Double-Sampling Designs for Dropouts

Dave Glidden
Division of Biostatistics, UCSF
26 February 2010

http://www.epibiostat.ucsf.edu/dave/talks.html

Outline

- Mbarra's Dropouts
- Inverse Prob. Weights/Horvitz Thompson
- Double Sampling for Dropouts
- An IPW Estimator and Variance
- Future Directions

PEPFAR

- Massive scale-up of ARV treatment
- 1.3 Million initiated ARV from 2004-2008
- Outcome results essential
- Required large-scale follow-up

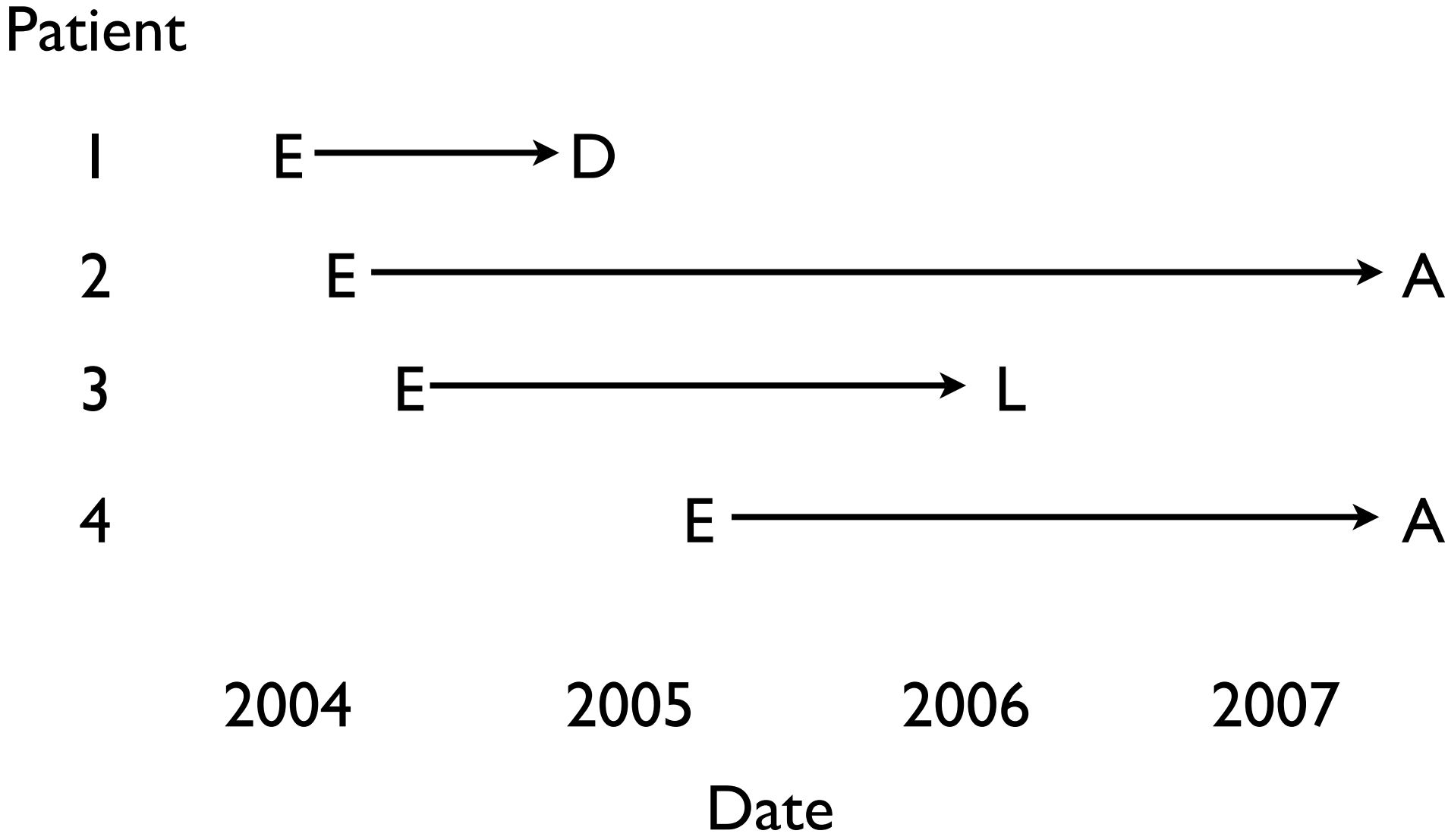
Mbarara ISS

- Cohort of 3,340 HIV+ infected individuals
- Started on ARV from 1/04 to 7/07
- Followed through 7/1/07
- 56 died
  2,530 alive as of 7/1/07
  715 lost to FU prior to 7/1/07

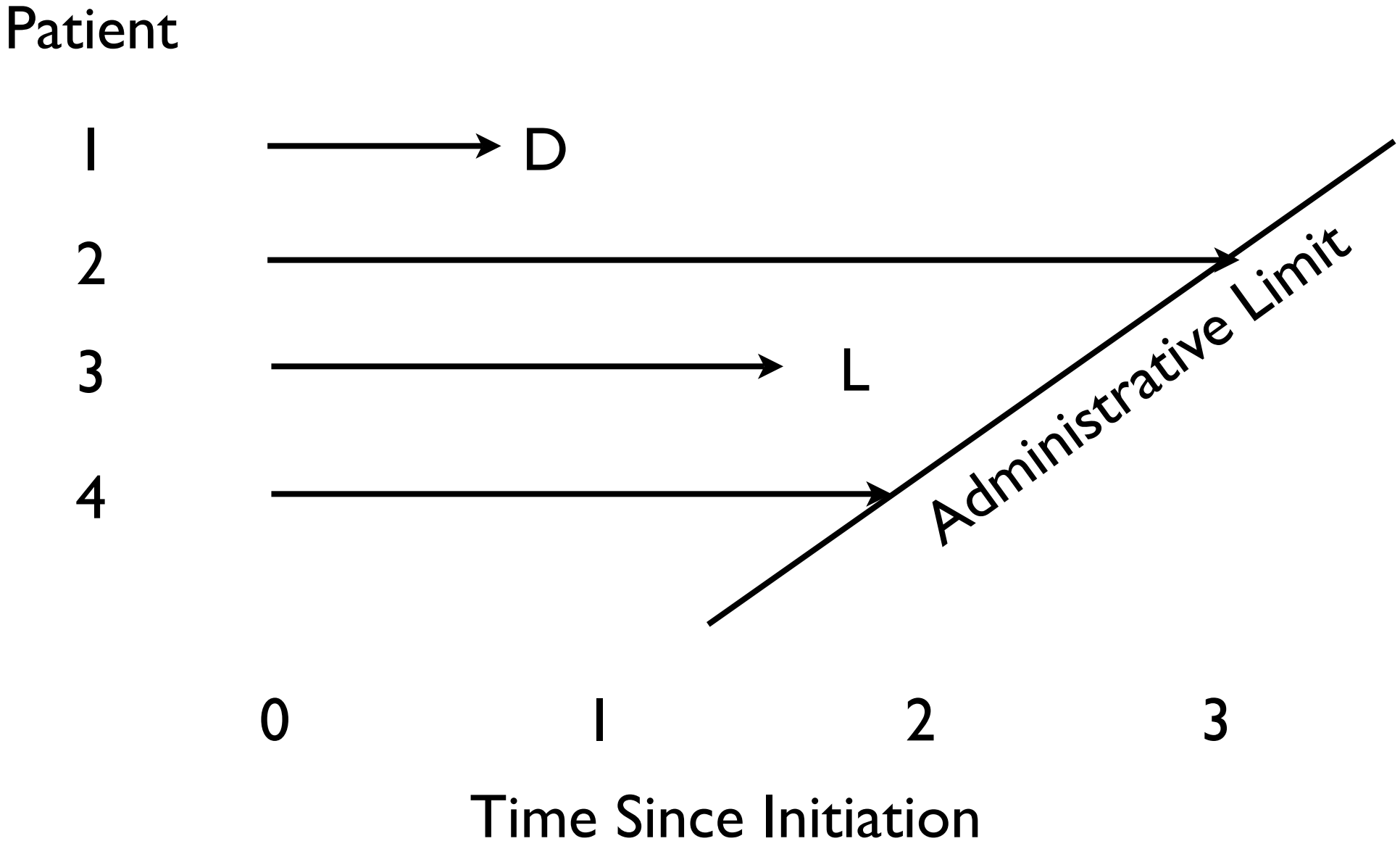
Multiple Events

- Three possible events
  - Death
  - Dropout
  - Administrative Censoring
- Competing risks

Calendar Time Scale



Follow Up Time Scale



Notation

- C_i: time to administrative censoring
  always known
- L_i: time to dropout
  censored by C_i and T_i
- T_i: time to death
  censored by C_i and L_i
- X_i: $\min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$
  data in absence of dropouts
- R^{obs}=0 dropout, R^{obs}=1 non-dropout

Administrative Censoring

- Patients can only dropout if
  $L_i < \min(T_i, C_i)$
- Patients can only die if
  $D_i < \min(L_i, C_i)$
- Some people may have dropout later....

Administrative Censoring

- (T,C) are independent
  very standard assumption
  violated if demographics change over time
- Can be relaxed to (T,C) independent
  given a series of covariates
- Conditional on R^{obs}, (T_i, C_i) *NOT* indep
  example of "collider" stratification
  creates dependent censoring

Dependent Dropout

- (T,L) are likely correlated
- Dropout suggest ARV discontinuation
- Hastens death
- Not easily handled
- What about observing after dropout
  how about sampling?

Sampling Plan

- 3,340 HIV+ initiated ART
- $R^{obs}=1$: $n_1=2,625$
  - 2,569 Alive and in FU as of 7/1/09
    - 56 Died in Follow-Up
- Robs=0: $n_0=715$, $\tilde{n}_0=79$
  - 95 sought, 79 vital status ascertained

Advantages of Sampling

- Very flexible
- Sampling prob can vary by individual
  using ancillary data
- Valid framework for dependent censoring
  in a way that is model-independent
  no need to specify cor(T,L)
  will get information on this

Horvitz Thompson

- Have finite population of size n
- Want to estimate $\tilde{\mu} = n^{-1} \sum_{i=1}^{n} x_i$
- ξ_i=1 indicated if ith person sampled
- Sample with probability $E(\xi_i=1)=\pi_i$
- HT estimator $\hat{\mu} = n^{-1} \sum_{i=1}^{n} \frac{\xi_i}{\pi_i} x_i$

Variance

$$\mathrm{var}\{\sqrt{n}(\tilde{\mu} - \hat{\mu})\}$$

$$= n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \right) \xi_i \xi_j x_i x_j$$

π_{ij} is the probability that i and j are selected

termed the second-order inclusion probability

Simple Sample

- Total population is n
- Sample with equal probability $\pi_i = \tilde{n}/n$
- Sample with replacement
  chose $\tilde{n}$ from n (putting balls back in jar)
- Sample with quota
  chose exactly $\tilde{n}$ different from n
- Sample without quota
  chose ith person with probability $\tilde{n}/n$

Same estimate, different variances!!

Second-Order Weights

under equal probability schemes

	π_{ii}	π_{ij}
w/ replace	$(\tilde{n}/n)^2$	$(\tilde{n}/n)^2$
quota	$\tilde{n}/n$	$\tilde{n}(\tilde{n}-1)/n(n-1)$
no quota	$\tilde{n}/n$	$(\tilde{n}/n)^2$

Variance for Finite Population

under "quota sampling"

$$\mathrm{var}\{\hat{\mu}\} = \frac{(\pi^{-1} - 1)\sigma^2}{n}$$

n: size of total popn

$$\mathrm{var}\{\hat{\mu}\} = \frac{(1 - \pi)\sigma^2}{\tilde{n}}$$

ñ: size of sampled popn

For the standard sample mean, π is effectively 0

$$\mathrm{var}\{\bar{x}\} = \frac{\sigma^2}{\tilde{n}}$$

Double Sampling

Neyman, 1938

- group of $i=1,...,n$ subjects
  sampled from a population
  (first level of sampling)
  some additional data avail. $(z_1,...,z_n)$
- second level of data
  richer data collection on subset
  sampling determined by $(z_1,...,z_n)$
  $pr(\xi_i=1|z_i)=\pi_i$

Example: Two-Stage Case Control Study

- Bladder cancer case/control study
- Exposure: metal working fluids (MWF)
- Phase I: Data collected including
  Q: *"Have you ever worked in metal industry?"*
- Phase II: Detailed work history/exposure collected
- Metal workers oversampled in Phase II

Other Examples

- Two-phase genotyping studies
- Nested case-control studies
- Case-cohort studies
- All valid, comparatively efft designs
- With significant cost savings

Back to Mbarara

Double Sample for Dropouts

- $R^{obs}=1$: observed death/admin censoring
  - data already completely observed
  - sampling fraction 1.00
- $R^{obs}=0$: observed dropouts
  - n_0 dropouts
  - sample $\tilde{n}_0$ for vital ascertainment
  - sampling fraction $\tilde{n}_0/n_0 = \pi$

Observed Data

- (X_i, Δ_i) if $\xi_i = 1$
- ignored otherwise
- $\xi = 1$ if $R^{obs} = 1$ or $R^{obs} = 0$ & sampled
- pr(sampled | $R^{obs} = 0$) = π
- Data is MAR conditional on R^{obs}

Frangakis and Rubin

- Double sampling estimate of survivor function
- Ignore data if $\xi=0$
  if dropout & not dble sampled, drop data
- Constructed survivor estimate
- Showed that (T,C) not independent conditional on the value of R^{obs}
- Estimate hazard and transform to survival

Frangakis Rubin Estimator

$$\hat{\Lambda}(t) := \sum_{g=0,1} \int_0^t \hat{w}_g(u) d\hat{\Lambda}_g^{\mathrm{crd}}(u),$$

where

$$\hat{w}_g(t) := \frac{\hat{\pi}_g(t)\hat{p}_g}{\sum_{g'=0,1} \hat{\pi}_{g'}(t)\hat{p}_{g'}}, \qquad \hat{\pi}_g(t) := Y_g(t)/n_g,$$

consistent with a Gaussian limiting distribution

Their representation

- Cumulative hazard is weighted sum
- Of crude hazards in 2 groups
- Weight varies with time
- Not an intuitive representation
  why does it work?

Double-Sampling

- $\Lambda(t)$ True cumulative hazard function
- $\tilde{\Lambda}(t)$ Nelson-Aalen estimator
  if complete data avail on cohort
- $\hat{\Lambda}(t)$ The FR estimator
  based on dble-sampled data

More Notation

- $N(t)=I(X \leq t, \Delta=1)$
- $Y(t) = I(X \geq t)$
- H_1 : set of people with $R^{obs}=1$
  has size $n_1, \tilde{n}_1$ *sampled* $(n_1 = \tilde{n}_1)$
  i in H_1 $\pi_i = 1$
- H_0 : set of people with $R^{obs}=0$
  has size $n_0, \tilde{n}_0$ *sampled*
  i in H_0: $\pi_i = \tilde{n}_0/n_0$

Complete Data on Cohort

$$\bar{N}_g(t) = n_g^{-1} \sum_{i \in \mathcal{H}_g} N_i(t)$$

$$\bar{Y}_g(t) = n_g^{-1} \sum_{i \in \mathcal{H}_g} Y_i(t)$$

$$\bar{N}(t) = \sum_g \frac{n_g}{n} \bar{N}_g(t)$$

$$\bar{Y}(t) = \sum_g \frac{n_g}{n} \bar{Y}_g(t)$$

Other Representation

$$\Lambda(t) = \int_0^t \frac{\mathcal{N}(du)}{\mathcal{Y}(u)}$$

$$\text{as} \qquad n \to \infty$$

$$\bar{N}(t) \to \mathcal{N}(t)$$

$$\bar{Y}(t) \to \mathcal{Y}(t)$$

$$\tilde{\Lambda}(t) = \int_0^t \frac{\bar{N}(du)}{\bar{Y}(u)}$$

Horvitz Thompson

$$\hat{N}_g(t) = \tilde{n}_g^{-1} \sum_{i \in \mathcal{H}_g} \frac{\xi_i}{\pi_i} N_i(t)$$

$$\hat{Y}_g(t) = \tilde{n}_g^{-1} \sum_{i \in \mathcal{H}_g} \frac{\xi_i}{\pi_i} Y_i(t)$$

$$\hat{N}(t) = \sum_g \frac{n_g}{n} \hat{N}_g(t)$$

$$\hat{Y}(t) = \sum_g \frac{n_g}{n} \hat{Y}_g(t)$$

Approximations

$$\hat{N}_1(t) = \bar{N}_1(t)$$

$$\hat{Y}_1(t) = \bar{Y}_1(t)$$

$$\hat{N}_0(t) \longrightarrow \bar{N}_0(t)$$

$$\hat{Y}_0(t) \longrightarrow \bar{Y}_0(t)$$

A Natural Estimator

$$\hat{\Lambda}(t) = \int_0^t \frac{\hat{N}(du)}{\hat{Y}(u)}$$

identical to FR estimator and to

$$\hat{\Lambda}(t) = \int_0^t \frac{\sum_i \frac{\xi_i}{\pi_i} N_i(du)}{\sum_i \frac{\xi_i}{\pi_i} Y_i(u)}$$

and has a IPW representation

But what about variance?

FR Variance

$$V_{\{p_1\}}(s,t) + \sum_{g=0,1} V_{\{\pi_g\}}(s,t) + V_{\{\Lambda_g\}}(s,t) + V_{\{\pi_g,\Lambda_g\}}(s,t) + V_{\{\pi_g,\Lambda_g\}}(t,s), \tag{A.2}$$

where

$$V_{\{p_1\}}(s,t) := p_1 p_0 \int_0^t d\Lambda^{\mathrm{crd}}_{\{p_1\}}(u) \int_0^s d\Lambda^{\mathrm{crd}}_{\{p_1\}}(u^*)$$

and, for $k_0 = 1/\{p_0 p^{(S)}\}$, $k_1 = 1/p_1$, and $g = 0, 1$,

$$V_{\{\pi_g\}}(s,t) := k_g \int_0^t \int_0^s \left[\pi_g\{\max(u,u^*)\} - \pi_g(u)\pi_g(u^*)\right] \times d\Lambda^{\mathrm{crd}}_{\{\pi_g\}}(u^*) d\Lambda^{\mathrm{crd}}_{\{\pi_g\}}(u),$$

$$V_{\{\Lambda_g\}}(s,t) := k_g \int_0^{\min(s,t)} \frac{\{w_g(u)\}^2}{\pi_g(u)} d\Lambda^{\mathrm{crd}}_g(u)$$

and

$$V_{\{\pi_g,\Lambda_g\}}(s,t) := -k_g \int_0^{\min(s,t)} \int_u^s \frac{\pi_g(u^*)}{\pi_g(u)} w_g(u) d\Lambda^{\mathrm{crd}}_{\{\pi_g\}}(u^*) d\Lambda^{\mathrm{crd}}_g(u).$$

Two-Part Variance

$$\sqrt{n}\{\hat{\Lambda}(t) - \Lambda(t)\} =$$

$$\sqrt{n}\{\hat{\Lambda}(t) - \tilde{\Lambda}(t)\} + \sqrt{n}\{\tilde{\Lambda}(t) - \Lambda(t)\}$$

last two terms are independent

$$\sigma^2(t) = \mathrm{var}\{\hat{\Lambda}(t) - \Lambda(t)\}$$

$$\sigma_1^2(t) = \mathrm{var}\{\hat{\Lambda}(t) - \tilde{\Lambda}(t)\}$$

$$\sigma_2^2(t) = \mathrm{var}\{\tilde{\Lambda}(t) - \Lambda(t)\}$$

$$\sigma^2(t) = \sigma_1^2(t) + \sigma_2^2(t)$$

Variance Decomposition

$\sigma^2(t)$: Total Variance

$\sigma_2^2(t)$: Variance if full cohort observed

Usual variance for Nelson-Aalen estimate

Easily estimated

$\sigma_1^2(t)$: Variance due to double sampling

HT type variance

Estimation more complicated

Nelson-Aalen Variance

$$\sigma_2^2(t) = \int_0^t \frac{\Lambda(du)}{\mathcal{Y}(u)}$$

$$\hat{\sigma}_2^2(t) = \int_0^t \frac{\hat{\Lambda}(du)}{\hat{Y}(u)}$$

$$= \frac{n_0}{n} \int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)} + \frac{n_1}{n} \int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}$$

$$= \int_0^t \frac{\hat{N}(du)}{\hat{Y}^2(u)}$$

Double-Sample Variance

looong HT-based variance arguments lead to

$$\hat{\sigma}_1^2(t) = (\pi^{-1} - 1)\frac{n_0}{n}\int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$$

$$= n^{-1}\sum_{i:R^{obs}=0}\int_0^t \frac{\xi_i}{\pi_i}\frac{(\pi_i^{-1} - 1)N_i(du)}{\tilde{Y}^2(u)}.$$

$$= n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\xi_i}{\pi_i}\frac{(\pi_i^{-1} - 1)N_i(du)}{\hat{Y}^2(u)}$$

The total variance

$$\hat{\sigma}^2(t) = \hat{\sigma}_1^2(t) + \hat{\sigma}_2^2(t)$$

$$= n^{-1} \sum_{i=1}^{n} \int_0^t \frac{\xi_i}{\pi_i} \frac{N_i(du)}{\hat{Y}^2(u)} +$$

$$n^{-1} \sum_{i=1}^{n} \int_0^t \frac{\xi_i}{\pi_i} \frac{(\pi_i^{-1} - 1) N_i(du)}{\hat{Y}^2(u)}$$

$$= n^{-1} \sum_{i=1}^{n} \int_0^t \frac{\xi_i}{\pi_i} \frac{\pi_i^{-1} N_i(du)}{\hat{Y}^2(u)}$$

Source of Variance

	$R^{obs}=0$	$R^{obs}=1$
$\hat{\sigma}_1^2(t)$	$(\pi^{-1}-1)\frac{n_0}{n}\int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$	0
$\hat{\sigma}_2^2(t)$	$\frac{n_0}{n}\int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$	$\frac{n_1}{n}\int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}$
$\hat{\sigma}^2(t)$	$\pi^{-1}\frac{n_0}{n}\int_0^t \frac{\hat{N}_0(du)}{\hat{Y}^2(u)}$	$\frac{n_1}{n}\int_0^t \frac{\hat{N}_1(du)}{\hat{Y}^2(u)}$

Some observations

- $R^{obs}=1$: no contribution to double-sample variance
- $R^{obs}=0$: ratio of NA variance to FR variance
  $= 1/\pi = n_0/\tilde{n}_0$
  just the ratio of sample size in Robs=0 between DS data and full data
- Intuitive look at the variance

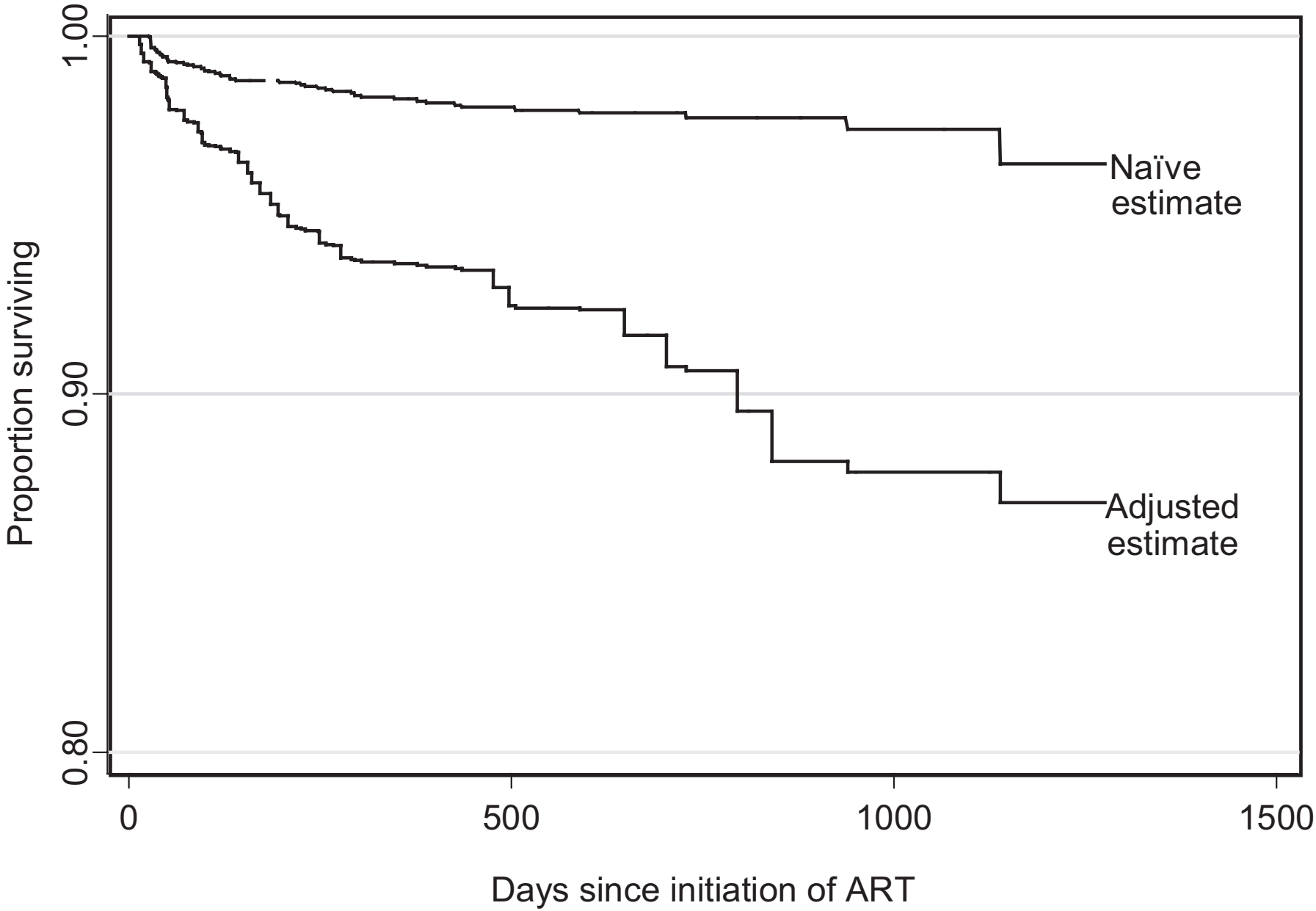
Variance Estimate

- Easily computed
- Demystifies the form
- Facilitates sample size calculations
  look at effect of various sample fractions
- Performs great in simulations

Data Example

- Cohort of 3,340 HIV+ infected individuals
- $R^{obs}=1$: $n_1=2,625$ (56 died)
- Robs=0: $n_0=715$, $\tilde{n}_0=79$ (26 died)
- $\pi = 1/9.18 = 0.109$
- Rate of death is about 16 times higher
  in dropouts compared to non-dropouts

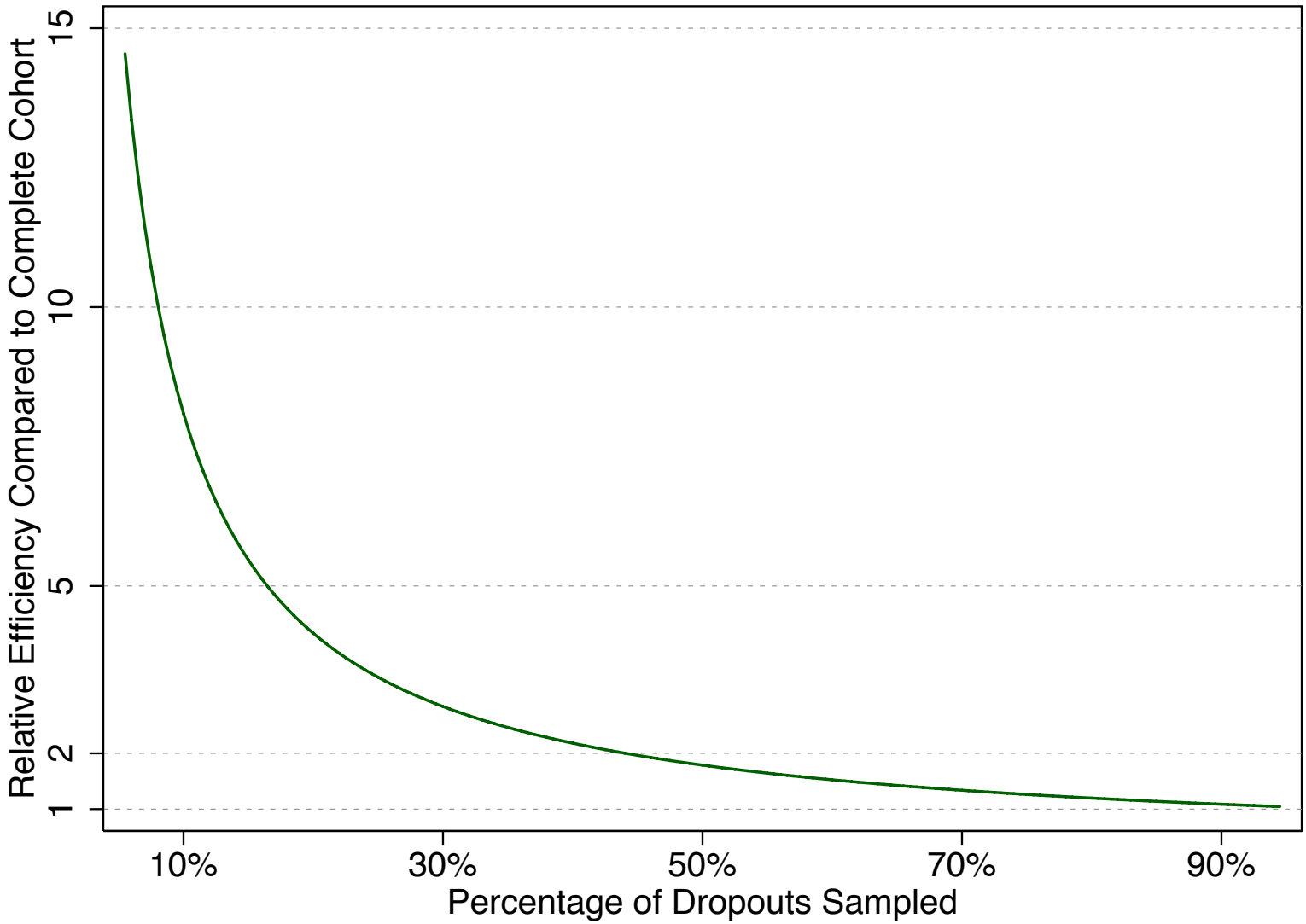
FR and Naive Survival



Relative Efficiency

- Trade-off between sampling fractions
- What is efficiency of sampling ρ dropouts compared to all dropouts
- Can be consistently estimated
- Based on Mbarara data

Relative Efficiency



Future Directions

- Apply insights from survey statistics
  - formulae and approximations
  - post-stratification, calibration
    auxiliary variables => more efficiency
- Look at using non-sample dropout data

Acknowledgement

- Jeff Martin
- Elvin Geng
- Constantin Yiannoutsos
- Mbarara ISS clinic
- East Africa Region of IeDEA
- Chuck McCulloch