Inference on Hazard Ratios and Survival Probabilities from Two-Phase Stratified Samples

Norman E Breslow[1], **Thomas Lumley**[2] **and Jon A Wellner**[1]

[1]U Washington, Seattle and [2]U Auckland, NZ

Fields Institute, Toronto

07 Dec 2012

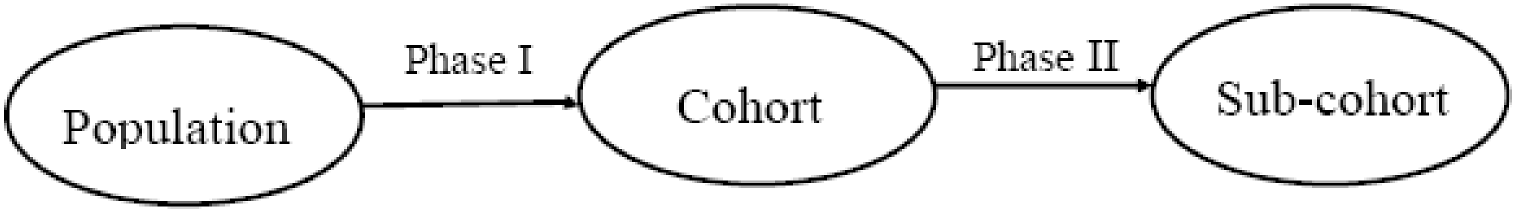
Sampling from Defined Cohort

- Large numbers of subjects in follow-up
    - Large cohort study
    - Surveillance of HMO population
- 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 tissue
        - biomarkers, gene expression levels
    - Detailed medical records abstraction
    - Second opinion on pathology specimens

Basic Questions

- How to select subjects for bioassay, or other detailed covariate ascertainment?
  - Simple random (validation) sample
  - Stratify on outcome (case-control, case-cohort study)
  - **Stratify jointly on outcome and covariates**
- How to analyze resulting data to provide “best” estimates of relative and absolute risk?
  - Maximum (pseudo)-likelihood
  - **Inverse probability weighting (IPW)**

Two Phase Sampling



Population may be finite or infinite

- finite ⇒ actual population (*e.g.*, population of Seattle)
- infinite ⇒ from probability model (superpopulation)

Sampling at phases I and II may be

- simple random or cluster sampling
- with or without stratification

R Survey Package (Lumley) accommodates all of above

Here consider only **infinite** population with

- simple random sampling at Phase I
- finite population stratified sampling at Phase II

Ex: National Wilms Tumor Study (NWTS)

- Main cohort: 3,915 patients from NWTS-3,4 (1980-94)
- Outcome: "event-free survival"
  - ▷ Event = relapse, progression or death from toxicity
- Covariates available for everyone (from institution)
  - ▷ "Favorable" (FH) *vs* "Unfavorable" (UH) histology
  - ▷ Stage (extent) of disease: I, II, III, IV
  - ▷ Age at diagnosis (years)
  - ▷ Tumor diameter (cm)
- More costly data only for selected subjects
  - ▷ Central Pathology evaluation of histology

Institutional vs Central Pathology

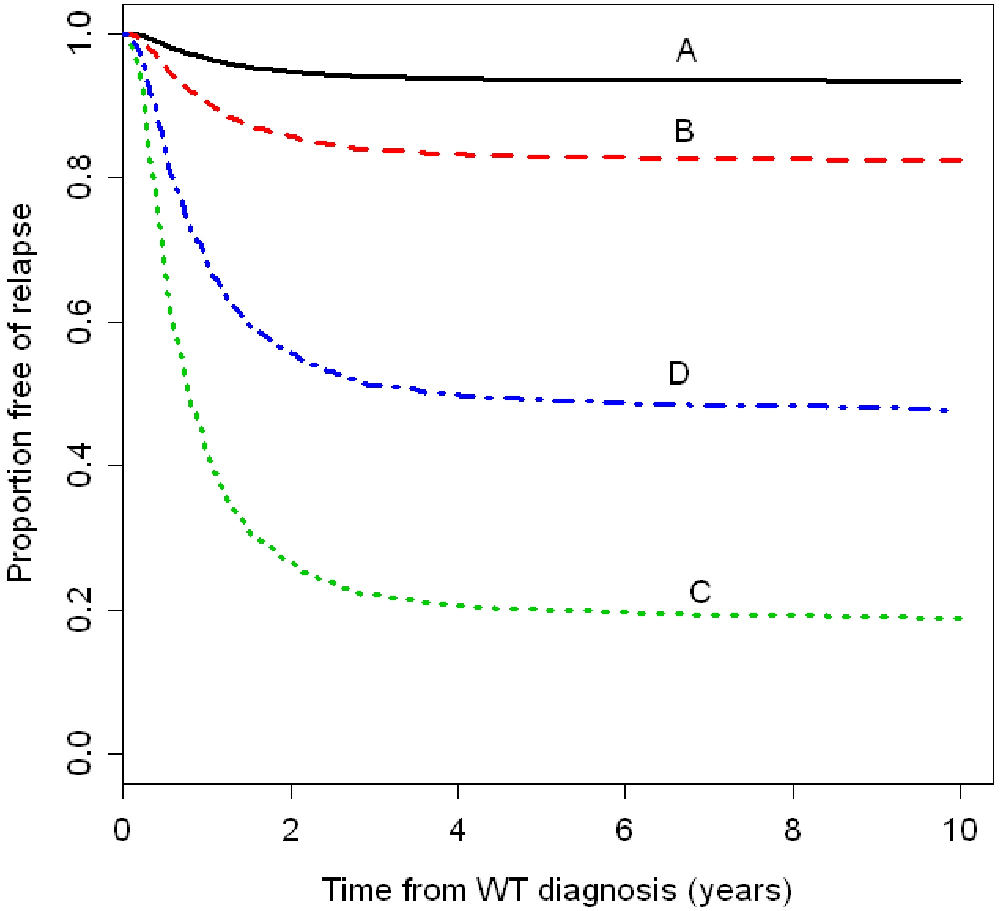
Central	Institutional Pathology		Percent
Pathology	Favorable	Unfavorable	missclassified
Favorable	3418	58	2%
Unfavorable	115	324	26%

Suggests Two Phase Design

- Phase I information
  - Institutional histology, stage, age, tumor diameter
  - Outcome: time to relapse or last seen
- Phase II information
  - Central Pathology histology
    - In fact available for everyone
    - Compare estimates based on full cohort data with those from simulated two-phase samples

Prediction of Relapse: Full Cohort

	Patient			
	A	B	C	D
UH	0	0	1	1
Age	1	4	$\frac{1}{2}$	7
Stg	0	1	0	1
Diam	8	10	10	16



Two Phase Stratified Sampling Design*

	Favorable Histol (Instit)				Unfavor Histol (Instit)				
	Stage I,II		Stage III,IV		Stage I,II		Stage III,IV		
Age	<1	≥ 1	<1	≥ 1	<1	≥ 1	<1	≥ 1	Tot
Cases	57	232	10	208	15	41	29	77	669
Controls	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
Phase II Sample									
Cases	57	232	10	208	15	41	29	77	669
Controls	**120**	**160**	40	**120**	12	107	2	99	660

- Sample 100% of cases, UH (institutional), stage III,IV babies
- Sample 27%, 10% and 13% of three remaining strata
- Total Phase II sample size $\sim$ 1/3 of Phase I
- **Goal:** Approximate Cox model fit to all Phase I subjects

* Kulich & Lin, *JASA*, 2004

Notation for Two-Phase Sampling

- Stratify cohort into J strata using variables V known for all
- Count numbers $N_1, \ldots, N_J$ of subjects in each stratum
- Sample n_j of N_j (without replacement)
- $R_i = 1$ if subject i sampled from cohort
  - ▷ $\pi_i = \text{Pr}(R_i = 1) = n_j/N_j$ if subject i in stratum j
    - ○ $n_j/N_j \to p_j$ as $N \uparrow \infty$
  - ▷ R_i dependent within strata
    - ○ but exchangeable

Semiparametric Model $P_{\theta,\eta}$

After vdV – Van der Vaart, *Asymptotic Statistics* (1998), §25

Model $P_{\theta,\eta}(x)$ for $X \in \mathcal{X}$

- Parametric part $\theta \in \Theta \subset R^p$
- Nonparametric part $\eta \in H \subset \mathcal{B}$

Assumptions to guarantee $\sqrt{N}\left(\widehat{\theta} - \theta_0, \widehat{\eta} - \eta_0\right)$ asymptotically Gaussian under iid random sampling (complete data)

Cox model has

- $X = (T, \Delta, Z)$; $\quad 0 \leq T \leq \tau, \quad \Delta \in \{0,1\}, \quad Z \in R^p$
- θ = regression coefficients (log hazard ratios)
- $\eta = \Lambda$ = baseline hazard function

Inverse Probability Weighted Empirical Measure

Random sample $X_1, \ldots, X_N$ from $P_0 = P_{\theta_0,\eta_0}(X)$

Empirical measure $\mathbb{P}_N$: uniform measure on N observations

IPW empirical measure $\mathbb{P}_N^{\pi}$ puts masses $\{1/(N\pi_i)\}$ on n selected observations ($R_i = 1$)

- Analogous to **bootstrap** (sample from $\mathbb{P}_N$)

Expectations are (for f in "Donsker" class $\mathcal{F}$)

$$P_0 f = \int f(x) dP_0(x)$$

$$\mathbb{P}_N f = \frac{1}{N} \sum_{i=1}^{N} f(X_i)$$

$$\mathbb{P}_N^{\pi} f = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} f(X_i)$$

IPW Estimating Equations

- Usual likelihood scores (for θ)

$$\dot{\ell}_{\theta,\eta} = \frac{\partial \text{log} p_{\theta,\eta}}{\partial \theta}$$

- Score operator $B_{\theta,\eta}$ acting on $h \in \mathcal{H}$: maps "directions" from which paths η_t approach η into scores for main model

- Solve IPW likelihood equations (infinite dimensional)

$$\mathbb{P}_N^{\pi} \dot{\ell}_{\theta,\eta} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} \dot{\ell}_{\theta,\eta}(X_i) = 0 \quad (1)$$

$$\mathbb{P}_N^{\pi} B_{\theta,\eta} h = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\pi_i} B_{\theta,\eta} h(X_i) = 0 \ \forall h \ \in \ \mathcal{H} \quad (2)$$

IPW Estimation for the Cox Model

Joint solution of (1) and (2) leads to IPW versions of

- Cox "partial likelihood" equations for θ
- "Breslow" estimator of Λ

Agree with methods proposed by (*inter alia*)

- Borgan *et al.*, *LIDA* 2000
- Lin, *Bioka* 2000

Asymptotic Properties of IPW Estimator of θ

$$\begin{aligned}
\sqrt{N}\left(\widehat{\theta}_N - \theta_0\right) &= \sqrt{N}\left(\widetilde{\theta}_N - \theta_0\right) + \sqrt{N}\left(\widehat{\theta}_N - \widetilde{\theta}_N\right) \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \widetilde{\ell}_0(X_i) + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i} - 1\right)\widetilde{\ell}_0(X_i) + o_p(1) \\
\mathsf{Var}_{\mathsf{Tot}}\left(\widehat{\theta}_N\right) &= \mathsf{Var}_{\mathsf{Phase\ I}} + \mathsf{Var}_{\mathsf{Phase\ II}}
\end{aligned}$$

- $\widetilde{\theta}_N$ is **unobserved** MLE based on complete data
- $\widetilde{\ell}_0$ is semiparametric **efficient influence function**
- $\mathsf{Var}_{\mathsf{Phase\ II}}$ is **design based**: normalized error in IPW estimation of unknown finite population total $\sum_{i=1}^{N} \widetilde{\ell}_0(X_i)$
- Phase I and II contributions asymptotically independent

Asymptotic Variances for Stratified Sampling

$$\begin{aligned}
\mathrm{Var}_{\mathrm{A}}\sqrt{N}(\widehat{\theta}-\theta_0) &= \begin{cases} \tilde{\mathcal{I}}_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} \mathrm{E}_j\left(\tilde{\ell}_0^{\otimes 2}\right) & \text{Bernoulli sampling} \\ \tilde{\mathcal{I}}_0^{-1} + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j} \mathrm{Var}_j\left(\tilde{\ell}_0\right) & \text{finite pop sampling} \end{cases} \\
\text{where } \nu_j &= \Pr(V \in \mathcal{V}_j) \quad \text{(size of stratum } j\text{)} \\
\tilde{\mathcal{I}}_0 &= \quad \text{usual information (complete data)}
\end{aligned}$$

- E_j and Var_j denote **within stratum** expectation & variance
- Potentially large difference if strata correlated with $\tilde{\ell}_0$

Theory Behind Asymptotics

Basic tools:

- Exchangeably weighted bootstrap empirical process of Præstgaard & Wellner (1993)
- Z-estimator theorem (3.3.1) of vdV & Wellner (1996)

Basic idea: Separate calculations for design and for model

- Sampling design gives properties of IPW empirical process

$$\mathbb{G}_N^\pi = \sqrt{N}(\mathbb{P}_N^\pi - P_0)$$

- Likelihood calculations, which are same as for complete data problem, give properties of efficient influence function for θ and other quantities of interest

Weak Convergence of IPW Empirical Process

$$\mathbb{G}_N^\pi = \sqrt{N}\,(\mathbb{P}_N^\pi - P_0) \rightsquigarrow \mathbb{G} + \sum_{j=1}^J \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j \quad \text{in} \quad \ell^\infty(\mathcal{F})$$

where

- $\nu_j = \text{Pr(stratum } j)$
- $p_j =$ sampling fraction stratum j (lim n_j/N_j)
- $\mathbb{G}$ is P_0-Brownian bridge
- $\mathbb{G}_j$ is $P_{0|j}$-Brownian bridge (restricted to stratum j)

Since sampling is independent in different strata

$$(\mathbb{G}, \mathbb{G}_1, \ldots, \mathbb{G}_J) \quad \text{mutually independent}$$

Breslow and Wellner, *Scand J Statist* **34**:88-102, 2007

Application to Cox Model

With

- $N(t) = \Delta \cdot \mathbf{1[T \leq t]}$ the counting process
- $Y = \mathbf{1}[T \geq t]$ the “at risk” process
- and M the usual martingale

$$M(t) = N(t) - \int_0^t e^{Z\theta_0} Y(s) d\Lambda_0(s)$$

the likelihood scores are

$$\begin{aligned}
\dot{\ell}_0(X) &= \Delta Z - Z e^{Z^\top \theta_0} \Lambda_0(T) = \int_0^\tau Z dM \\
B_0 h(X) &= \Delta h(T) - e^{Z^\top \theta_0} \int_0^T h d\Lambda_0 = \int_0^\tau h dM \quad \forall \; h \in \mathcal{H}
\end{aligned}$$

where $h \in \mathcal{H} = \mathrm{BV}[0,\tau]$ corresponding to one-dimensional submodels of form $d\Lambda_t = (1 + ht) d\Lambda$

van der Vaart (1998, §25.12.1)

Application to Cox Model

Defining

$$S_0^{(0)} = P_0\left(e^{Z\theta}Y\right) \quad \text{and} \quad S_0^{(1)} = P_0\left(Ze^{Z\theta}Y\right)$$

the adjoint and information operators (vdV, 1998 §25.12.1) are

$$B_0^*\dot{\ell}_0 \;=\; S_0^{(1)},\;\; B_0^*B_0h \;=\; hS_0^{(0)} \quad \text{and} \quad \left(B_0^*B_0\right)^{-1}h \;=\; h/S_0^{(0)}.$$

Setting $m(t) = S_0^{(1)}/S_0^{(0)}(t) = P_0(Z|T = t, \Delta = 1)$, we obtain **efficient score**

$$\ell_0^* = \left[I - B_0\left(B_0^*B_0\right)^{-1}B_0^*\right]\dot{\ell}_0 = \int_0^\tau [Z - m(t)]\,dM(t),$$

efficient information

$$\tilde{\mathcal{I}}_0 = P_0\left(\ell_0^*\ell_0^{*\top}\right) \;=\; P_0e^{Z^\top\theta_0}\int_0^\tau [Z - m(t)]^{\otimes 2}Y(t)d\Lambda_0(t)$$

and **efficient influence function**

$$\tilde{\ell}_0 = \tilde{\mathcal{I}}_0^{-1}\ell_0^*$$

in agreement with Cox (1972)

Improve Efficiency of $\widehat{\theta}$ via Survey Methods

Problem: Design and analyze the Phase II sample to estimate the unknown finite population total $\tilde{\ell}_{\mathrm{Tot}} = \sum_{i=1}^{N} \tilde{\ell}_0(X_i)$

Solution: Construct **auxiliary** variables $C = C(V)$ **correlated with** $\tilde{\ell}_0(X)$ and use to

- Construct strata for Phase II sampling
- **Adjust** sampling weights (*design* weights) $d_i = 1/\pi_i$ to bring in Phase I information
  - ▷ **Calibration** of weights to Phase I totals of C

    (Deville & Särndal, *JASA*, 1992)
  - ▷ **Estimate** weights with parametric model $\pi_i = \pi(V_i; \alpha)$ *e.g.*, logistic regression of R on C and stratum indicators

    (Robins *et al.*, *JASA*, 1994)

Calibration of Sampling Weights

Choose *new weights* $w_i = g_i d_i$ as close as possible to *design weights* $d_i = \pi_i^{-1}$ in sense of

Distance measure $G(w, d)$, *e.g.*

$$G(w,d) = \begin{cases} (w-d)^2/2d & \text{(least squares)} \\ w\log(w/d) - w + d & \text{(raking)} \end{cases}$$

such that total of auxiliary variables exactly estimated, *i.e.*,

Minimize $\sum_{i=1}^N R_i G(w_i, d_i)$ subject to **constraints** known as

Calibration equations: $\sum_{i=1}^N R_i w_i C_i \; = \; \sum_{i=1}^N C_i$

Lagrange multipliers $\lambda = \widehat{\lambda}_N$ obtained in minimization

Adjusted weights: $g_i = 1 - \widehat{\lambda}_N^\top C_i$ or $e^{-\widehat{\lambda}_N^\top C_i}$

Asymptotic Variance with Calibrated Weights

$$\begin{aligned}
\mathsf{Var}_{\mathsf{A}}\left(\widehat{\theta}(\widehat{\lambda}_N)\right) &= \mathsf{Var}_{\mathsf{Phase\ I}} + \mathsf{Var}_{\mathsf{Phase\ II}} \\
&= \frac{1}{\sqrt{N}}\left[\mathsf{Var}\,\tilde{\ell}_0(X) + \sum_{j=1}^{J}\nu_j\left(\frac{1-p_j}{p_j}\right)\mathsf{Var}_j\left(\tilde{\ell}_0 - QC\right)\right]
\end{aligned}$$

$$\text{where}\quad QC = P_0\left(\tilde{\ell}_0 C^{\mathsf{T}}\right)P_0^{-1}\left(CC^{\mathsf{T}}\right)C$$

is **population least squares regression** of $\tilde{\ell}_0$ on C.

Choosing $C = \mathsf{E}\left(\tilde{\ell}_0|V\right)$ achieves **optimality** within class of *augmented inverse probability weighted* (AIPW) estimators under Bernoulli (iid) sampling

Estimated weights have similar asymptotic properties

Choice of Auxiliary Variables C for Calibration

Goal: Find C for all main cohort (Phase I) subjects to approximate E$(\tilde{\ell}_0|V)$

Suggestion from work of Kulich & Lin (*JASA*, 2004)

1. Develop (rich) parametric model $[X|V]$ (goal: prediction)
   - fit model $[X|V]$ to Phase II sample using IPW
2. Impute values $\widehat{X}_i$ for all in main cohort using above model
3. Fit model $P_{\theta,\eta}(X)$ to main cohort using imputed $\widehat{X}_i$
4. Construct C as "delta-beta" residuals from model 3)
   - surrogates for $\tilde{\ell}_0(X_i)$
5. Estimate θ using adjusted weights based on $\{C_i\}$

Imputation model 1) need not be correct for procedure to yield asymptotically valid inferences (model assisted)

Simulation Study based on Wilms Tumor Cohort

- Fit Cox model to entire cohort of 3,915 subjects
  - ▷ Central path lab histology in fact available for all
- Draw 10,000 independent Phase II samples, each containing all 669 cases and 660 sampled controls using stratified design
  - ▷ Fit prediction model $[X|V]$ using IPW logistic regression
  - ▷ Impute X for Phase I subjects and fit Cox model
  - ▷ Extract "delta-beta" residuals as calibration variables C
  - ▷ Fit Cox model to Phase II data using standard, calibrated and estimated weights
- RMSE of coefficients $\widehat{\theta}$ from two-phase samples, considered as estimates of coefficients $\tilde{\theta}$ from fit to Phase I sample (already obtained), are **empirical Phase II standard errors**

Stratified Case-Control Sampling Design*

Main Cohort

	Favorable Histology				Unfavorable Histology				
	Stage I,II		Stage III,IV		Stage I,II		Stage III,IV		
Age	<1	≥ 1	<1	≥ 1	<1	≥ 1	<1	≥ 1	Tot
Cases	57	232	10	208	15	41	29	77	669
Controls	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 + Cohort Random Sample									
Cases	57	232	10	208	15	41	29	77	669
Controls	**120**	**160**	40	**120**	12	107	2	99	660

- Sample 100% of cases, UH (institutional), stage III,IV babies
- Sample 27%, 10% and 13% of three remaining strata
- May combine strata sampled at 100% for analysis

* Kulich & Lin, *JASA*, 2004

Two Models Suggested by Kulich & Lin

1) Semiparametric model $P_{\theta,\eta}(X)$

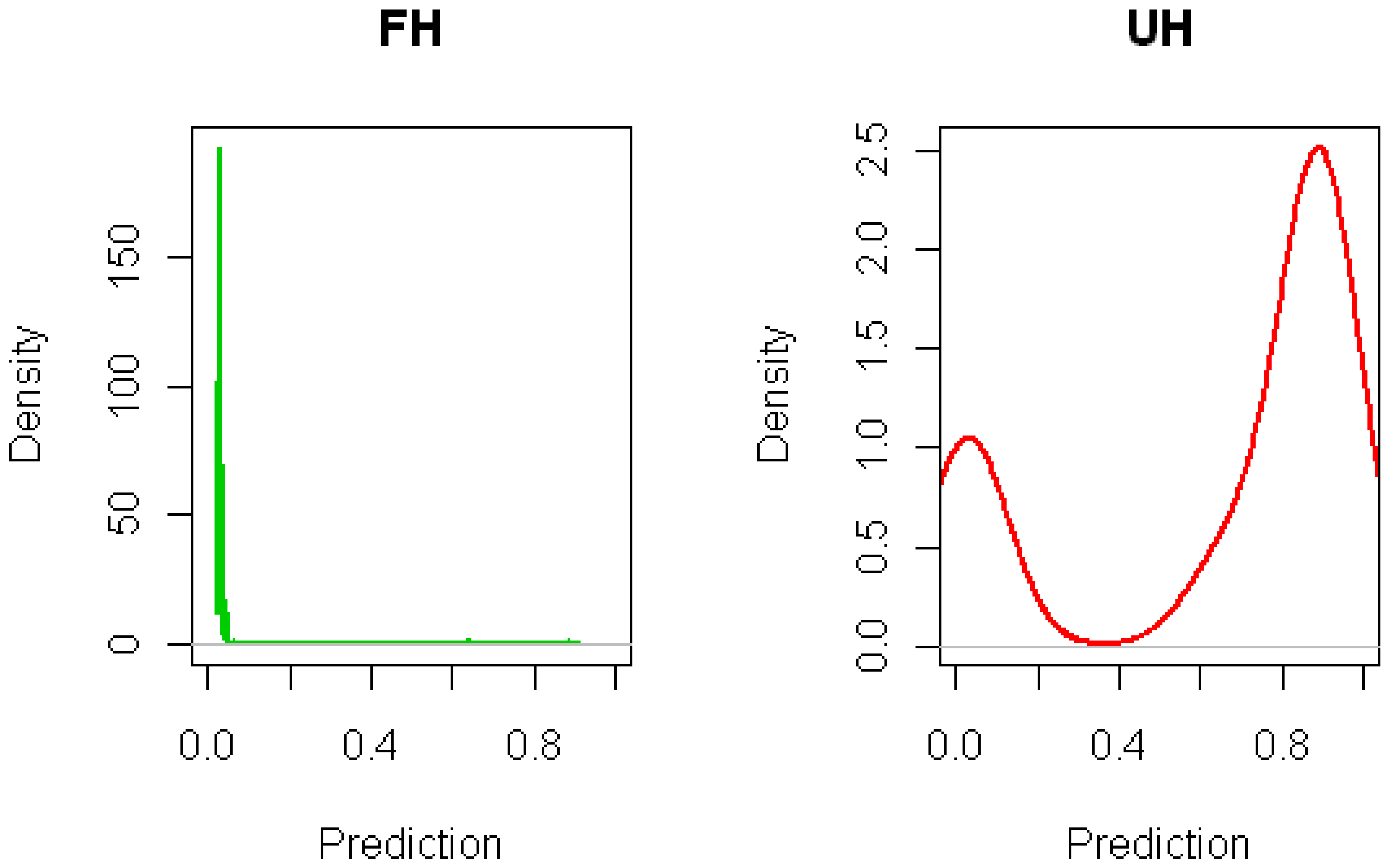
- Cox regression model for prognosis (event free survival) using covariates
  - ▷ Histology: unfavorable (UH) *vs* favorable – Central Path
  - ▷ Age: linear spline, knot at 1 yr
  - ▷ Stage: III-IV *vs* Stage I-II
  - ▷ Tumor diameter: linear
  - ▷ Interactions: histology × age; stage × diameter

Two Models Suggested by Kulich & Lin

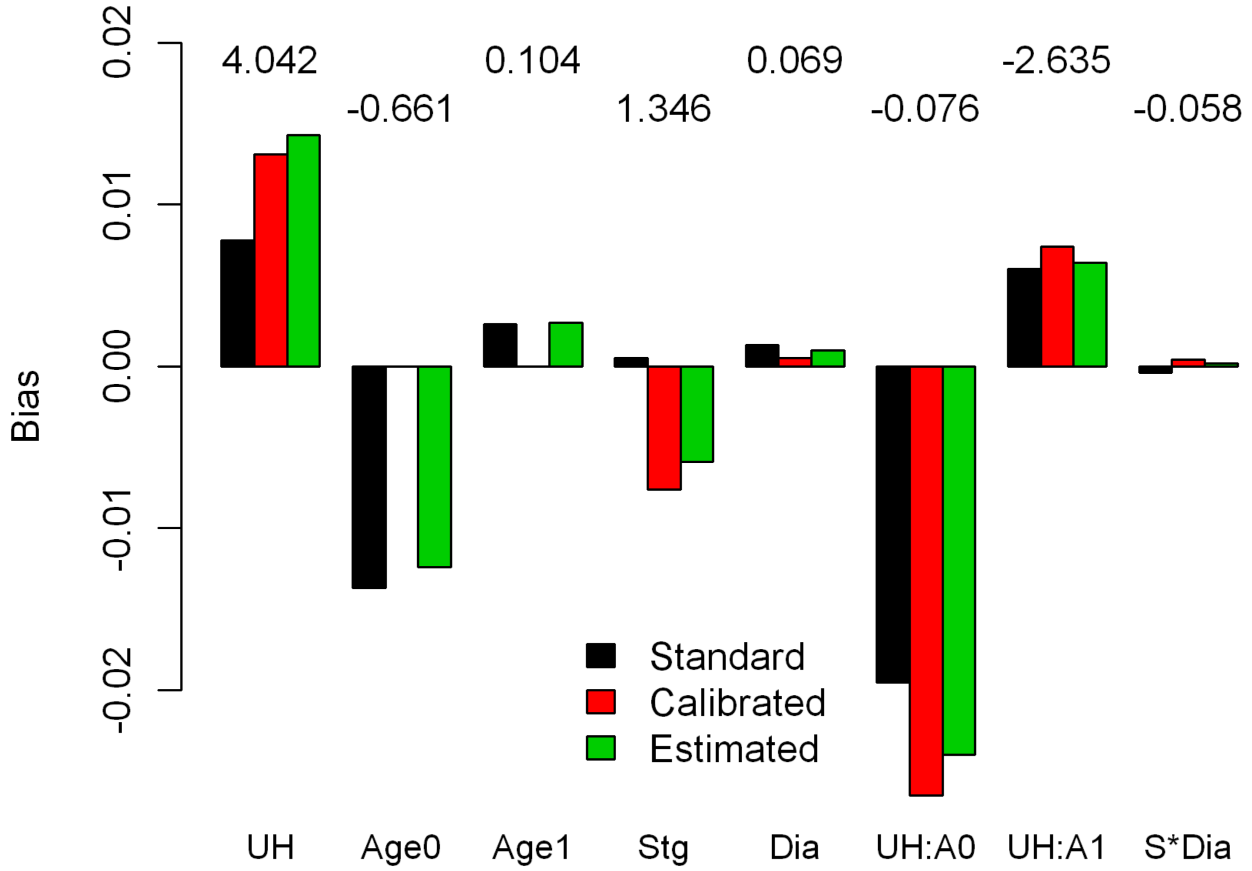
2) Parametric imputation model $[X|Z]$

- Logistic model for histology (1=UH) as function of
  - ▷ Local institutional histology
  - ▷ Stage IV *vs* Stage I-III
  - ▷ Age > 10 *vs* age ≤ 10
  - ▷ Study: NWTS-4 *vs* NWTS-3
  - ▷ Interaction of local histology and stage
- Other X's known for everyone

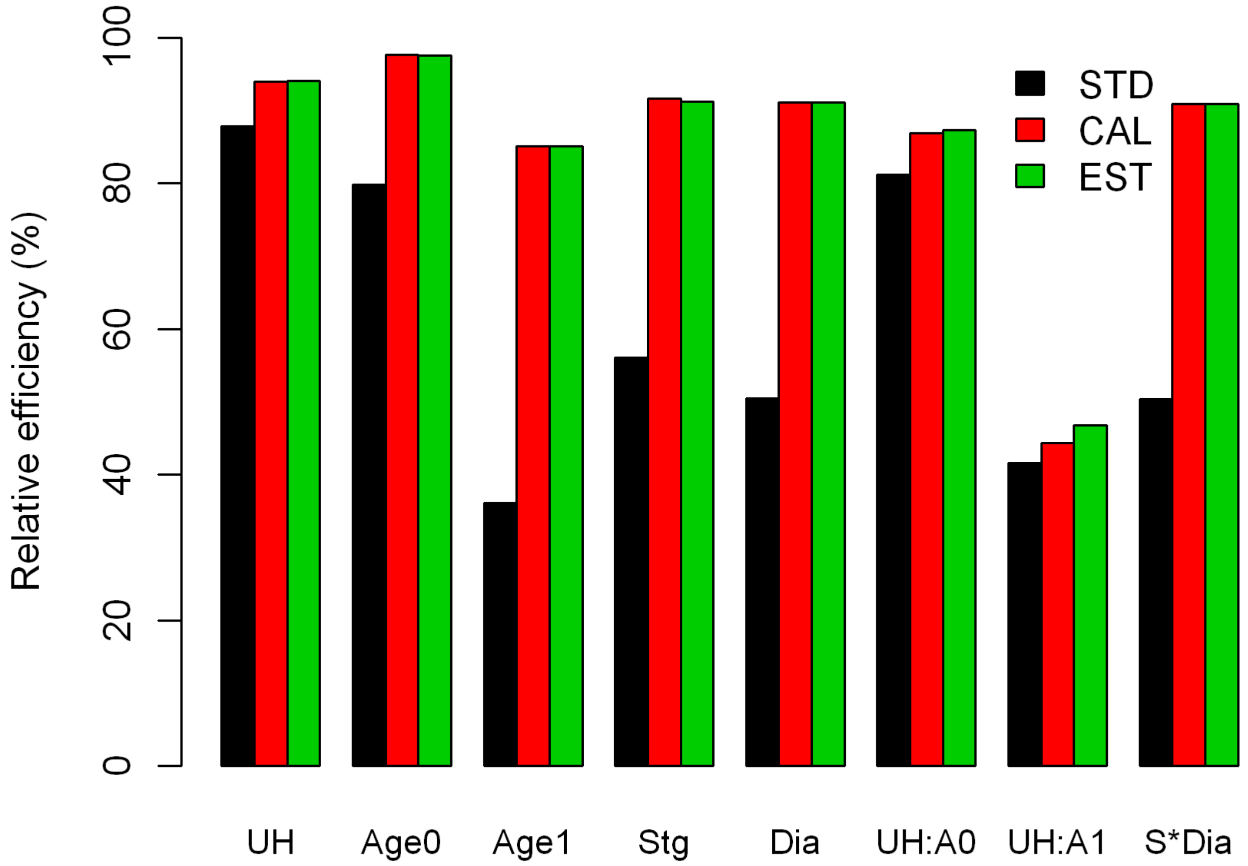
Two Models Suggested by Kulich & Lin



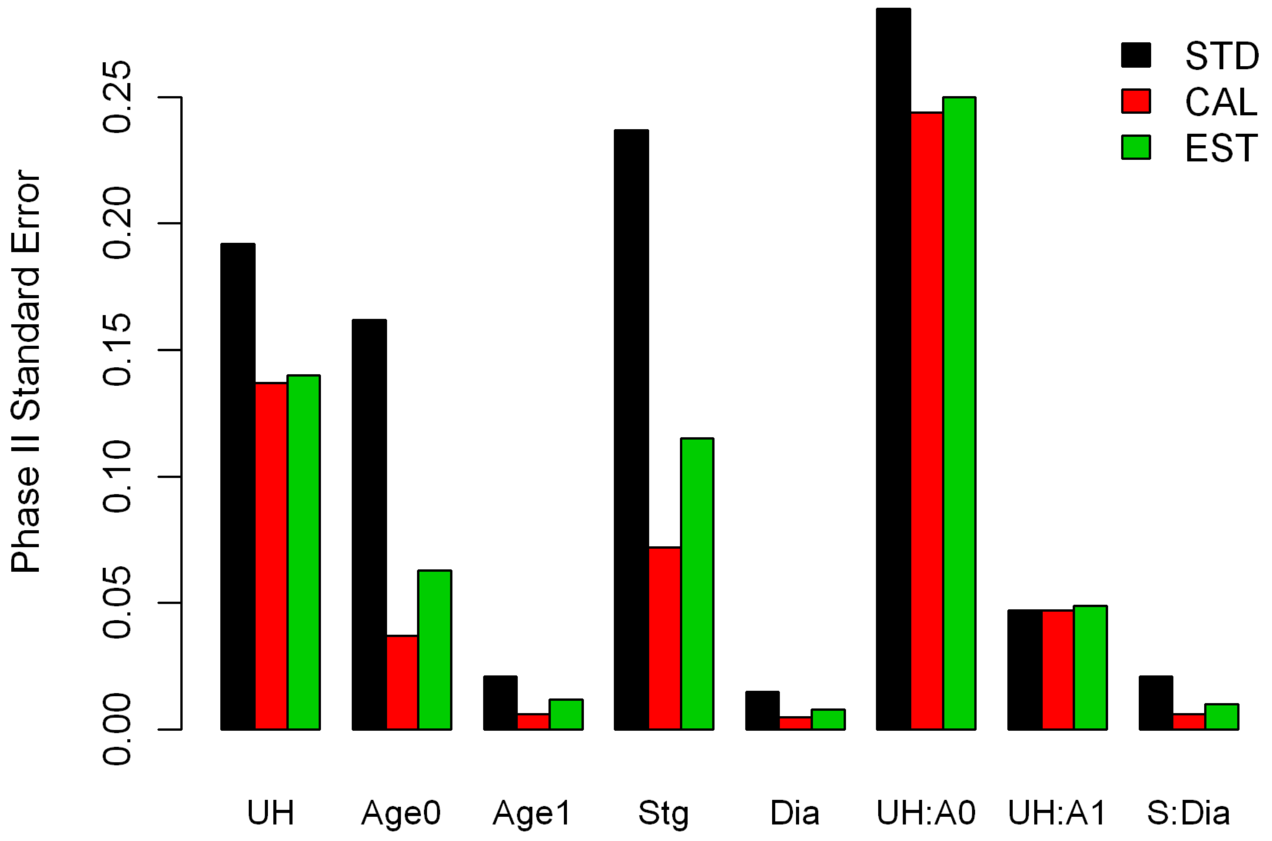
Results of Simulations of $\widehat{\theta}$: Bias



Efficiency relative to full data: $100 \cdot \left(\widehat{\text{var}}\hat{\theta}/\widehat{\text{var}}\tilde{\theta}\right)$



Empirical Phase II Std Error: RMSE $\widehat{\theta}$



Asymptotic Properties of IPW Estimator of η

Extending the results in vdV (1998) §25 (η a measure)

$$\sqrt{N}\,(\widehat{\eta}_N - \eta_0)\,h \;=\; \mathbb{G}_N^{\pi} Ah + o_p(1)$$

$$\sqrt{N}\left(\widehat{\eta}_N(\widehat{\lambda}_N) - \eta_0\right)h \;=$$

$$\mathbb{G}_N Ah + (\mathbb{G}_N^{\pi} - \mathbb{G}_N)\left\{Ah - P_0\left(AhC^{\mathsf{T}}\right)\left[P_0\left(CC^{\mathsf{T}}\right)\right]^{-1} C\right\} + o_p(1)$$

where the operator $A : \mathcal{H} \mapsto L_2(P_0)$ is given by

$$Ah \;=\; B_0\,(B_0^* B_0)^{-1}\,h - P_0\left[B_0\,(B_0^* B_0)^{-1}\,h\dot{\ell}_0^{\mathsf{T}}\right]\tilde{\ell}_0.$$

Conclusion: estimators of η with and without calibration of weights using variables C are asymptotically Gaussian.
Asymptotic variance for calibrated estimator is

$$\mathsf{Var}_{\mathsf{A}}\sqrt{N}\left(\widehat{\eta}_N(\widehat{\lambda}_N) - \eta_0\right)h$$

$$= \;\mathsf{Var}_0(Ah) + \sum_{j=1}^{J} \nu_j \frac{1-p_j}{p_j}\mathsf{Var}_j\left[Ah - \Pi(Ah|C)\right]$$

Application to Cox Model

For baseline hazard (Λ) estimation with calibrated weights:

$$\sqrt{N}\left(\widehat{\Lambda}_N - \Lambda_0\right)(t) \;\rightsquigarrow\; \mathbb{G}A_t + \sum_{j=1}^{J} \sqrt{\nu_j}\sqrt{\frac{1-p_j}{p_j}}\mathbb{G}_j\left(A_t - \Pi(A_t|C)\right)$$

$$\text{where} \quad A_t \;=\; \int_0^t \frac{dM}{S_0^{(0)}} - \left(\int_0^t m\,d\Lambda_0\right)\tilde{\ell}_0$$

$$\text{and} \quad \Pi(A_t|C) \;=\; P_0\left(A_t C^\top\right) P_0^{-1}\left(CC^\top\right) Z$$

(projection of A_t on $[C]$)

suggests using **additional calibration variables** of form

$$C_t = \widehat{\mathsf{E}}\left[\int_0^t \frac{dM}{S_0^{(0)}}\right] \quad \text{for} \quad t = 1, 2, 5, 10$$

Estimation of Individual Survival Proportions

Delta method applied to estimated cumulative hazard for subject with covariates $Z = z_0$ gives

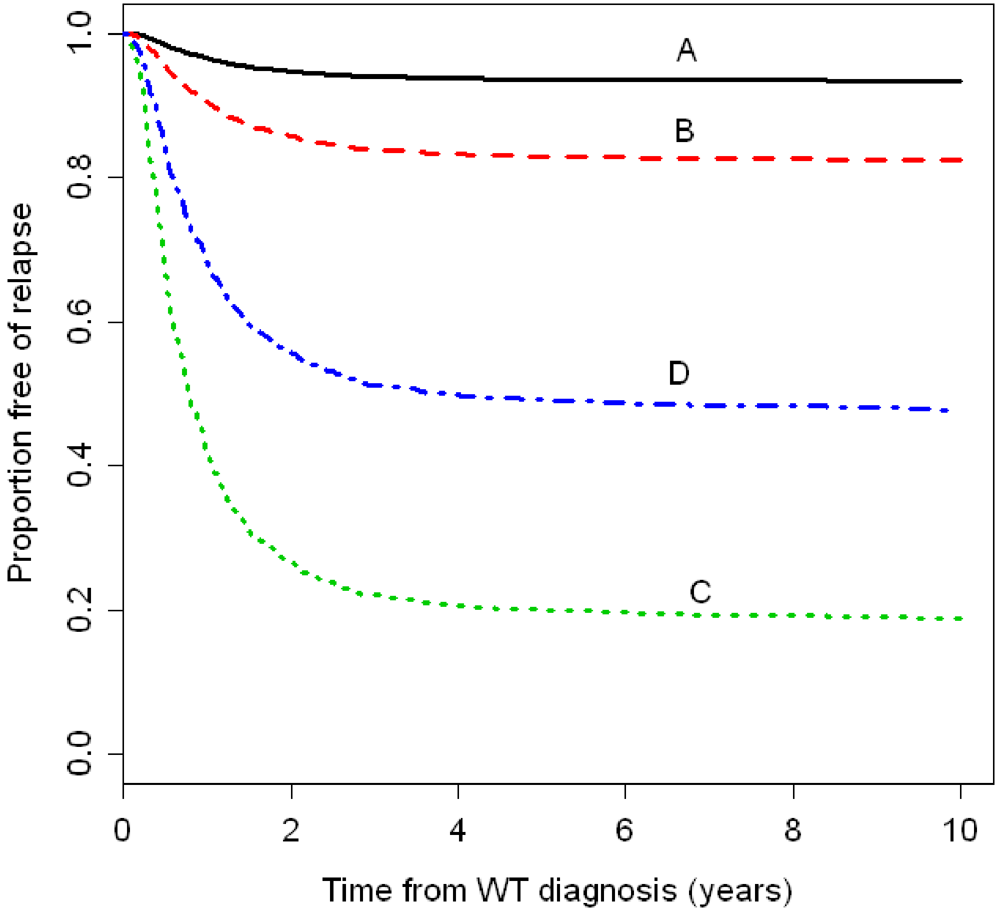
$$\sqrt{N}\left[e^{z_0\widehat{\theta}}\widehat{\Lambda}(t) - e^{z_0\theta}\Lambda_0(t)\right]$$

$$= e^{z_0\theta}\mathbb{G}_N^{\pi}\left[\int_0^t \frac{dM}{S_0^{(0)}} + \left(\int_0^t [z_0 - m]d\Lambda_0\right)\tilde{\ell}_0\right] + o_p(1)$$

- Results for simple random sampling obtained by replacing $\mathbb{G}_N^{\pi}$ with $\mathbb{G}_N$
- Generalizes work of Tsiatis (1981), Andersen and Gill (1982) and Begun, Hall, Huang and Wellner (1983) to IPW estimation with two phase stratified samples.

Prediction of Relapse: Full Cohort

	Patient			
	A	B	C	D
UH	0	0	1	1
Age	1	4	$\frac{1}{2}$	7
Stg	0	1	0	1
Diam	8	10	10	16



Simulation Study (Continued)

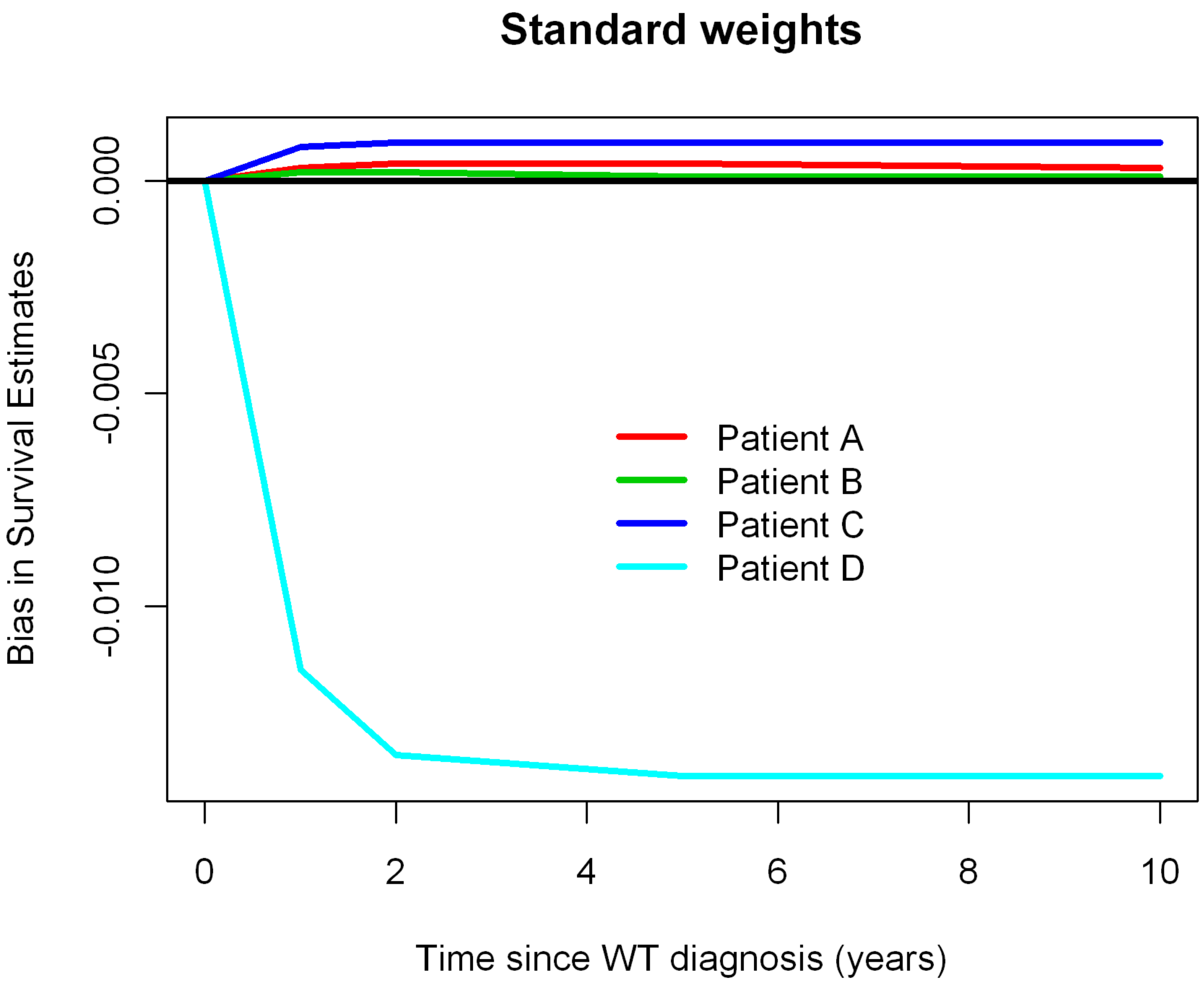
For each of 10,000 Phase II samples

- Construct calibration variables as "delta-betas" for Cox model fit to imputed Phase I data (as before)
- Using $\widehat{\Lambda}$ and $\widehat{S}_0^{(0)}$ from same imputed data fit, construct additional calibration variables ($t = 1, 2, 5, 10$)

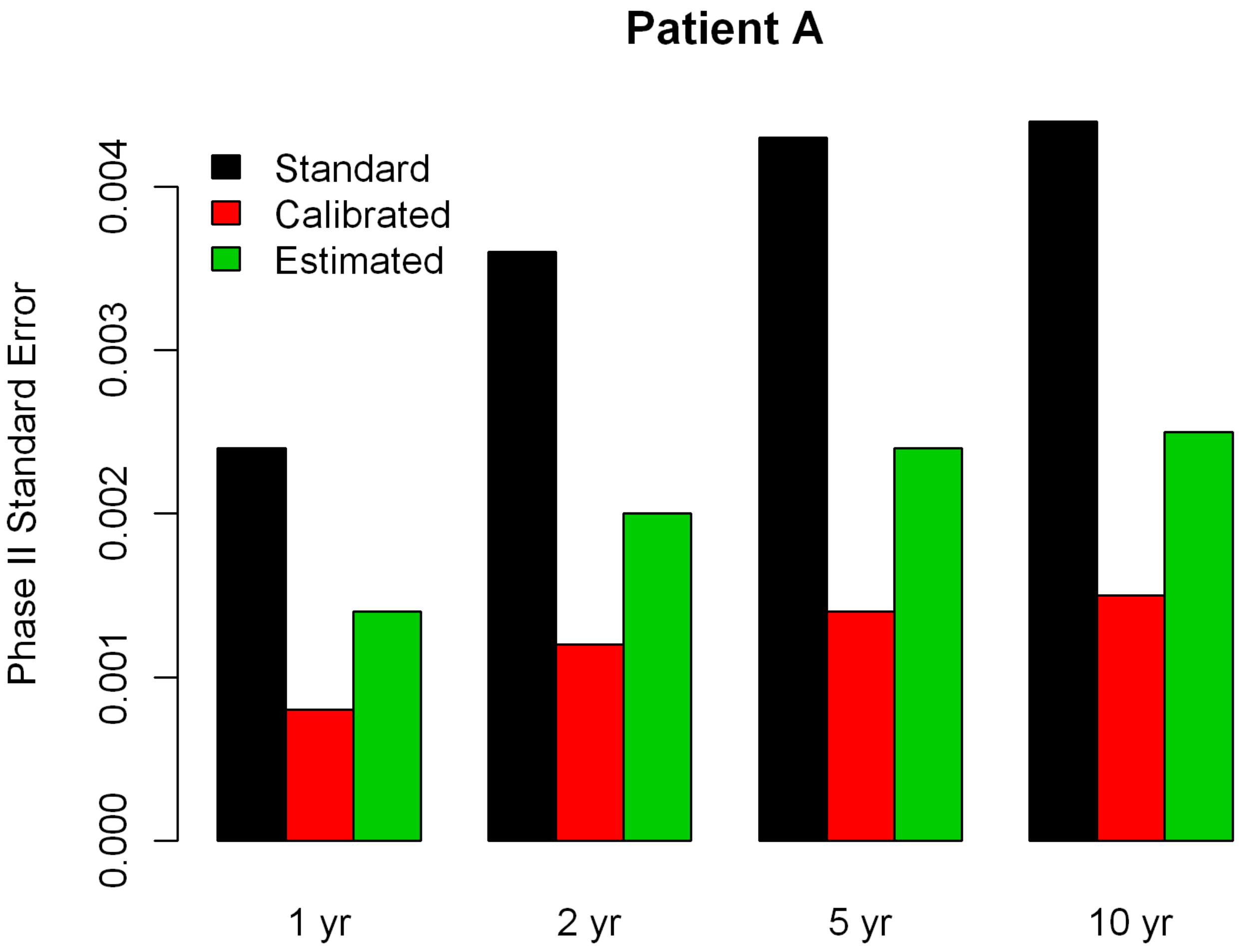
$$C_t = \int_0^t \frac{d\widehat{M}_i}{\widehat{S}_0^{(0)}} = \frac{\Delta_i \mathbf{1}[T_i \leq t]}{\widehat{S}_0^{(0)}(T_i)} - e^{Z_i^\top \widehat{\theta}} \int_0^{t \wedge T_i} \frac{d\widehat{\Lambda}}{\widehat{S}_0^{(0)}}$$

- For estimation of $e^{z_0 \theta_0} \Lambda_0(t)$, add C_t to "delta-betas" to calibrate the weights
- Fit the Cox model to Phase II data using standard, calibrated and estimated weights
- Determine RMSE of estimates of survival proportions estimated using standard and adjusted weights

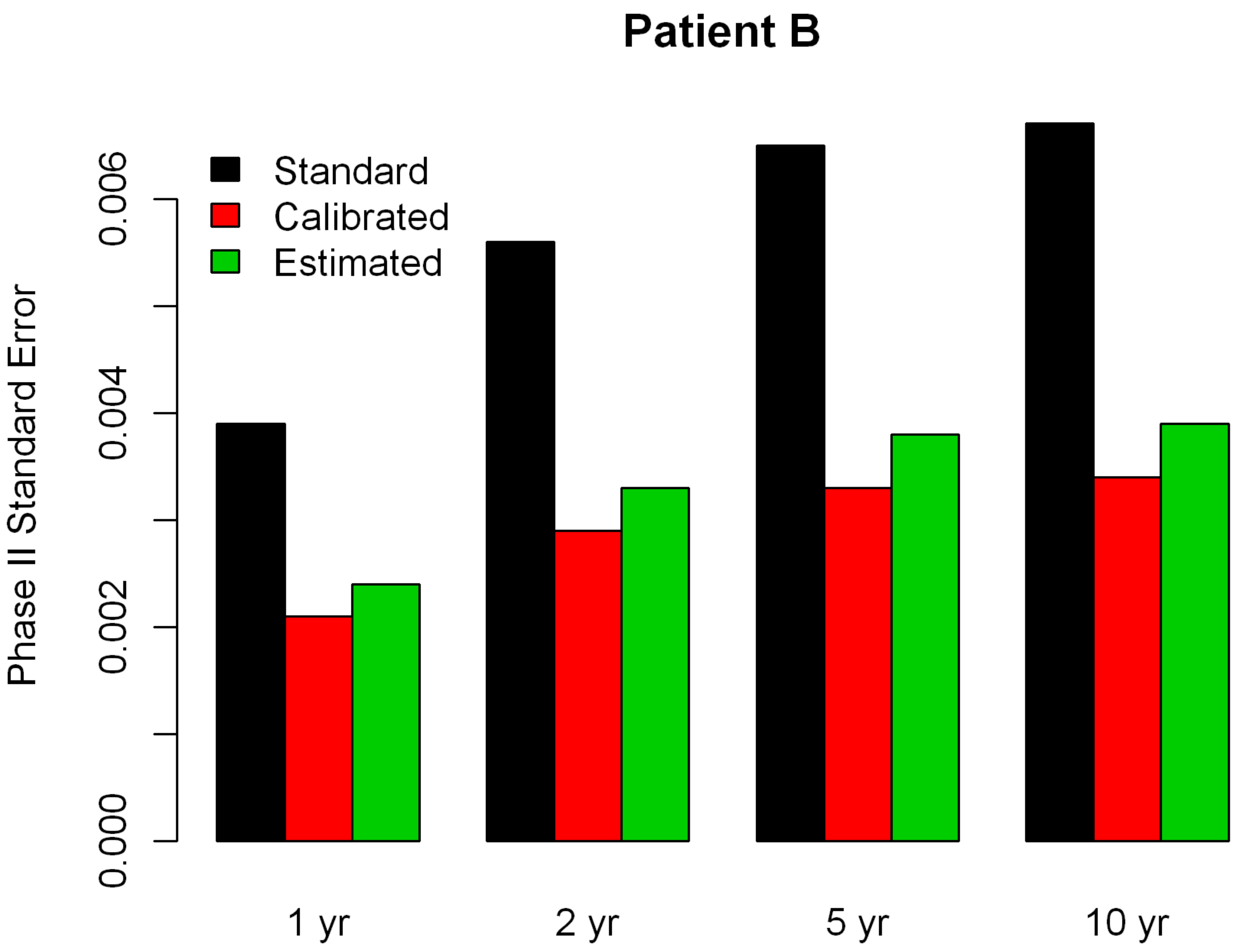
Bias in Survival Proportions: Standard Weights



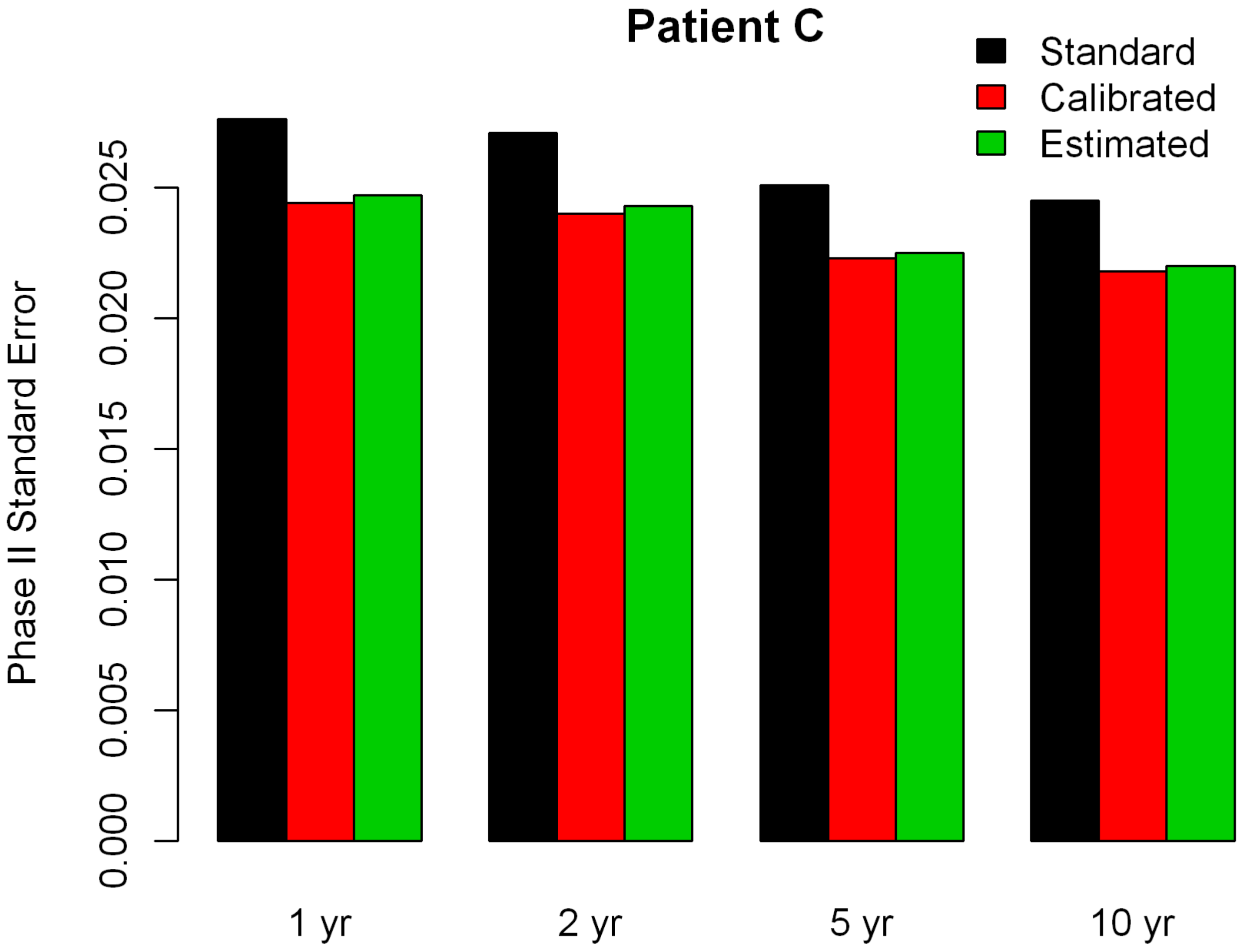
RMSE in Survival Proportions: Patient A



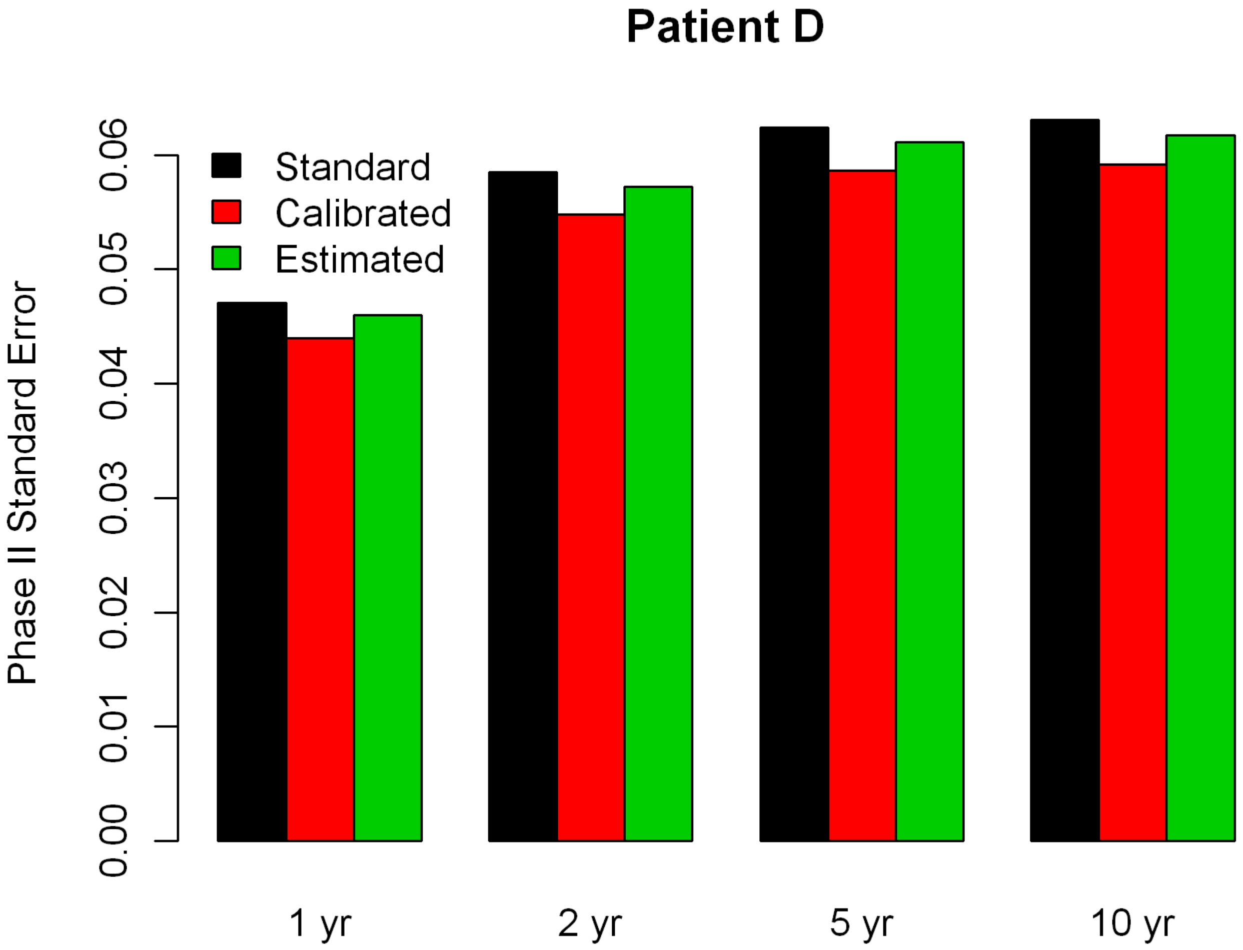
RMSE in Survival Proportions: Patient B



RMSE in Survival Proportions: Patient C



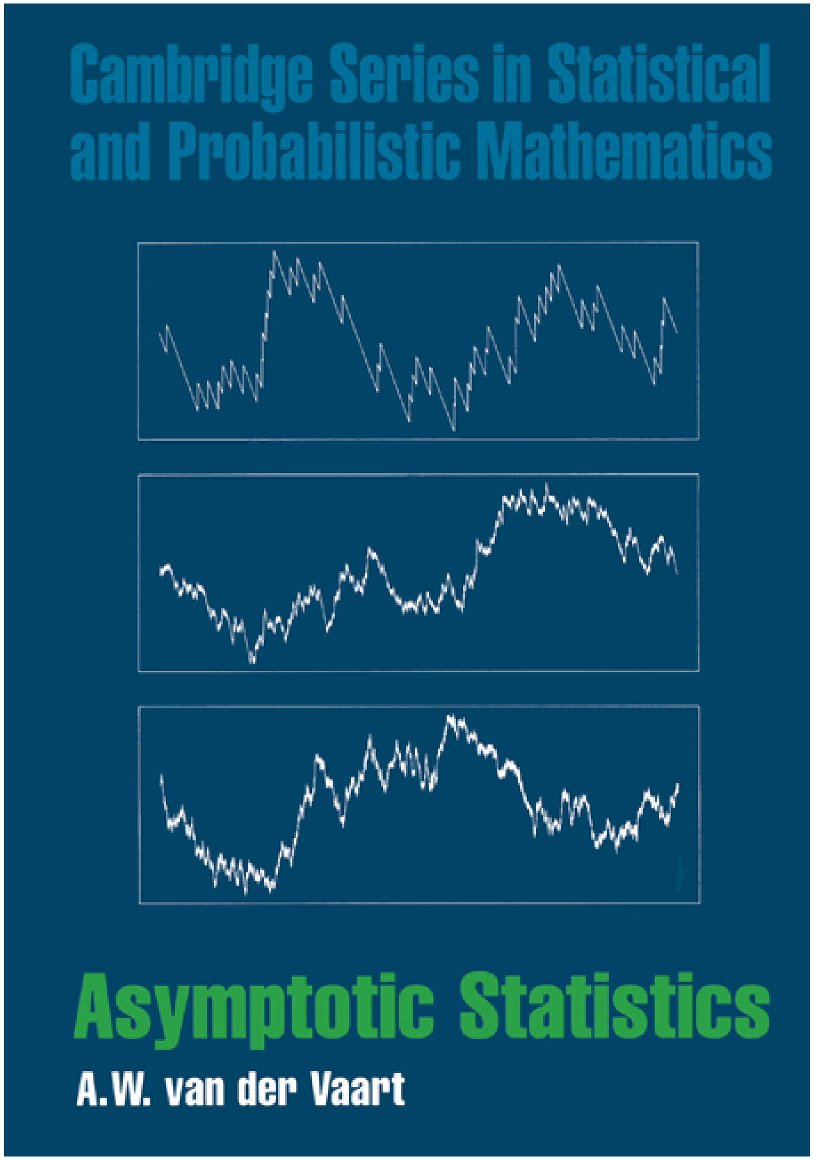
RMSE in Survival Proportions: Patient D



Conclusions

- Calibration/estimation of weights improves efficiency
  - ▷ Reduces Phase II error, often to neglible levels, for coefficients of covariates known at Phase I
  - ▷ Reduces Phase II error moderately for coefficients of other covariates **provided** good surrogates available for them
    - ○ No improvement otherwise
  - ▷ Robust to misspecification of imputation model
  - ▷ **Improves model based predictions**
  - ▷ Should be more widely used
- More research needed
  - ▷ Complex sampling designs generally
  - ▷ Other models (GEE)
  - ▷ Other choices for calibration variables

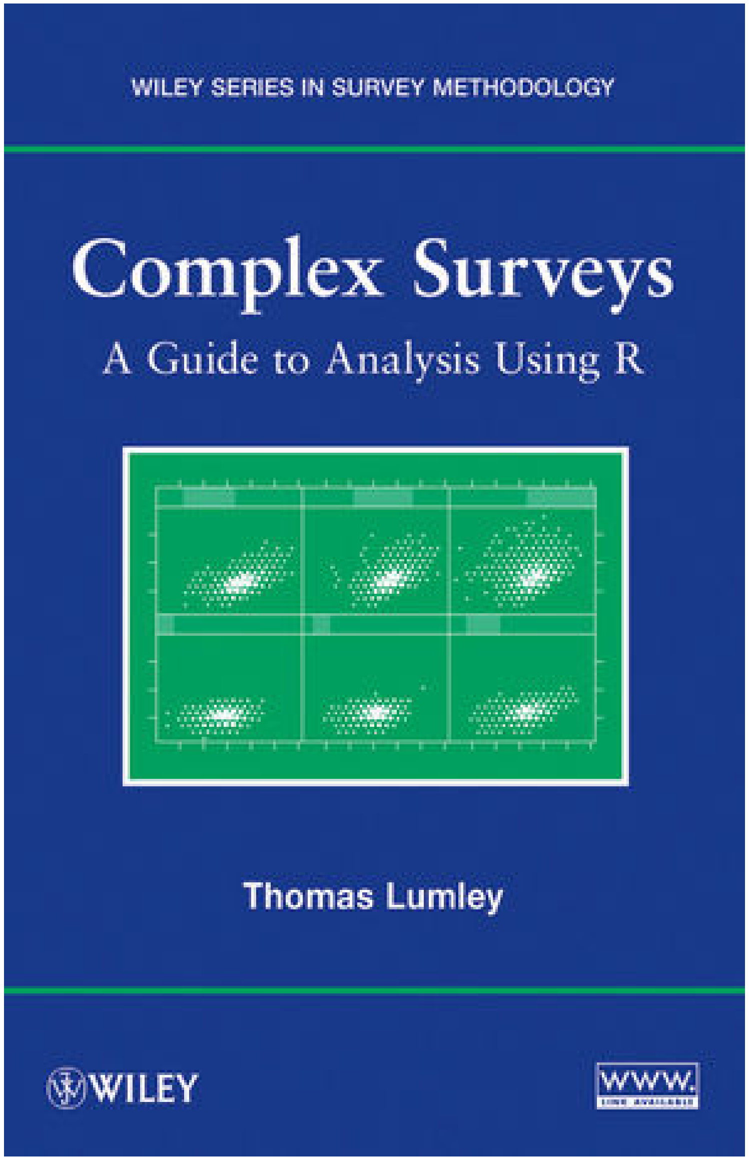
Major References



Cambridge Series in Statistical and Probabilistic Mathematics

Asymptotic Statistics

A.W. van der Vaart



WILEY SERIES IN SURVEY METHODOLOGY

Complex Surveys

A Guide to Analysis Using R

Thomas Lumley

WILEY

Additional References

- Borgan, Langholz *et al.* *Lifetime Data Anal* **6**:39-58, 2000
- Lin *Biometrika* **87**:37-47, 2000
- Mark, Katki *J Amer Statist Assoc* **101**:460-471, 2006
- Mark, Katki *NestedCohort* R package `http://dceg.cancer.gov/tools/analysis/nested-cohort`
- Breslow, Wellner *Scand J Statist* **34**:88-102, 2007
- Breslow, Wellner *Scand J Statist* **35**:186-192, 2008
- Lumley *Survey Analysis in R* `http://faculty.washington.edu/tlumley/survey/`
- Breslow, Lumley *et al.* *Amer J Epi* **169**:1398-1405, 2009
- Breslow, Lumley *et al.* *Statist Biosc* **1**:32-49, 2009
- Lumley *Complex Surveys*, Wiley, 2010
- Breslow, Lumley *IMS Monograph Series* (Wellner Festschrift)

Issues for Discussion

- What is the population parameter of primary scientific interest?
- How does one identify this parameter in the inevitable situation where the specified statistical model is at best an approximation to the truth?

Issues for Discussion

- What is the appropriate tradeoff between statistical efficiency, assuming the model is correct, and robustness to model misspecification?

Issues for Discussion

- How important is it to optimize the sampling design using approximations to Neyman allocation, or other similar criteria, instead of using a more intuitive approach that attempts to sample roughly equal numbers of subjects within phase two strata?

Issues for Discussion

- How often does one encounter situations where multiple analyses are required for the same sample, e.g., using different time scales in Cox regression, or where secondary analyses are required of phase two data collected initially for another purpose?
- How flexible are different designs and proposed methods of analysis for dealing with such situations?

Issues for Discussion

- What software is available for implementation of the various design and analysis proposals?
- How convenient (and safe) is it for end users who may have limited statistical background?

Issues for Discussion

- How can two-phase stratified sampling designs and associated methods of statistical analysis best be promoted within the scientific community so that there is greater appreciation of their importance and potential?