

Choosing good subsamples

Doing more with fewer

Thomas Lumley, University of Auckland, New Zealand

Setting

- You have a data set: cohort, case-control, EHR database,...
- You want to measure some more variables (or the same ones, more accurately)
- You want to fit a regression model and estimate associations/effects

Notation

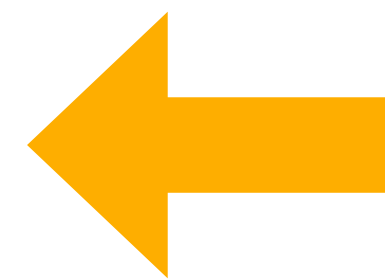
- Y: outcome (typically in everyone)
- X: predictors you want to measure
- Z: predictors in the model you want to fit available for everyone
- A: other available variables **not** in the model you want to fit
- R: indicator of being sampled

Executive Summary

- You can choose people based on any information you have
- You need to account for the sampling in the analysis
- Weighted complete-case analysis does pretty well in a wide variety of settings
- We have software

Measuring additional variables

- Stored blood/tissue samples
- Free-text questionnaire
- Summaries of clinical notes
- Panel-adjudicated outcomes
- Actually talking to people



**Additional
non-response**

Who to measure?

- Everyone!
- Simple random sample
- Random sample stratified by eg site
- Case-control sample for one outcome
- Case-cohort samples for multiple outcomes
- Who all is most informative!

Information heuristics

Simple linear regression:

- maximise variation in exposure

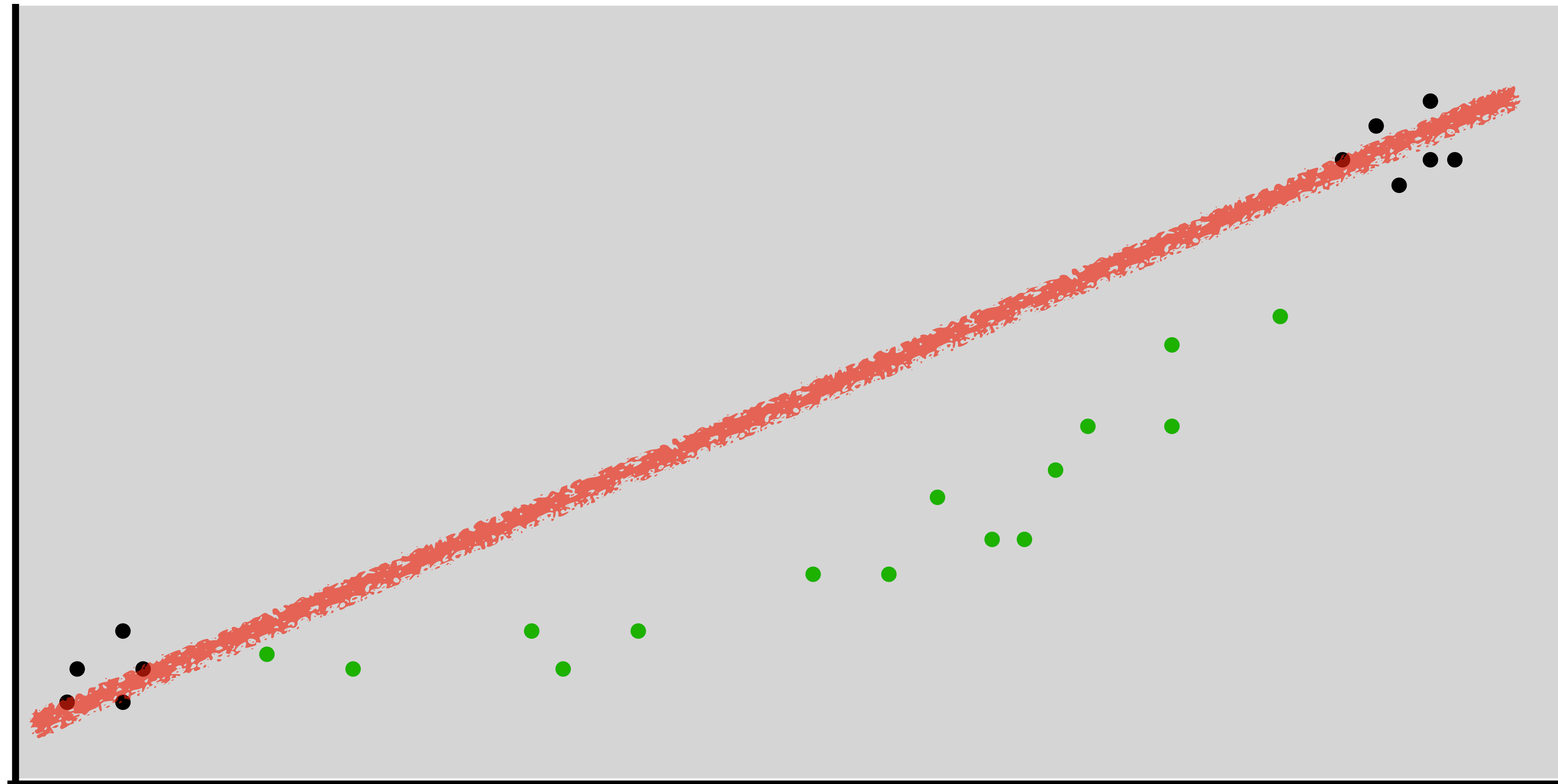
Multiple linear regression

- maximise variation in exposure conditional on other predictors

Logistic regression

- also try to get outcome prevalence near 0.5

Model-based vs design-based



Our goal

You pick a model $Y \sim X + Z$ that **you would fit to the full cohort**

We want to estimate the full-cohort best-fitting parameters, with

- little to no bias
- as small variance as feasible
- even if the model isn't perfectly accurate

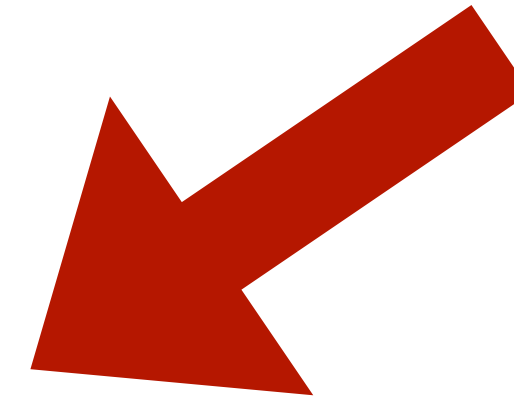
(Other goals lead to other sampling designs!)

Estimators

Weighted regression

- with sampling weights
- or estimated weights (Robins et al)
- or raking/calibration of weights (Deville & Särndal, et al)
- or AIPW (Robins et al)

All the same
thing



Weighted estimation

Complete-cohort estimator would solve

$$\sum_{i=1}^N U(X_i, Y_i, Z_i, A_i; \beta) = 0$$

Weighted estimator solves

$$\sum_{i=1}^N \frac{R_i}{\pi_i} U(X_i, Y_i, Z_i, A_i; \beta) = 0$$

where π_i is the probability of sampling person i given all the complete variables

Use the R survey package (or Stata)

Example simulations

Data from National Wilms' Tumor Group trials (via Norm Breslow)

Binary Y: relapse

Binary X: histology favorable or not (Central lab)

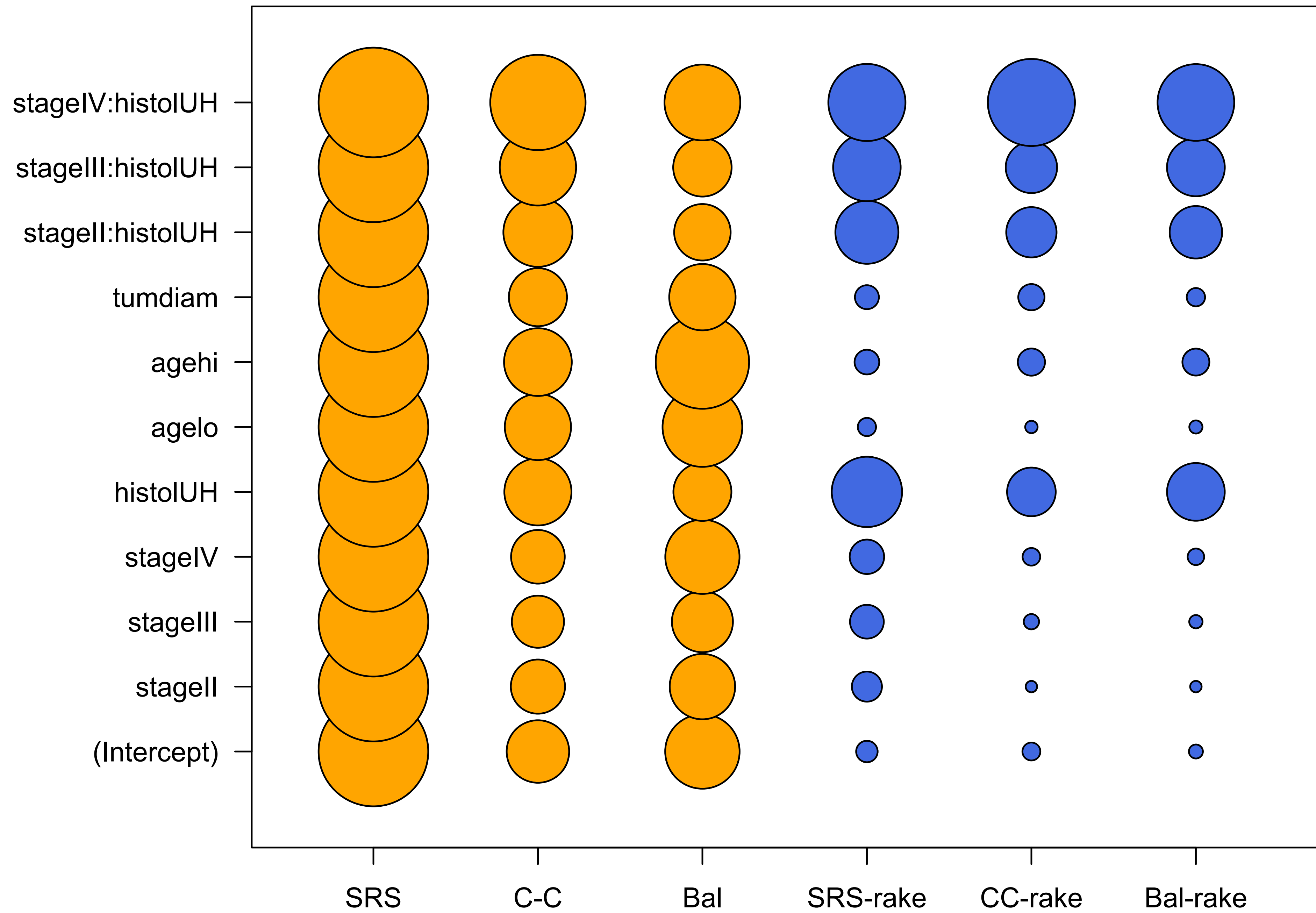
Binary A: histology favorable or not (Local lab)

Covariates Z: Stage, age at diagnosis, tumor diameter

	FH	UH
Relapse	450	153
Not	374 / 3083	229

Model: relapse~stage*hist+agelo+agehi+tumdiam

Subsample: 1200 out of 3915



Optimal design

Secret weapon: influence functions

Influence function:

$$\hat{\beta} - \beta_0 = \sum_{i=1}^N h(X_i, Y_i, Z_i, A_i; \beta_0) + O_p(\text{small})$$

eg linear regression

$$h_i(X, Y; \beta) = E[X^T X]^{-1} x_i (y_i - x_i \beta)$$

Survey sampling theory is **good** at estimating population sums

Secret weapon: Neyman allocation

Given a division of the cohort into K strata of size N_k , the optimal sample size n_k for each stratum to estimate a population total satisfies

$$n_k \propto N_k \sigma_k$$

where σ_k is the standard deviation of the variable. Also, ideal strata have n_k approximately equal

Do this with estimates of the influence functions! (might need some imputation)

A-optimality: use $\sigma_k = \sqrt{\sum_p \sigma_{kp}^2}$

NWTS again

Data from National Wilms' Tumor Group trials (via Norm Breslow)

Binary Y: relapse

Binary X: histology favorable or not (Central lab)

Binary A: histology favorable or not (Local lab)

Covariates Z: Stage, age at diagnosis, tumor diameter

	FH	UH
Relapse	224 / 450	153
Not	600 / 3083	229

Use imputed X (mostly from A) to approximate inf funs, optimal design for **histology** coef

	Bal	rake	Opt	rake
(Intercept)	1	0.188	1.151	0.220
stageII	1	0.175	1.335	0.328
stageIII	1	0.218	1.416	0.380
stageIV	1	0.223	1.144	0.305
histolUH	1	0.995	0.994	0.894
ageI	1	0.165	1.107	0.176
agehi	1	0.289	1.007	0.282
tumdiam	1	0.276	1.188	0.268
stageII:histolUH	1	0.929	1.146	1.055
stageIII:histolUH	1	0.990	1.011	0.895
stageIV:histolUH	1	1.011	0.941	0.895

```
ccs2.twophase.if <- twophase(id=list(~1, ~1),  
  strata=list(NULL, ~interaction(rel3, instit))  
  subset=~in.ccs2,  
  data=nwts.if,  
  method="simple")
```

```
model3 <- svyglm(rel3~stage*histol+agelo+agehi+tumdiam,  
  family=quasibinomial,  
  design=ccs2.twophase.if)
```

```
cal.ccs2 <- calibrate(ccs2.twophase.if,  
  formula=if.formula,  
  calfun="raking")
```

```
model3.cal <- svyglm(rel3~stage*histol+agelo+agehi+tumdiam,  
  family=quasibinomial,  
  design=cal.ccs2)
```

Multiwave design

Multiple waves

Optimal design depends on unknown parameters

- Take a small sample, estimate parameters, optimal design
- Take another small sample, re-estimate

Two waves is definitely helpful, three sometimes is.

Multiple waves also help if you get something wrong!

R package `optimal` helps with multiwave design

Received: 1 April 2022






Accepted: 16 June 2022

DOI: 10.1111/biom.13713

BIOMETRIC PRACTICE

Biometrics WILEY
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

Multiwave validation sampling for error-prone electronic health records

Bryan E. Shepherd¹  | **Kyunghee Han**²  | **Tong Chen**³ | **Aihua Bian**¹ |
Shannon Pugh⁴ | **Stephany N. Duda**⁵ | **Thomas Lumley**³  |
William J. Heerman⁶  | **Pamela A. Shaw**⁷ 

Biometrics best paper award, 2023

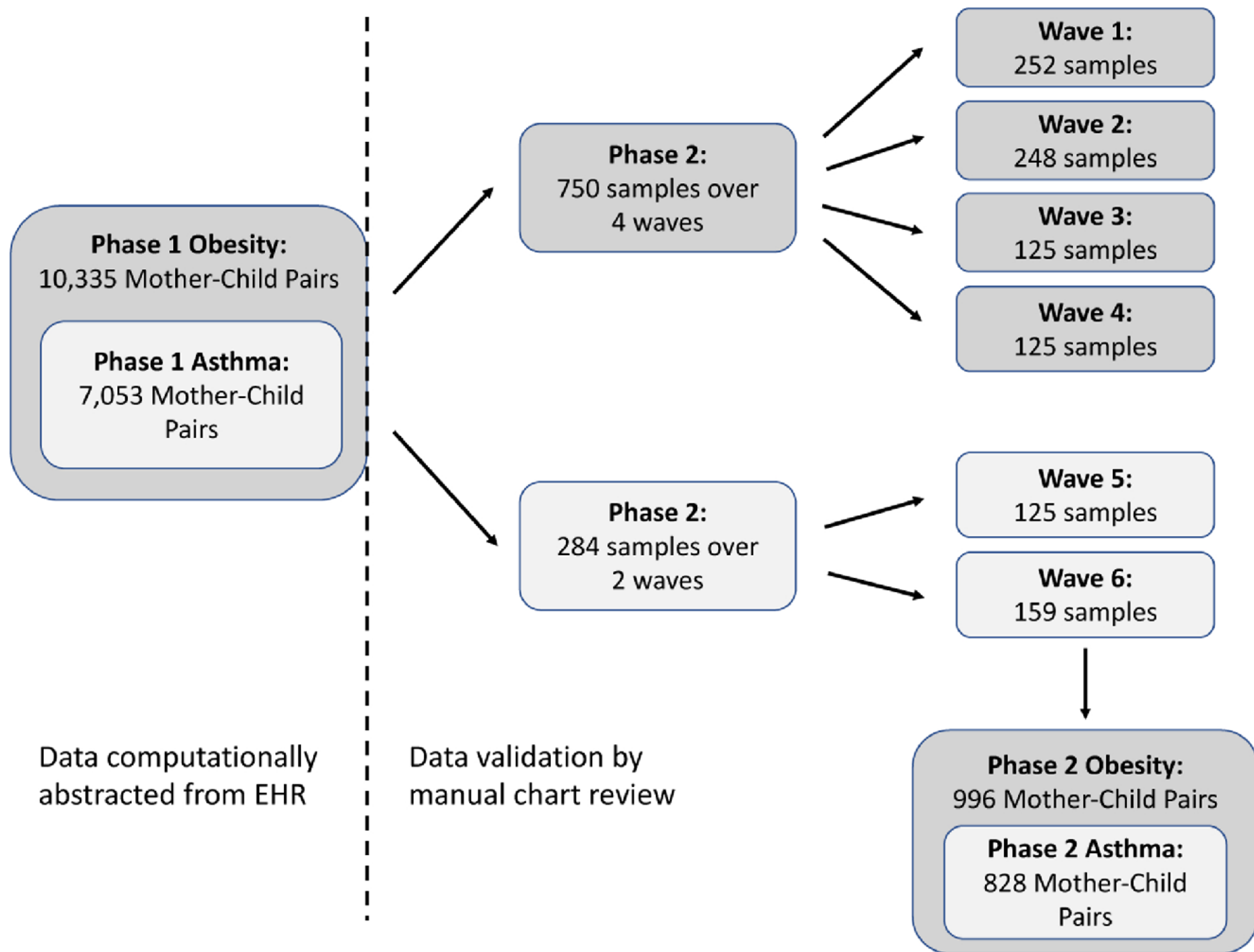
...estimate the association between maternal weight gain during pregnancy and the risks of her child developing obesity or asthma.

The optimal validation sampling design depends on the unknown efficient influence functions of regression coefficients of interest.

In the first wave of our multiwave validation design, we estimate the influence function using the unvalidated (phase 1) data to determine our validation sample;

then in subsequent waves, we re-estimate the influence function using validated (phase 2) data and update our sampling.

For efficiency, estimation combines obesity and asthma sampling frames while calibrating sampling weights using generalized raking. We validated 996 of 10,335 mother-child EHR dyads in six sampling waves.



Take-home message

- You can choose people based on any information you have
- You need to account for the sampling in the analysis
- Weighted complete-case analysis does pretty well in a wide variety of settings
- We have software (survey, optima11)

Any questions?

