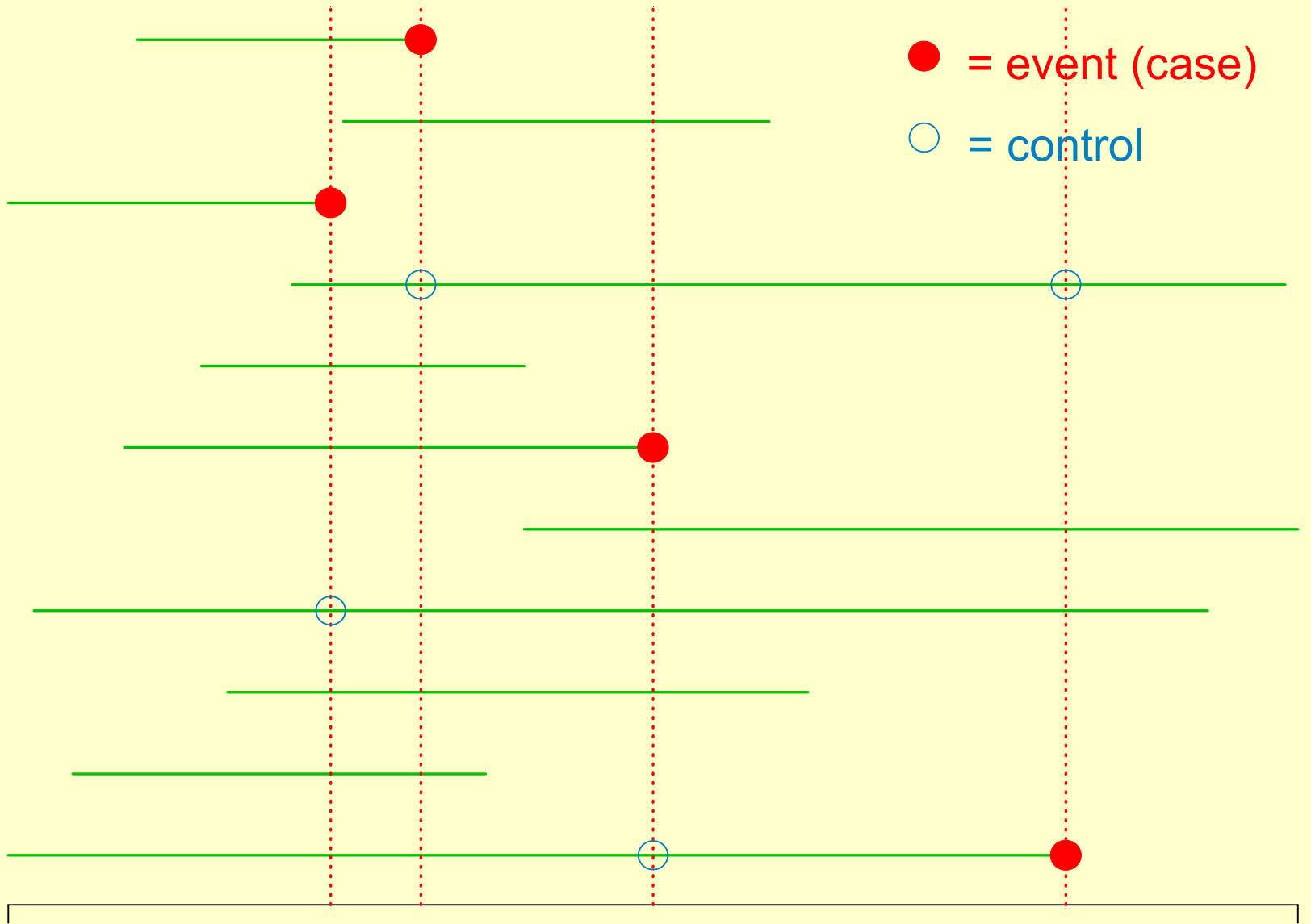
- Large numbers of subjects in follow-up
- Some data available for everyone
 - Outcomes
 - Demographics (gender, age, ethnicity)
 - Covariates (possibly subject to measurement error)
- Additional, costly data potentially available
 - Assays of stored biological tissues
 - LOH, LOI, gene expression levels, genotypes
 - Detailed medical records abstraction.

Basic Questions

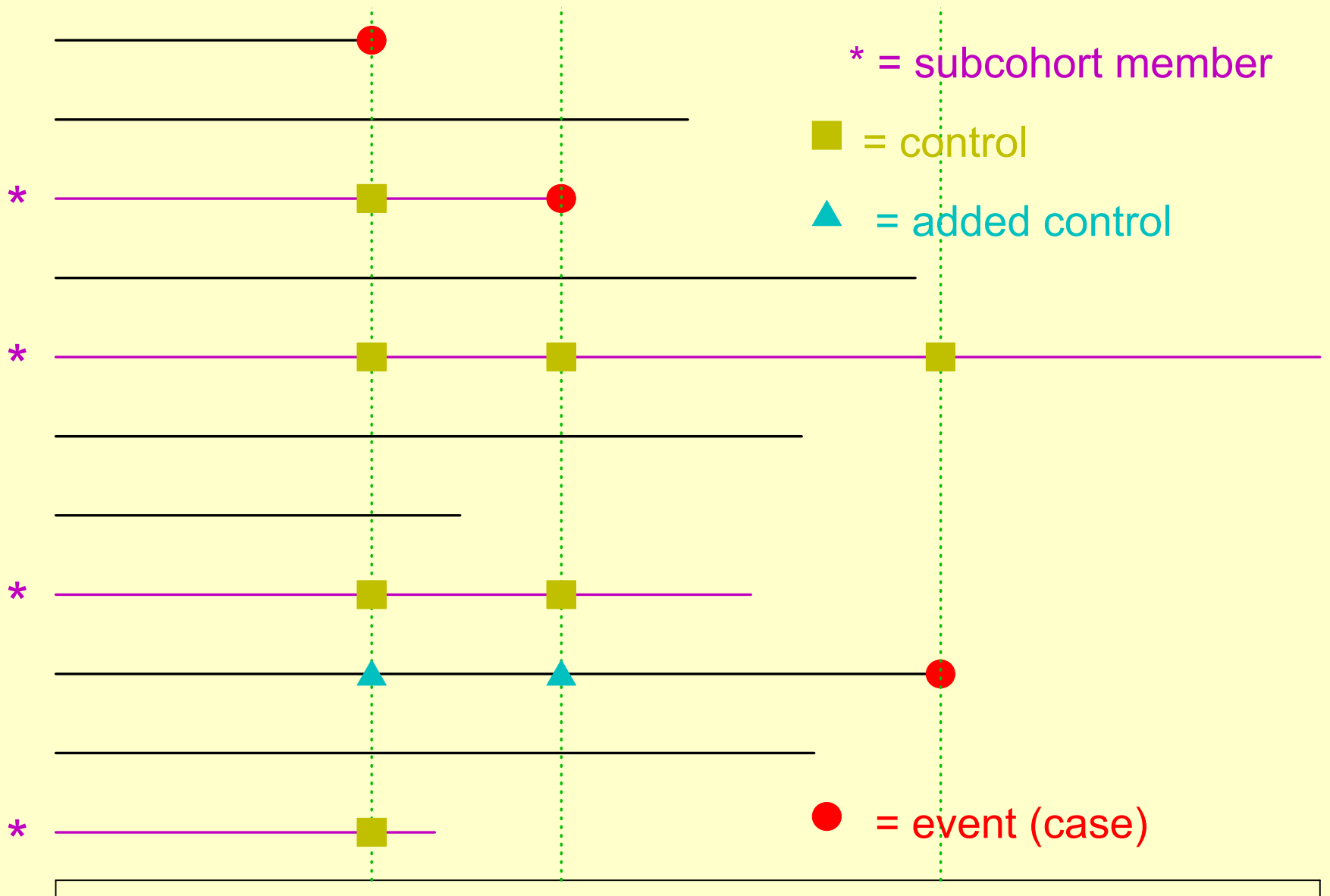
- How to select subjects for bioassay, or other detailed covariate ascertainment?
- How to analyze resulting “two phase” data to provide most precise estimates of hazard ratios?
 - *i.e.*, of exponentiated regression coefficients in Cox proportional hazards model for population

Design Options for Follow-up Studies

- Nested case-control study (Mantel, '73, Thomas '77)
 - “Risk set” sampling in context of Cox regression
 - 1:M matched samples of cases and controls
- Case-cohort study (Kupper et al. '75, Prentice, '86)
 - Random subcohort (cohort random sample)
 - Additional data for subcohort and “outside” cases
- Two phase stratified versions of both designs
 - Over sample “most informative” subjects
 - Disease cases (usually sampled at 100%)
 - Rare or unusual patterns of exposure
 - Here consider stratified case-cohort studies
 - See Langholz & Clayton (*EHP*, 1994) and Langholz & Borgan (*Biometrika*, 1995) for “counter-matching” in nested case-control studies



Risk sets for nested case-control study



Risk sets for case-cohort study

Relative Advantages

- Nested case-control study (risk-set sampling)
 - Bioassays at same time for cases and matched controls
 - Important if biological samples change with time
- Case-cohort study: subcohort (CRS) facilitates
 - estimation of population frequencies of risk factors
 - genotypes
 - control of multiple endpoints
 - heart disease & diabetes
 - analyses using alternative time scales
 - time-on-study & attained age

Example I: ARIC*

- 15,792 participants, 45-64 yr, examined 1986-89
- Randomly selected from 4 U.S. communities
 - 2 white, 1 black, 1 mixed (12% black)
- Baseline data from interview, blood, ultrasound
 - Education, smoking history, medication use
 - Lipids, hemostasis, chemistries
 - Imaging (wall thickness) of carotid artery (& others)
- Follow-up exam q. 3 yr + community surveillance
 - MI, CHD, angina, stroke, cause-specific mortality

* The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. *Am J Epidemiol*, 1989 & many publications since

ARIC: Stratified Case-Cohort Studies

- *GSTM1/GSTT1* x Smoking interaction and CHD
 - Li *et al.*, *Atherosclerosis*, 2000
- *HFE* C282Y mutation and CHD
 - Rasmussen *et al.*, *Atherosclerosis*, 2001
- C-reactive protein and CHD
 - Folsom *et al.*, *Am Heart J*, 2002
- *GSTM1/GSTT1* x Smoking and atherosclerosis
 - Olshan *et al.*, *Epidemiology*, 2003
- GP IBA α VNTR alleles and CHD
 - Afshan-Kharghar *et al.*, *Blood*, 2004
- Lp-PLA $_2$ x C-reactive protein and ischemic stroke
 - Ballantyne *et al.*, *Arch Int Med*, 2005
-

ARIC: *GSTM/GSTT* x Smoking & CHD

- $n = 14,239$ in main cohort (after exclusions)
- $n = 986$ in sub-cohort from *stratified* random sample

Carotid wall	Male		Female		Total
	Age<55	Age≥55	Age<55	Age≥55	
Thin	86	289	59	152	586
Thick	50	150	50	150	400
Total	136	439	109	302	986

- $n = 458$ incident CHD cases 1987-1993 (36 in sub-cohort)

“Over-sampling on thin carotid walls permits study of persons with minimal evidence of atherosclerosis, and over-sampling on ages ≥ 55 strata ensures a sufficient number of non-cases in those strata for comparison”

ARIC: Critique of Design and Analysis

- For study of gene \times environment interaction, best to stratify on (rare) environmental factor*
 - Approx equal no. of smokers/non-smokers by gender
- Analysis *ignores* data on sex, age, smoking, lipids, etc. available for $14,239 - 986 - 458 + 36 = 12,831$ controls *not* sampled for sub-cohort
 - Information on “main effect”, CHD rates for smokers vs. non-smokers, available for entire cohort
 - Baseline data available for entire cohort most valuable when they *predict* “missing” data

* Goldstein & Andrieu, *JNCI Monographs*, 1999

QUERY: How can we utilize information ignored in current, conventional analyses of case-cohort data?

ANSWER: Adjust the sampling weights used in the conventional analyses.

NEED: Some concepts and notation from survey sampling

Two Phase Sampling



- Population may be finite or infinite
 - Finite => actual population (*e.g.* US population)
 - Infinite => from probability model
- Sampling at Phases I and II may be
 - simple random or cluster sampling
 - with or without stratification
- Case-cohort design involves
 - simple random sampling from model at Phase I
 - finite population stratified sampling at Phase II

Two Phase Sampling

- Phase I: N subjects in cohort are classified into K strata on basis of information known for everyone: N_k in stratum k , $k=1, \dots, K$, so that $N = N_1 + N_2 + \dots + N_K$
- Phase II: For each k , $n_k \leq N_k$ subjects are sampled at random, without replacement from stratum k , and more information is collected for resulting $n = n_1 + n_2 + \dots + n_K$

Illustration: ARIC Case-Cohort Study*

- $N=12,345$ in main cohort, followed 6-8 yrs
 - Plasma collected at second visit (start of FU)
 - Free of CHD, transient ischemic stroke
- $n=1,336$ at Phase II (604 CHD, 732 controls)
 - Plasma assayed for C-reactive protein (CRP) and lipoprotein-associated phospholipase A₂ (Lp-PLA₂)
- Focus on association of Lp-PLA₂ with CHD after adjustment for traditional risk factors

* Ballantyne CM et al. *Circulation* **109**:837-42, 2004

ARIC Case-Cohort Study

	Non CHD cases (controls)								CHD cases	Totals
Race	Black				White					
Sex	Female		Male		Female		Male			
Age	<55	≥55	<55	≥55	<55	≥55	<55	≥55		
Stratum (<i>k</i>)	1	2	3	4	5	6	7	8	9	
Cohort N_k	1,133	719	598	393	2,782	2,213	1,959	1,818	730	$N=12,345$
Sample n_k	59	54	42	71	88	154	117	147	604	$n=1,336$
Weights N_k/n_k	19.2	13.3	14.2	5.5	31.6	14.4	16.7	12.4	1.2	

Available Data

- X = variables in Cox regression model
 - time to development of CHD or time followed
 - main risk factor Lp-PLA₂ (known only at Phase II)
 - adjustment variables: age, sex, race, SBP, DBP, HDL-C, LDL-C, ...
- V = variables known for entire cohort
 - used to stratify Phase II sampling or adjust the weights
 - includes adjustment variables for ARIC CHD study
 - in general includes variables *not* in regression model

Horvitz-Thompson Estimator

- Inverse probability weighting (IPW) -- notation
 - $\xi_i = 1$ if i^{th} subject sampled at Phase II, 0 otherwise
 - $\pi_i =$ known sampling probability

$$\pi_i = \frac{n_k}{N_k} \text{ if subject } i \text{ in stratum } k$$

- Probability model $P_{\theta, \eta}(X)$
 - $\theta =$ regression coefficients in Cox model
 - $\eta =$ baseline hazard function

Horvitz-Thompson Estimator

- Solve IPW likelihood equations for $(\hat{\theta}_N, \hat{\eta}_N)$

$$\frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} \dot{\ell}_{\theta, \eta}(X_i) = 0 \quad (\dot{\ell}_{\theta, \eta}, \text{ scores for } \theta)$$

$$\frac{1}{N} \sum_{i=1}^N \frac{\xi_i}{\pi_i} B_{\theta, \eta} h(X_i) = 0 \quad (B_{\theta, \eta} \text{ is score operator})$$

$h \in \mathcal{H}$, directions from which η may approach η_0

- Yields Barlow's (1994) method of analysis of case-cohort data and "robust" variance
 - solve weighted partial-likelihood equations
 - most common method in current use

Sampling Properties of HT Estimator*

$$\begin{aligned}\sqrt{N}(\hat{\theta}_N - \theta_0) &= \sqrt{N}(\tilde{\theta}_N - \theta_0) + \sqrt{N}(\hat{\theta}_N - \tilde{\theta}_N) \\ &\approx \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{\ell}_0(X_i) + \frac{1}{\sqrt{N}} \sum_{i=1}^N \left(\frac{\xi_i}{\pi_i} - 1 \right) \tilde{\ell}_0(X_i)\end{aligned}$$

$$\text{Var}_{\text{TOT}} = \text{Var}_{\text{PHASE I}} + \text{Var}_{\text{PHASE II}}$$

- $\tilde{\theta}_N$ is **unobserved** MLE based on complete data
- $\tilde{\ell}_0$ is semi parametric **efficient influence function**
- $\text{Var}_{\text{PHASE II}}$ is **design based**: normalized error in HT estimation of unknown finite pop. total $\tilde{\ell}_{\text{TOT}} = \sum_{i=1}^N \tilde{\ell}_0(X_i)$
- Phase I and Phase II contributions approx. **independent**

* Breslow & Wellner, *Scandinavian J Statist*, 2007/8, and others

Two Components of Variance

- **Phase I** variance represents usual uncertainty in generalizing results for N cohort subjects to target population
 - only variance if complete data for all
 - **cannot** be reduced by adjustment of weights
- **Phase II** variance represents additional uncertainty from not having complete data for all N cohort members, but only for n
 - **can** be reduced by adjustment of weights

Improving Efficiency: Survey Techniques

- Construct auxiliary vars $\tilde{V} = \tilde{V}(V)$ to **adjust weights**
 - Post-stratification (finer than needed for biased sampling)
 - Calibration (generalized raking: Deville, Särndal, Sautory, *JASA* 93)
 - Estimation using correct parametric model: $\pi_i = \pi(\tilde{V}_i; \alpha)$
(Robins, Rotnitzky, Zhao, *JASA* 94)
- One possibility for auxiliary variables \tilde{V}
 1. Impute missing X values using parametric model $[X|V]$
 2. Fit model $P_{\theta, \eta}(X)$ to main cohort using imputed data
 3. Construct \tilde{V}_i as “delta-betas” for above model
 - surrogates for unknown $\tilde{\ell}_0(X_i)$
- Estimate θ using adjusted weights based on \tilde{V}_i

Calibration vs Estimation of Sampling Weights*

- Weights **calibrated** to Phase I totals solve

$$\sum_{i=1}^N \xi_i w_i \tilde{V}_i = \sum_{i=1}^N \tilde{V}_i$$

- Weights **estimated** using Phase I variables solve

$$\sum_{i=1}^N \xi_i \tilde{V}_i = \sum_{i=1}^N \tilde{V}_i / w_i$$

- For auxiliary stratum indicators, \tilde{V}_i binary indicators of stratum membership, two sets of weights agree

$$w_i = \frac{1}{\pi_i} = \frac{N_j}{n_j} \quad \text{for } i \in \text{stratum } j$$

- Both sets converge to true π_i^{-1} in large samples

* after Lumley (2007)



Survey Package

- Implements adjustment of weights in weighted Cox regression analysis of stratified case-cohort data (and a whole lot more)
- See author Thomas Lumley's website <http://faculty.washington.edu/tlumley/survey/>
- Datasets and sample R code used for NWTs simulations reported below are at my site <http://faculty.washington.edu/norm/software.html>

Adjustment of Sampling Weights

- If variables used for calibration or estimation are the **only** variables in the probability model, then weighted estimate same as estimate from fit to main cohort and Phase II variance component is zero
- Illustrate by exploring relationship between Lp-PLA₂ and HDL-C with ARIC data

Association between Lp-PLA₂ and HDL-C: Standard Sampling Weights

HDL-C (mg/L)	Lp-PLA ₂ (μG/L)			Total
	0-0.309	0.310-0.421	0.422-1	
< 40	701.4 (105.7)	938.6 (111.3)	1,561.3 (138.9)	3,201.3 (185.4)
40-59.0	1,764.4 (166.9)	2,310.2 (187.5)	2,175.3 (170.0)	6,249.9 (234.1)
≥ 60.0	1,569.6 (164.2)	909.4 (124.1)	414.8 (81.9)	2,893.8 (197.7)
Total	4,035.4 (217.3)	4,158.2 (222.2)	4,151.4 (206.1)	12,345

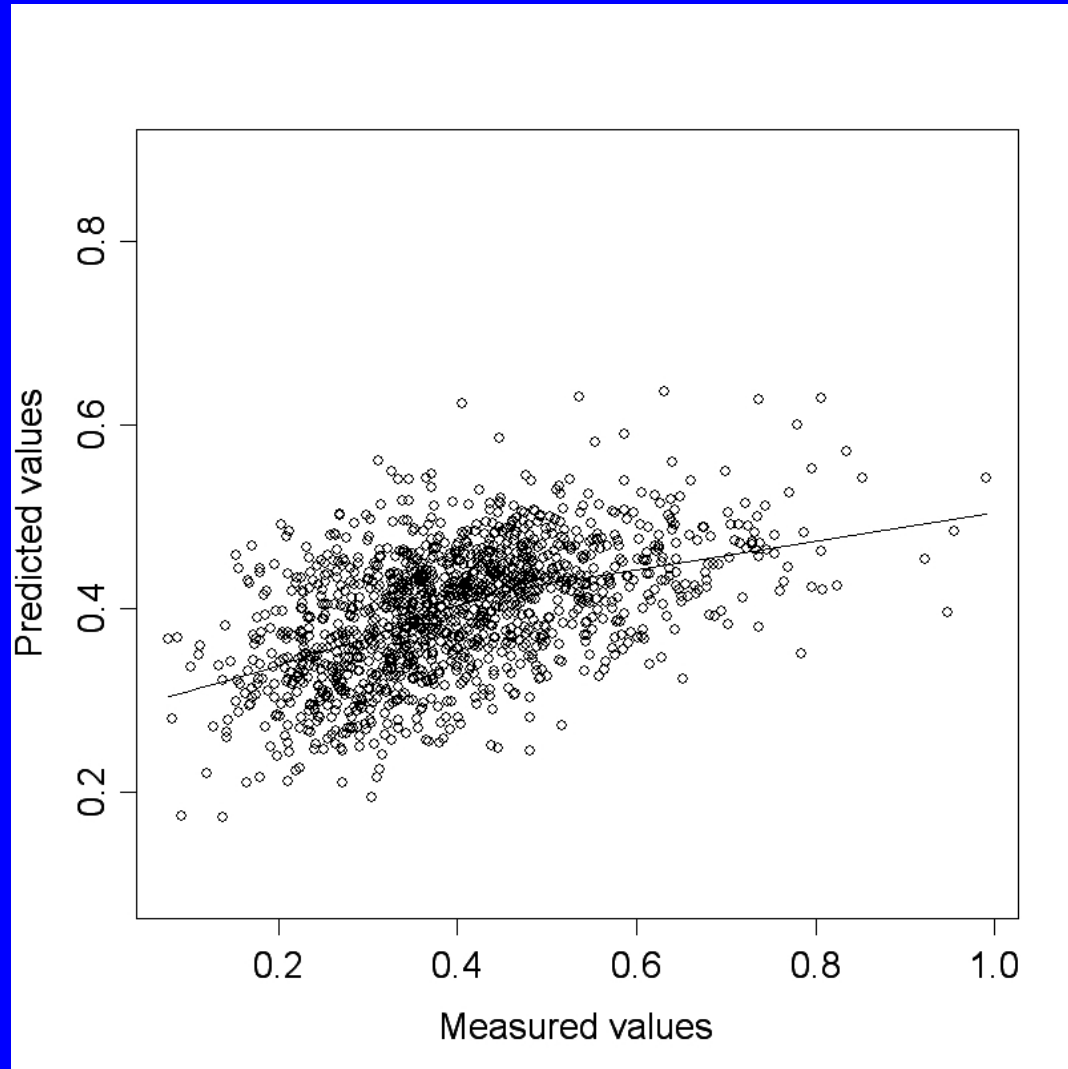
Association between Lp-PLA₂ and HDL-C: Calibrated (to HDL-C) Weights

HDL-C (mg/L)	Lp-PLA ₂ (μG/L)			Total
	0.0.309	0.310-0.421	0.422-1	
< 40	739.0 (99.7)	988.9 (105.0)	1,645.0 (117.7)	3,373 (49.5)
40-59.0	1,665 (144.4)	2,180.1 (155.8)	2,059.9 (146.7)	5,898 (55.5)
≥ 60.0	1,667.4 (128.0)	966.0 (117.0)	440.6 (83.1)	3,074 (48.0)
Total	4,071.4 (212.9)	4,135.1 (220.3)	4,138.5 (201.8)	12,345

Horvitz-Thompson Estimation of Hazard Ratios by Lp-PLA₂ Tertile

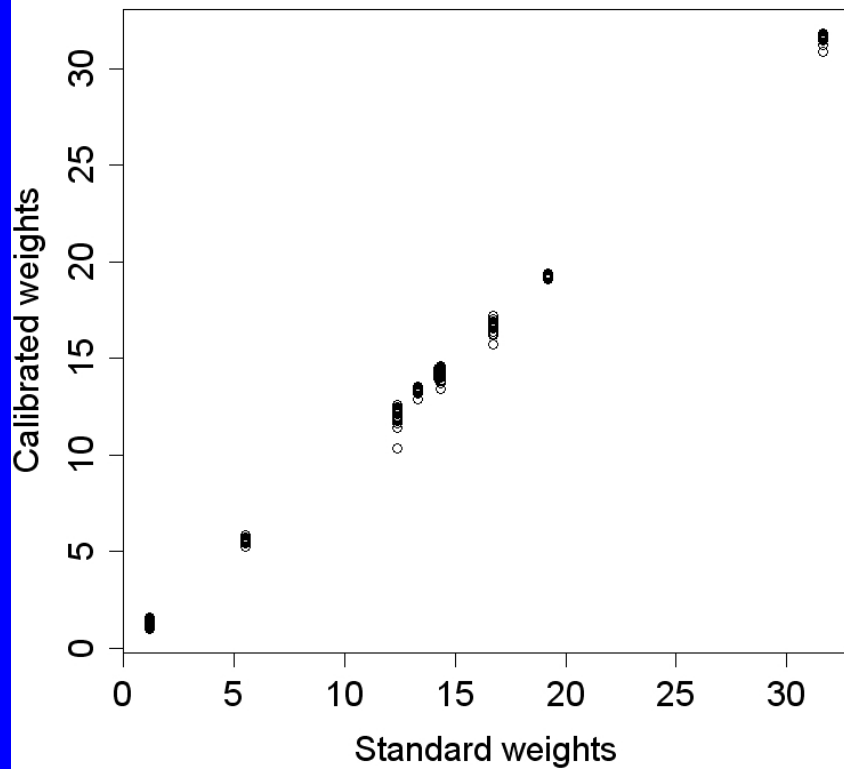
- Compare with Table 4, Model 2 of Ballantyne
 - results using standard weights virtually identical
- Prediction model for Lp-PLA₂ : (weighted) linear regression on sex*race, LDL-C, HDL-C, SBP and DBP using Phase II data
- Impute values of Lp-PLA₂ for all in main cohort
- Fit Cox model to cohort using imputed values
- Extract “delta-betas” and use for calibration or estimation of weights

Prediction of Lp-PLA₂ : $R^2 = 0.28$

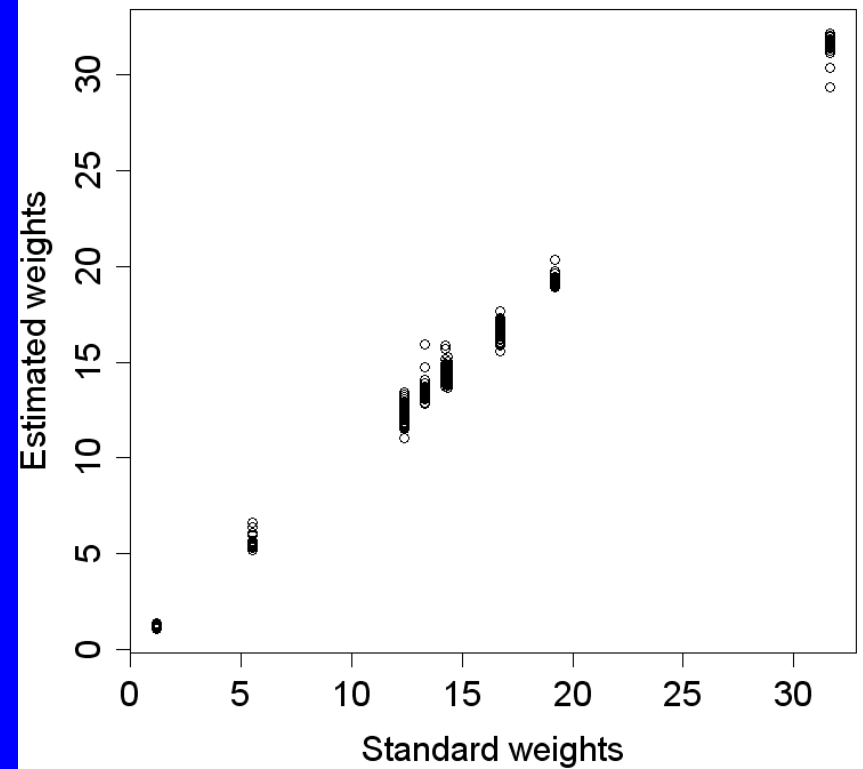


Comparison of Weights

Calibrated vs Standard



Estimated vs Standard



ARIC Case-Cohort Study of Lp-PLA₂

Results of Cox Regression Analyses

Model term*	Standard weights			Calibrated weights			Estimated weights		
	Coef.	SE1	SE2	Coef.	SE1	SE2	Coef.	SE1	SE2
Age/10	0.420	.073	.075	0.393	.073	.012	0.431	.073	.015
Male	0.762	.088	.091	0.791	.088	.019	0.742	.088	.022
White	0.037	.098	.090	0.159	.099	.016	0.101	.100	.029
Frmr smoker	-0.421	.093	.126	-0.464	.092	.017	-0.459	.092	.020
Never smoke	-0.552	.099	.129	-0.557	.099	.016	-0.622	.099	.020
SBP/100	1.554	.207	.267	1.539	.208	.046	1.580	.207	.048
LDL-C/100	0.777	.106	.151	0.786	.106	.045	0.748	.107	.048
HDL-C/100	-2.539	.329	.392	-2.361	.329	.052	-2.736	.334	.060
Diabetes	0.572	.092	.127	0.738	.090	.019	0.531	.093	.026
Lp-PLA₂ (2)*	0.052	.110	.126	0.054	.111	.127	0.050	.111	.127
Lp-PLA₂ (3)*	0.163	.108	.129	0.182	.108	.130	0.154	.108	.130

* 2nd and 3rd tertiles of lipoprotein-associated phospholipase A₂

ARIC Case-Cohort Study of Lp-PLA₂

Results of Cox Regression Analyses

Model term*	Standard weights			Calibrated weights			Estimated weights		
	Coef.	SE1	SE2	Coef.	SE1	SE2	Coef.	SE1	SE2
Age/10	0.420	.073	.075	0.393	.073	.012	0.431	.073	.015
Male	0.762	.088	.091	0.791	.088	.019	0.742	.088	.022
White	0.037	.098	.090	0.159	.099	.016	0.101	.100	.029
Frmr smoker	-0.421	.093	.126	-0.464	.092	.017	-0.459	.092	.020
Never smoke	-0.552	.099	.129	-0.557	.099	.016	-0.622	.099	.020
SBP/100	1.554	.207	.267	1.539	.208	.046	1.580	.207	.048
LDL-C/100	0.777	.106	.151	0.786	.106	.045	0.748	.107	.048
HDL-C/100	-2.539	.329	.392	-2.361	.329	.052	-2.736	.334	.060
Diabetes	0.572	.092	.127	0.738	.090	.019	0.531	.093	.026
Lp-PLA₂ (2)*	0.052	.110	.126	0.054	.111	.127	0.050	.111	.127
Lp-PLA₂ (3)*	0.163	.108	.129	0.182	.108	.130	0.154	.108	.130

* 2nd and 3rd tertiles of lipoprotein-associated phospholipase A₂

ARIC Case-Cohort Study: Interaction of Lp-PLA₂ and SBP

Model term*	Standard weights			Calibrated weights			Estimated weights		
	Coef.	SE1	SE2	Coef.	SE1	SE2	Coef.	SE1	SE2
Lp-PLA ₂ (2)	0.137	.118	.130	0.139	.119	.131	0.138	.118	.131
Lp-PLA ₂ (3)	0.303	.121	.132	0.306	.122	.131	0.299	.121	.131
Lp-PLA-SBP	-0.672	.204	.302	-0.681	.205	.274	-0.692	.205	.274

* Results for adjustment variables not shown

Summary of ARIC Analyses

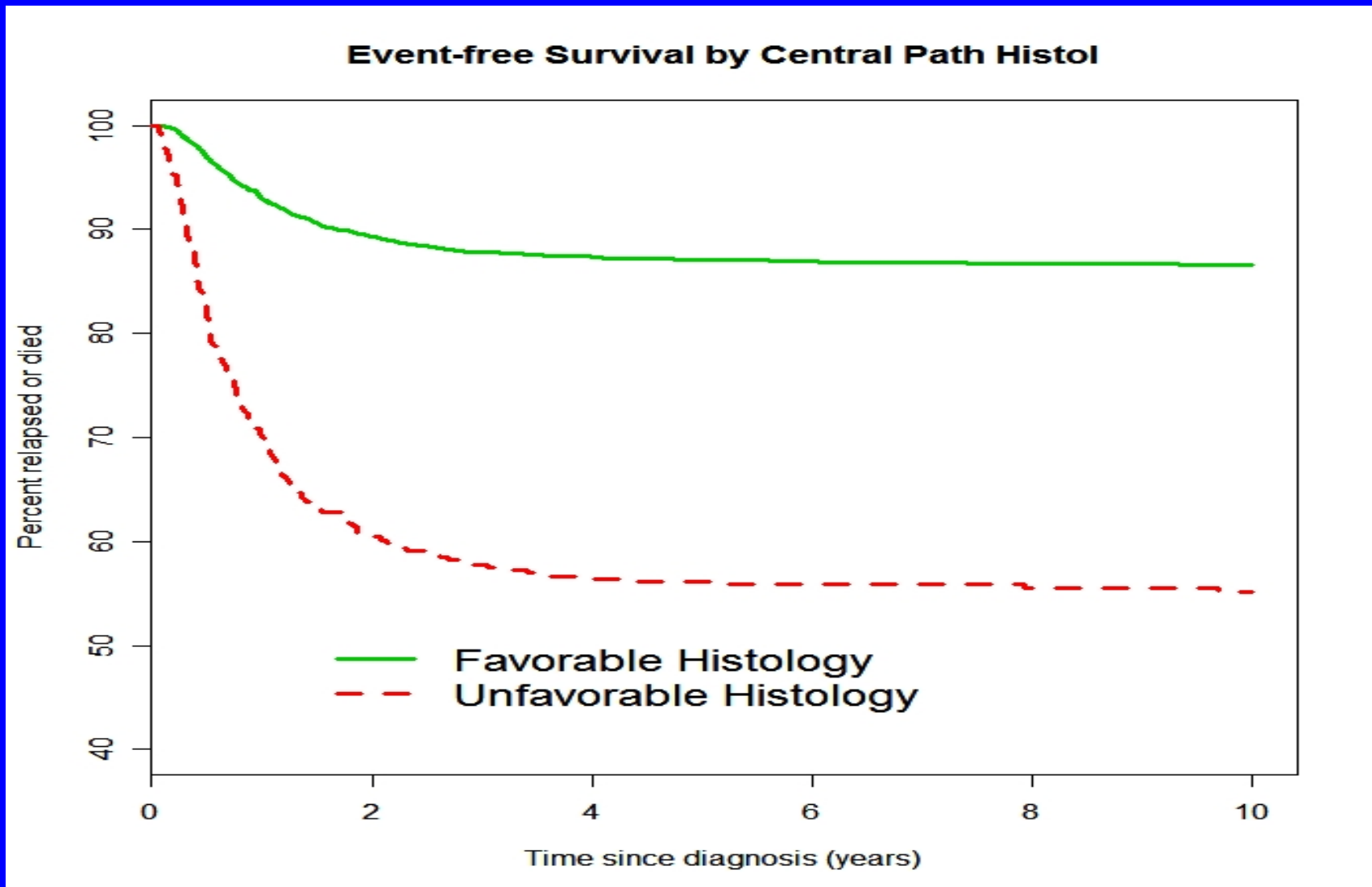
- Weak prediction of Phase II variable (Lp-PLA₂)
- No improvement in precision of main effect
- Dramatic improvement in precision of coefficients of adjustment variables
- Modest but significant improvement in precision of interaction between Phase I variable known for all and Phase II variables
 - Reduction of 10% in standard error
- **Adjustment of weights adds value to analysis**

Example II: National Wilms Tumor Study

- Cohort: 3,915/4,088 patients from NWTs-3,4
 - known values for outcome and selected covariates
- Outcome: “event free survival”
 - time to relapse, progression or (toxic) death
- Covariates available for everyone (from institution)
 - Stage (extent of disease)
 - “Favorable” vs. “Unfavorable” histology
 - age (years) at diagnosis
 - tumor diameter (cm)
- More costly data potentially available
 - Central Pathology evaluation of histology
 - in fact known for all 3,915 subjects

Relapse-free Survival by Histology

NWTS-3,4 (1980-94), Stage I-IV



Institutional vs. Central Pathology

Central Lab Pathology	<u>Institutional Pathology</u>		Pct. Miss- classified
	Favor	Unfavor	
Favorable	3,418	58	2%
Unfavorable	115	324	26%

Suggests Two Phase Design

- Limit number of subjects for whom slides sent to Central Laboratory
- Stratify sampling on basis of outcome and rare covariate patterns (esp. unfavorable histology)

NWTS: Stratified Sampling Design*

Main Cohort

	Favorable Histology (inst.)				Unfavorable Histology (inst.)				Tot
	Stage I,II		Stage III,IV		Stage I,II		Stage III,IV		
	<1	≥1	<1	≥1	<1	≥1	<1	≥1	
Case	57	232	10	208	15	41	29	77	669
Cont	452	1620	40	914	12	107	2	99	3246
% relap	11.2	12.5	20.0	18.5	55.5	27.7	93.5	43.8	17.1

Cases + Sub-cohort

Case	57	232	10	208	15	41	29	77	669
Cont	120	160	40	120	12	107	2	99	660

- Sample 100% of cases, UH, young stage III,IV
- Sample remaining 3 strata @ 27%, 10% and 13%
- Four strata for analysis (combine those @ 100%)

Simulations using R Survey Package*

- Fit Cox model to main cohort, robust SE
 - Central Lab pathology in fact available for all
- Replication: 1000 stratified case-control samples
- Three analyses of each of 1000 Phase II datasets
 - **Stratified** case-cohort analysis (Borgan, *LIDA*, 2000)
 - **Calibrated** to totals of imputed *delta-betas*
 - **Estimated** using strata and imputed *delta-betas*
- Summary statistics over 1000 replications
 - mean coefficients, mean estimated SE, root mean squared error compared with main cohort coefficients

Prediction of Missing Variable

- Logistic model for $\text{Pr}(\text{UH})$ has covariates
 - Local institutional histology
 - Stage (IV vs 1-III)
 - Age (>10 vs ≤ 10 years)
 - Study (NWTS-4 vs NWTS-3)
 - Interaction of local instit histol and stage
- Since these covariates are known for all in main cohort, can use fitted model to impute (predict) Central Lab histology there

R Code to Construct Sampling Design

```
> dstrat<-
twophase(id=list(~1,~1),strata=list(NULL,~strt),subset=~ins,data=nwt)
> summary(dstrat)
Two-phase design: twophase(id = list(~1, ~1), strata = list(NULL,
~strt), subset = ~ins, data = nwt)
Phase 1: Independent Sampling design (with replacement) ...
Phase 2: Stratified Independent Sampling design
svydesign(id = ~1, strata = ~strt, fpc = `*phase1*`)
Probabilities:
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.09877 0.26550 1.00000 0.74670 1.00000 1.00000
Stratum Sizes:
      1  2  3  4  5  6  7  8  9
obs   120 160 40 120 12 107 2 99 669 ...
Population stratum sizes (PSUs):
      6  5  9  2  8  3  4  1  7
107  12 669 1620 99 40 914 452 2
Data variables:
[1] "trel"      "tsur"      "relaps"    "dead"      "study"     "stage"
    ...
```

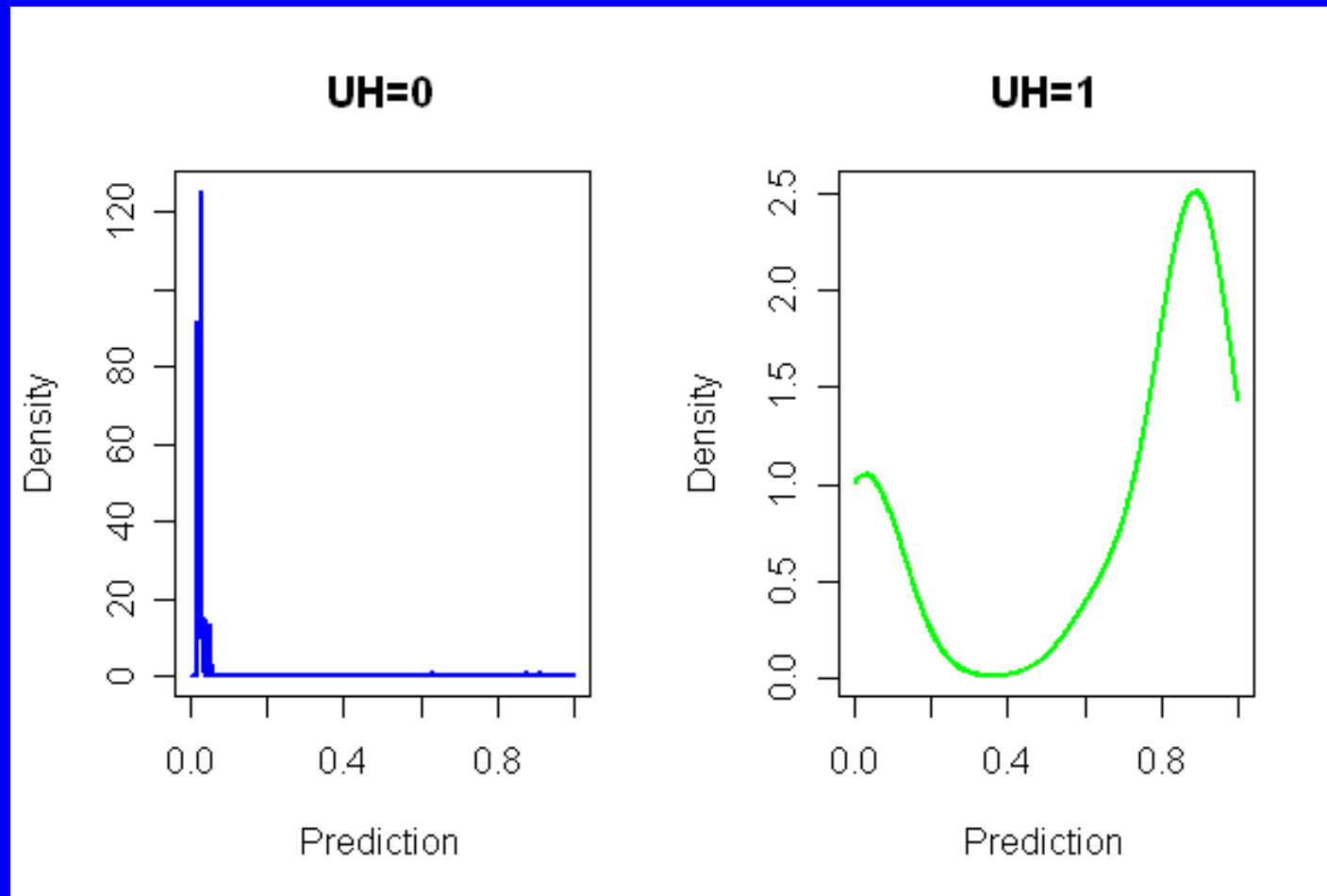
R Code to Predict Missing Variable

```
> Hmodel<-
svyglm(histol~instit*I(stage>3)+I(age>10)+factor(study),
+ family=quasibinomial,design=dstrat)
> sumH<-summary(Hmodel)
>
> round(sumH$coef,4)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.8761	0.3331	-11.6357	0.0000
instit	5.8490	0.3330	17.5642	0.0000
I(stage > 3)TRUE	0.6388	0.5693	1.1220	0.2621
I(age > 10)TRUE	0.2474	0.5329	0.4642	0.6426
factor(study)4	0.3370	0.3503	0.9620	0.3362
instit:I(stage > 3)TRUE	-2.0649	0.6499	-3.1771	0.0015

```
>
> nwt$estH<-predict(Hmodel,type="response",newdata=nwt,se=F)
```

Prediction of Central Lab Histology



R Code to Fit Calibration Model

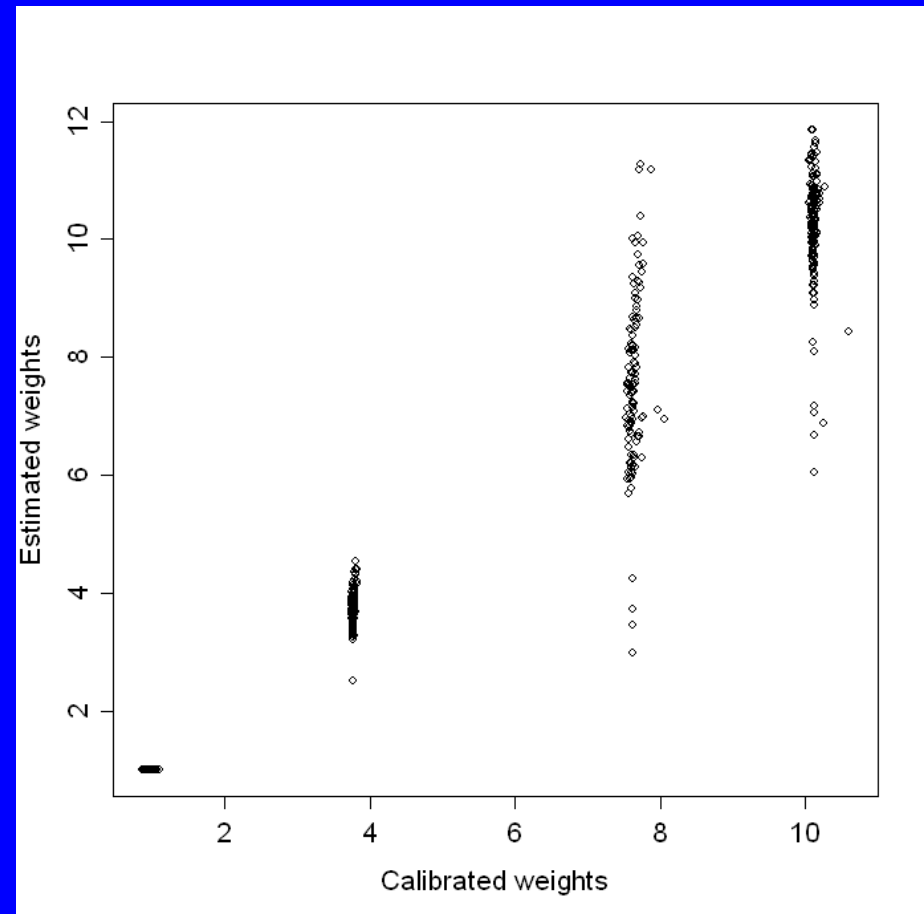
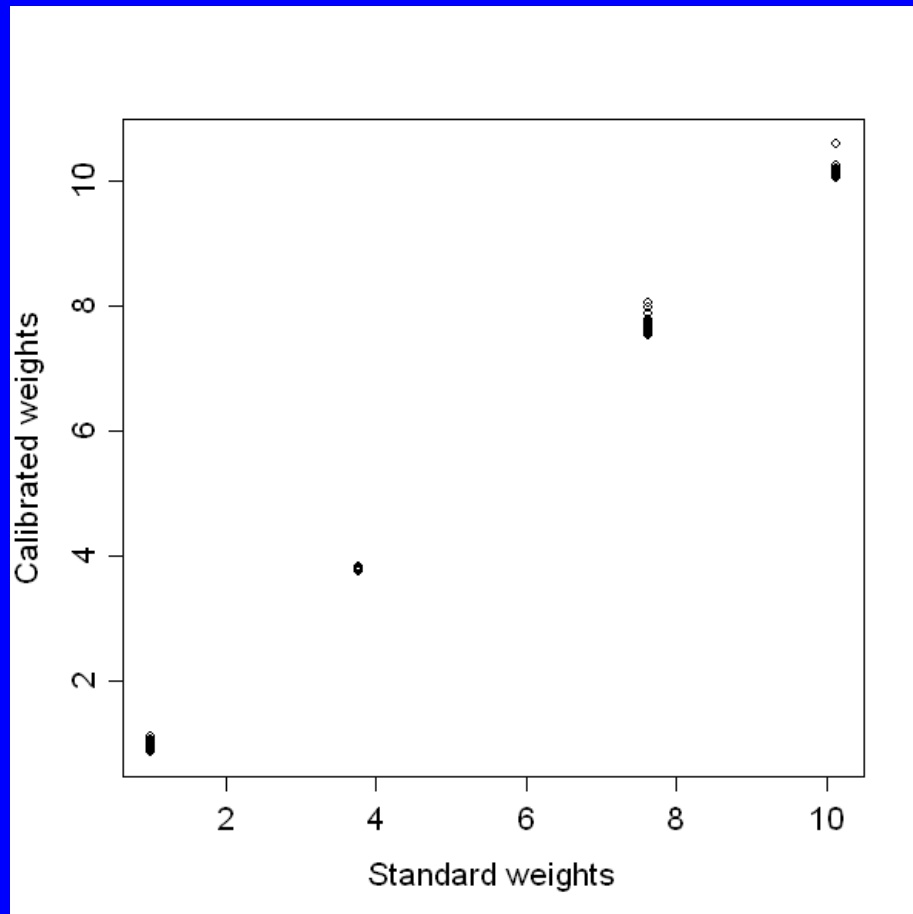
```
> calmodel<-  
coxph(Surv(trel,relaps)~estH*(age1+age2)+stg34+tumdiam+stgdiam,  
+ data=nwt)  
> sumcal<-summary(calmodel)  
> round(sumcal$coef,4)
```

	coef	exp(coef)	se(coef)	z	p
estH	4.5053	90.4989	0.4725	9.5354	0.00
age1	-0.7273	0.4832	0.3123	-2.3286	0.02
age2	0.1062	1.1121	0.0166	6.4090	0.00
stg34	1.4407	4.2235	0.2429	5.9311	0.00
tumdiam	0.0717	1.0743	0.0138	5.1994	0.00
stgdiam	-0.0841	0.9193	0.0192	-4.3772	0.00
estH:age1	-3.0089	0.0493	0.5490	-5.4805	0.00
estH:age2	-0.0711	0.9314	0.0512	-1.3886	0.16

R Code to Calibrate Sampling Model

```
> db<-resid(calmodel,"dfbeta")+1
> colnames(db)<-paste("db",1:ncol(db),sep=" ")
> nwtDB<-cbind(nwt,db)
> dstrt<-
twophase(id=list(~1,~1),strata=list(NULL,~strt),subset=~ins,
data=nwtDB)
> dcal<-calibrate(dstrt,formula=make.formula(colnames(db)),
pop=c(`(Intercept)`=3915,colSums(db)),calfun="raking",eps=0.0001
)
```

Comparison of Weights



NWTS: 1000 Simulated Phase II Samples

Summary Statistics for Phase Two Estimates

Model terms	Phase One Estimates		Standard Weights		Calibrated Weights		Estimated Weights	
	Coeff.	SE	ASE	RMSE	ASE	RMSE	ASE	RMSE
UH	4.042	0.503	.537	.188	.520	.143	.518	.144
Age ₀	-0.661	0.321	.360	.158	.326	.041	.324	.063
Age ₁	0.104	0.015	.026	.021	.017	.017	.017	.011
Stage	1.346	0.259	.346	.231	.271	.077	.271	.114
Diameter	0.069	0.015	.021	.015	.015	.012	.015	.008
Stg-Diam	-0.076	0.020	.029	.020	.021	.008	.021	.010
UH-Age ₀	-2.635	0.552	.612	.287	.592	.251	.590	.257
UH-Age ₁	-0.058	0.033	.051	.048	.049	.047	.048	.050

NWTS: 1000 Simulated Phase II Samples

Summary Statistics for Phase Two Estimates

Model terms	Phase One Estimates		Standard Weights		Calibrated Weights		Estimated Weights	
	Coeff.	SE	ASE	RMSE	ASE	RMSE	ASE	RMSE
UH	4.042	0.503	.537	.188	.520	.143	.518	.144
Age ₀	-0.661	0.321	.360	.158	.326	.041	.324	.063
Age ₁	0.104	0.015	.026	.021	.017	.017	.017	.011
Stage	1.346	0.259	.346	.231	.271	.077	.271	.114
Diameter	0.069	0.015	.021	.015	.015	.012	.015	.008
Stg-Diam	-0.076	0.020	.029	.020	.021	.008	.021	.010
UH-Age ₀	-2.635	0.552	.612	.287	.592	.251	.590	.257
UH-Age ₁	-0.058	0.033	.051	.048	.049	.047	.048	.050

Summary and Conclusions

- Excellent results even with stratified estimator
- Noticeable improvement with adjusted weights
 - Reduction in RMSE, sometimes to negligible levels
 - Little difference between calibration and estimation
 - Comparable to complicated “combined doubly weighted” estimator of Kulich & Lin (2004)
- Methods should be considered for general use
 - Especially useful when good surrogates available
 - Work needed on optimal choice of auxiliary variables

Selected References

1. Borgan O, Langholz B, Samuelsen SO, Goldstein G, Pogoda J: Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6**:39-58, 2001
2. Lumley T: *Survey Analysis in R*
<http://faculty.washington.edu/tlumley/survey>
3. Kulich M, Lin DY: Improving the efficiency of relative-risk estimation in case-cohort studies. *JASA* **99**:832-844, 2004
4. Mark SD, Katki HA: Specifying and implementing nonparametric and semiparametric survival estimators in two stage (nested) cohort studies with missing data *JASA* **101**:460-471, 2006
5. Breslow NE, Wellner JA: Weighted likelihood for semiparametric models and two phase stratified samples, with applications to Cox regression. *Scand J Statist* **34**:86-102, 2007 & **35**:186-192, 2008

Thank You!