Sentinel

Multi-Wave Validation Sampling to Improve Estimates Derived from Electronic Health Record Data

Sentinel Innovation and Methods Seminar Series

Bryan Shepherd

bryan.shepherd@vanderbilt.edu

2023/07/11

Paper and Acknowledgments

Shepherd BE, Han K, Chen T, Bian A, Pugh SK, Duda SN, Lumley T, Heerman WJ, Shaw PA. Multi-wave validation sampling for error-prone electronic health records. *Biometrics* (in press).

Others

- Gustavo Amorim, Ran Tao, Sarah Lotspeich, Mark Giganti

Funders

- PCORI (R-1609-36207)
- NIH (R01AI131771)

Background

- Routinely collected clinical data, including electronic health records (EHRs), are increasingly used as a data source for medical studies

- These data are often prone to errors

Data Quality

BBC | Sign in | Home | News | Sport | Reel | Worklife | Travel

NEWS

Home | War in Ukraine | Coronavirus | Climate | Video | World | US & Canada | UK | Business | Tech | Science

England | Regions | Liverpool

Covid: Man offered vaccine after error lists him as 6.2cm tall

18 February 2021

Coronavirus pandemic

Liam Thorp was wrongly classed as morbidly obese according to his height and weight

A man in his 30s with no underlying health conditions was offered a Covid vaccine after an NHS error mistakenly listed him as just 6.2cm in height.

Liam Thorp was told he qualified for the jab because his measurements gave him a body mass index of 28,000.

He told BBC Radio 5 Live: "I've put on a few pounds in lockdown but I was surprised to have made it to clinically, morbidly-obese.

"It really made me rethink what I was going to do for pancake night."

Example: Floyd et al. (2012), *JAMA* 307: 1580–1582

Incident rate ratios (IRR) for statin-related rhabdomyolysis, a rare adverse drug reaction

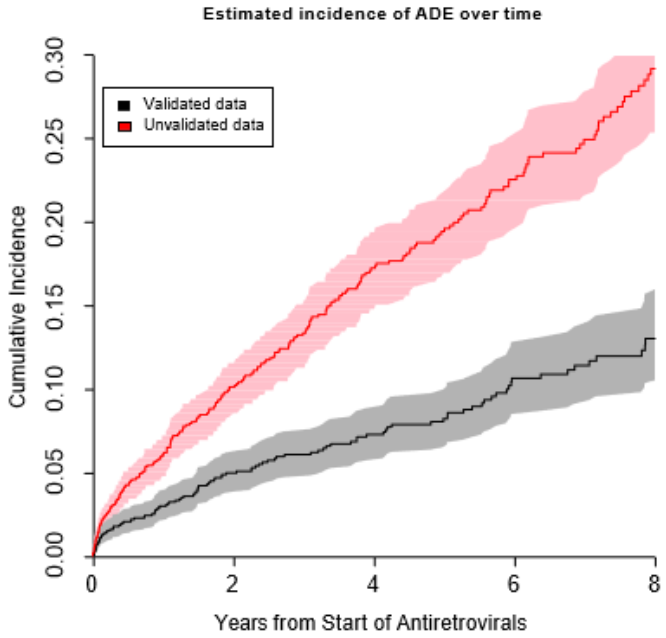
	simvastatin vs. other statins (95% CI)	high vs. low doses of simvastatin (95% CI)
Unvalidated	1.03 (0.80, 1.34)	1.77 (1.05, 2.88)
Validated	2.6 (1.03, 7.84)	12.2 (3.6, 52.3)

Vanderbilt Comprehensive Care Clinic

- 4217 HIV-positive adults who established care from 1998-2011
- Extensive chart reviews are performed to validate key variables for all patients
- Pre- and post-validation datasets available
- Incidence of ADE after starting ART and association with CD4 at ART initiation

Data	Estimated Incidence at 5 years (95% CI)	Estimated Hazard Ratio for 100 cell CD4 increase (95% CI)
Unvalidated	0.196 (0.171, 0.221)	0.80 (0.74, 0.86)
Validated	0.083 (0.066, 0.100)	0.63 (0.55, 0.72)

Giganti et al. (2020) *Ann Appl Stat* 14: 1045–1061.



Validation Sampling

Validate subsamples of records

- Validation of all records is resource-intensive and often unrealistic
- An alternative is to validate data on a random subset of records
- Goal is to obtain estimates that are efficient and are close to estimates had the entire dataset been validated

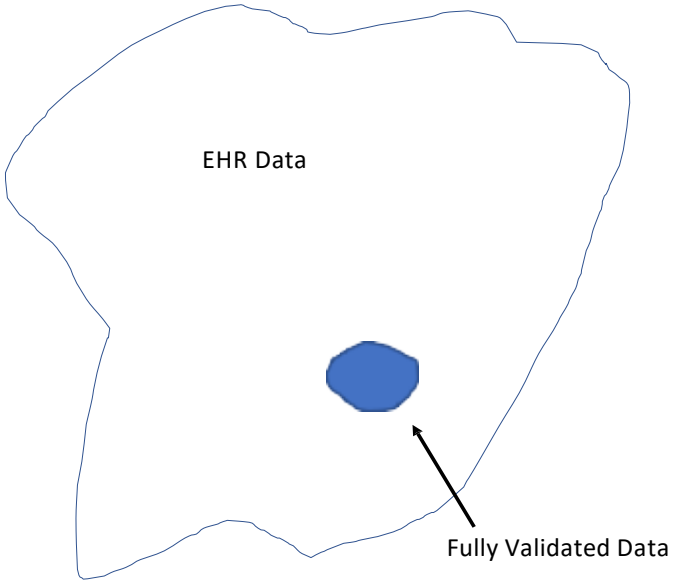
Research Agenda

- Estimation: How to best combine validated and unvalidated data?
- Design: How to best select which records to validate?
- Applying new methods and designs in practice

Methods for Incorporating Validation Data

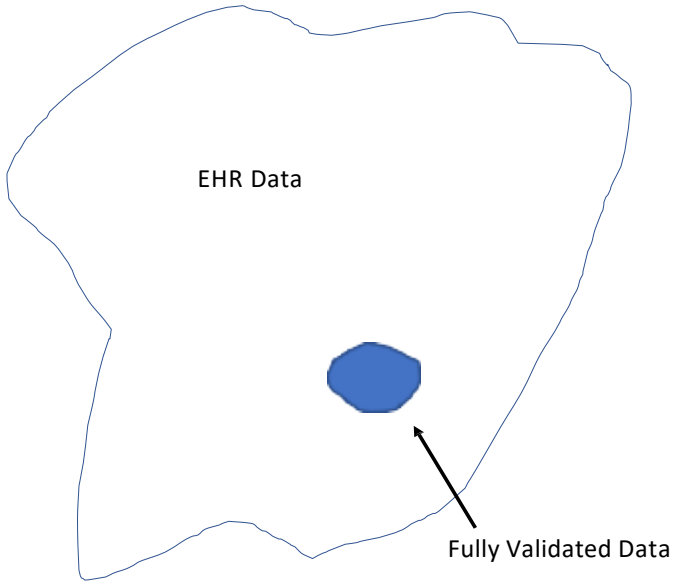
Errors across multiple variables

- Traditional measurement error methods
  - Moment-based estimation: Shepherd & Yu (2011) *Biometrics* 67: 1100-1110.
  - Regression calibration: Shaw et al. (2021) *Stat Med* 40: 271–286.
  - SIMEX: Oh et al. (2018) *Stat Med* 37: 1276–1289.
- Full likelihood approaches
  - Tao et al. (2020) *Stat Med* 40: 725–738.
  - Lotspeich et al. (2022) *Biometrics* 78: 1674–1685.
- Multiple imputation
  - Giganti et al. (2020) *Ann Appl Stat* 14: 1045–1061+
- Generalized raking
  - Oh et al. (2021) *Stat Med* 40: 631–649



Generalized Raking Estimators

- D = childhood obesity (validated)
- D^* = error-prone childhood obesity (EHR)
- IPW estimate of $Pr(D = 1)$
  - unbiased for population estimate, but high variance
- IPW estimate of $Pr(D^* = 1)$
  - unbiased for population estimate, but not exact due to sampling error
  - but $Pr(D^* = 1)$ already known because D^* is available for everyone in phase 1
- Tweak our IP weights so that IPW estimate = known value in phase 1
  - Keep weights as close to possible as original IPW but with this new constraint
- Now apply those new weights to obtain a raked estimator of $Pr(D = 1)$
- If D^* is correlated with D then raked estimator more efficient than IPW estimator

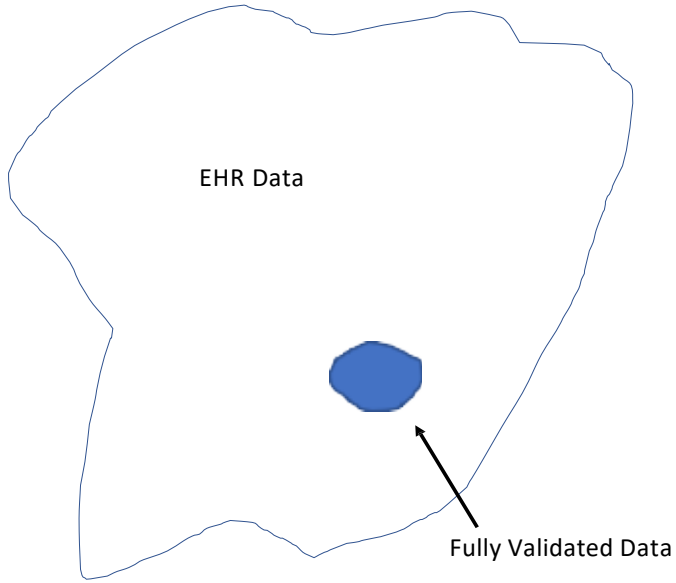


Generalized Raking Estimators

- Generalized raking estimator is more efficient than IPW estimator
  - efficiency improves with auxiliary variable closer to truth
- Idea extends to more complicated estimators
  - e.g., regression coefficients
  - auxiliary variable is the influence function
- Generalized raking makes same assumptions as IPW estimator and fewer than MI or likelihood-based methods
- Well-known in survey sampling literature, but less known in biostatistics
- Also known as generalized regression or calibration
  - Sarndal et al. (2003) *Model Assisted Survey Sampling*
  - Lumley et al. (2011) *Int Stat Rev* 79: 200-220.
- Generalized Raking Estimators $\subset$ Augmented Inverse Probability Weighted (AIPW) Estimators

Designs for Sampling Validation Records

- Simple random sampling
- Case-control sampling
  - Breslow, Chatterjee (2002) *JRSS-C* 48: 457–468.
- Optimal sampling
  - Tao et al. (2020) *JASA* 115: 1946–1959.
  - Amorim et al. (2021) *JRSS-A* 184: 1368–1389
- Multi-wave sampling
  - McIsaac, Cook (2015) *Stat Med* 34: 2899–2912.
  - Han et al. (2021) *Stat Methods Med Res* 30: 857–874.
  - Lotspeich et al. (in press) *Can J Stat*.



Our experience designing and carrying out a multi-wave validation study

Mother-Child Obesity Study

What is the association between maternal weight gain during pregnancy and the time to childhood obesity?

Secondary:

What is the association between maternal weight gain during pregnancy and a child's risk of developing asthma?

Study Variables

Variables $(Y, D, X, \boldsymbol{Z})$

- Y = time from birth to childhood obesity or censoring
- D = indicator of childhood obesity
- X = maternal weight change during pregnancy
- $\boldsymbol{Z}$ = other covariates

Unvalidated Variables $(Y^*, D^*, X^*, \boldsymbol{Z}^*)$

$(Y^*, D^*, X^*, \boldsymbol{Z}^*)$ are available for all subjects in the EHR,
$(Y, D, X, \boldsymbol{Z})$ will only be available for those records that are validated

Model of Interest

Cox model:

$$h(t|X,\boldsymbol{Z}) = h_0(t)exp(\beta X + \beta_Z \boldsymbol{Z}),$$

where $h(t|X, \boldsymbol{Z})$ is hazard of childhood obesity at age t conditional on X and $\boldsymbol{Z}$, $h_0(t)$ is unspecified baseline hazard, β is parameter of interest.

Phase 1 Data (Y^*, D^*, X^*, Z^*)

Inclusion criteria:

- Mothers in VUMC EHR who gave birth between Dec 2005 and Aug 2019
- Linked child also in VUMC EHR
- Mother had at least one height measurement and one weight measurement $\in (-1.75, 0)$ years
- Child had at least one pair of height/weight measurements > 2 years

$N = 10,335$ mother-child pairs.

Data extracted by programers from EHR including demographics, ICD-9/ICD-10 diagnoses, labs, encounters, and insurance data.

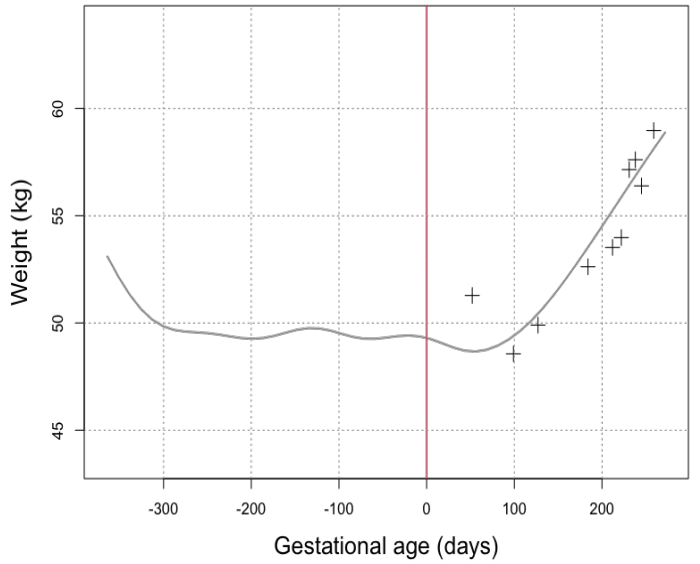
Published Phecodes used to determine asthma, diabetes, and depression. Childhood obesity defined as BMI ≥ 95th percentile based on age and sex using US CDC growth curves.

Deriving Maternal Weight Gain

X = average maternal weight change per week during pregnancy
=(weight just before delivery – weight at conception)/ pregnancy length

Challenge: We don't know most of these variables

- Estimate using functional principal components analysis (FPCA)
- FPCA borrows information across mothers while fitting mother-specific weight trajectory
- Based on procedure proposed by Yao et al. (2005) *JASA* 100: 577–590.
- Initially assume all pregnancies were 273 days



Validation Procedures to get Phase 2 Data (*Y, D, X, Z*)

Thorough review of complete EHR by research nurse.

n = 996 linked mother-child records.

Phase 1 data extracted computationally. Phase 2 involved looking at free text fields, other data not easily extracted.

For example, estimated gestational age was not in phase 1 data but extracted in phase 2.

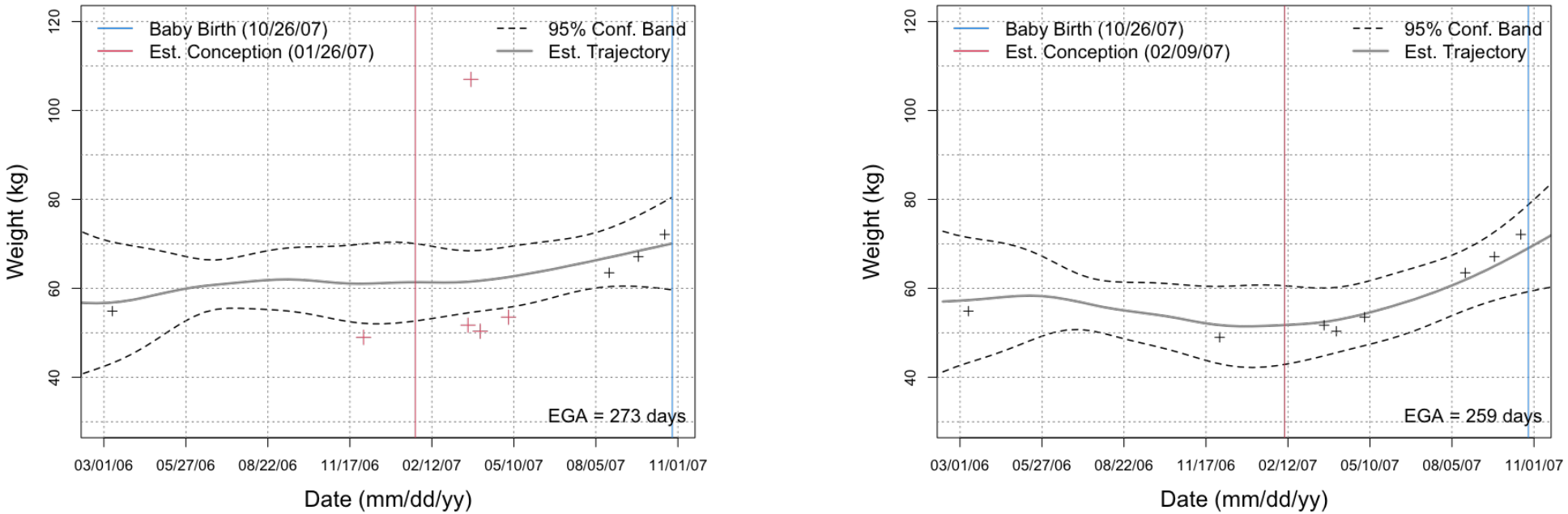
REDCap forms, Excel spreadsheets.

Pilot validation of 10 records.

Too many weights / heights to validate all. Chose a subset to validate and then flagged outliers for validation.

Screenshot of REDCap form

Maternal Race	White
Maternal Ethnicity	Non-Hispanic
Mother's height (cm) *Most representative value*	162.56
Marital Status *At time of child's birth*	Single
Any Tobacco Use (ever)	Yes
Tobacco Use during Pregnancy	No
Alcohol Use during Pregnancy	No
Type 1 or Type 2 Diabetes (ever) *Not gestational, prior to study birth*	No
Gestational Diabetes	No
Depression (ever)	No
Asthma (Mother) *Is there evidence that the mother has asthma?*	No
Pregnancy / Delivery Data	
Insurance category *At time of birth*	Medicaid



The estimated weight trajectory and 95%-confidence band derived using FPCA for one of the mothers based on phase 1 (left) and phase 2 (right) data; dates have been shifted for de-identification. Red crosses in the left panel were identified as potential outliers and were manually validated. After validation, we updated the weight trajectory (right panel); the outlier weight > 100 kg was found to be erroneous and removed.

Selecting which Records to Validate

Stratified random sampling

With fixed strata and a fixed number to validated ($n \approx$ 1000), the optimal way to validate across strata for an IPW estimator is via *Neyman allocation*:

$$n_s = n\frac{N_s\sigma_s}{\sum_s N_s\sigma_s},$$

where N_s is population size of stratum s

σ_s is the standard deviation in stratum s. (Neyman (1934) *J R Stat Soc* 97: 558–625.)

For optimal design for a regression coefficient (e.g., log hazard ratio), σ_s is standard deviation of the influence function for β.

Multi-Wave Sampling

Choice of strata matters

- Choose strata to minimize σ_s within, maximize between
- Generally, more strata are better
- Optimality achieved with
  - $n_1 = n_2 = \cdots = n_S$

σ_s is generally not known

- Approximate σ_s with estimate from EHR data, σ_s^*, for first sampling wave.
- Estimate σ_s from first wave of validated data, and then recalculate optimal allocation.

Defining Strata

Based on (Y^*, D^*, X^*), the variables that will have the largest influence on β.

Over-sample records with the largest influence on β.

- Those experiencing childhood obesity early ($D^* = 1, Y^*$ small)
- Those with lots or little weight gain (X^* small or large)
- Standard deviation of influence function will likely be large in these strata

Fit naive Cox model to error-prone data:

$$h(t \mid X^*, \boldsymbol{Z}^*) = h_0(t) exp(\beta^* X^* + \beta_Z^* \boldsymbol{Z}^*),$$

Compute the influence function for β^* for each observation.

Play around with strata boundaries such that

$$n_{(1),s} = n_{(1)} \frac{N_s \sigma_s^*}{\sum_s N_s \sigma_s^*},$$

such that $n_{(1),s} \approx n_{(2),s} \cdots \approx n_{(1),S}$.

Wave 1 Strata and Numbers

Original Strata	Obesity	Follow-up Time (yrs)	Maternal Gestational Weight Gain (kg)	N_s	$n_{(1),s}$
A	0	(2, 5]	≤ 5.14	190	7
B	0	(2, 5]	(5.14, 20.5]	3786	8
C	0	(2, 5]	> 20.5	177	8
D	0	(5, 6]	≤ 5.14	208	14
E	0	(5, 6]	(5.14, 20.5]	3904	16
F	0	(5, 6]	> 20.5	225	17
G	1	(2, 2.5]	≤ 5.14	49	17
H	1	(2, 2.5]	(5.14, 20.5]	547	20
I	1	(2, 2.5]	> 20.5	33	17
J	1	(2.5, 3]	≤ 5.14	13	12
K	1	(2.5, 3]	(5.14, 20.5]	258	12
L	1	(2.5, 3]	> 20.5	19	12
M	1	(3, 4]	≤ 5.14	21	10
N	1	(3, 4]	(5.14, 20.5]	378	13
O	1	(3, 4]	> 20.5	28	13
P	1	(4, 5]	≤ 5.14	22	9
Q	1	(4, 5]	(5.14, 20.5]	261	10
R	1	(4, 5]	> 20.5	24	11
S	1	(5, 6]	≤ 5.14	14	8
T	1	(5, 6]	(5.14, 20.5]	167	8
U	1	(5, 6]	> 20.5	11	10
Total				10335	252

After Completing Wave 1 Validation

Fit a new Cox regression model incorporating validated data

- Weighted Cox model of validated data

Standard deviation of influence function, $\hat{\sigma}_{s,1}$ re-estimated.

Neyman allocation to select wave 2:

$$n_{(2),s} = \left(\sum_{j=1}^{\Sigma 2} n_{(j)} \right) \frac{N_s \hat{\sigma}_{s,1}}{\sum_s N_s \hat{\sigma}_{s,1}} - n_{(1),s},$$

If $n_{(2),s} < 0$, then that stratum is closed and Neyman allocation is recalculated for the total number to be validated in the remaining strata.

Wave 2 Sampling

Original Strata	Obesity	Follow-up Time (yrs)	Maternal Gestational Weight Gain (kg)	N_s	$n_{(1),s}$
A	0	(2, 5]	≤ 5.14	190	7
B	0	(2, 5]	(5.14, 20.5]	3786	8
C	0	(2, 5]	> 20.5	177	8
D	0	(5, 6]	≤ 5.14	208	14
E	0	(5, 6]	(5.14, 20.5]	3904	16
		...			
Total				10335	252

Neyman allocation for $n_{(1)} + n_{(2)} = 500$ suggested fairly different sampling scheme

- Neyman allocation for stratum A was 6.
- Neyman allocation for stratum E was 105.
- Some (9) strata were closed.
- Some (4) strata were split.

Wave 2 Sampling

Original Strata	Obesity	Follow-up Time (yrs)	Maternal Gestational Weight Gain (kg)	N_s	$n_{(1),s}$
A	0	(2, 5]	≤ 5.14	190	7
B	0	(2, 5]	(5.14, 20.5]	3786	8
C	0	(2, 5]	> 20.5	177	8
D	0	(5, 6]	≤ 5.14	208	14
E	0	(5, 6]	(5.14, 20.5]	3904	16
		...			
Total				10335	252

became

Wave 2 Strata	Obesity	Follow-up Time (yrs)	Maternal Gestational Weight Gain (kg)	N_s	$n_{(1),s}$	$n_{(2),s}$	$n_{(1),s} + n_{(2),s}$
A	0	(2, 5]	≤ 5.14	190	7	0	7
B	0	(2, 5]	(5.14, 20.5]	3786	8	21	29
C	0	(2, 5]	> 20.5	177	8	2	10
D	0	(5, 6]	≤ 5.14	208	14	18	32
E1	0	(5, 6]	(5.14, 8.6]	429	3	22	25
E2	0	(5, 6]	(8.6, 12]	1478	5	15	20
E3	0	(5, 6]	(12, 20.5]	1997	8	18	26
		...					
Total				10335	252	248	500

Waves 3 and 4

Process repeated

- $n_{(3)} = 125$ across 30 strata
- $n_{(4)} = 125$ across 33 strata

Final strata

Original Strata	Final Strata	Obesity	Follow-up Time (yrs)	Maternal Gestational Weight Gain (kg)	N_s	n_s
A	1	0	(2, 5]	≤ 5.14	190	7
B	2	0	(2, 5]	(5.14, 12]	1904	24
	3	0	(2, 5]	(12, 16]	1356	34
	4	0	(2, 5]	(16, 20.5]	526	37
C	5	0	(2, 5]	> 20.5	177	13
D	6	0	(5, 6]	≤ 5.14	208	33
E	7	0	(5, 6]	(5.14, 8.6]	429	25
	8	0	(5, 6]	(8.6, 12]	1478	39
	9	0	(5, 6]	(12, 14]	846	44
	10	0	(5, 6]	(14, 16]	563	40
	11	0	(5, 6]	(16, 20.5]	588	35
			...			
Total					10335	750

Sampling for Asthma Endpoint

A total of 250 mother-child pairs were targeted for sampling for the asthma endpoint

Table: Multi-wave Sampling Design for Childhood Asthma Endpoint

Original Strata	Final Strata	Asthma	Maternal Gestational Weight Gain (kg)	N_s	$n_{(1),s}$	$n_{(2),s}$	n_s
A	1	0	< 5	306	31	27	31
	2	0	[5, 10)	1251		4	31
B	3	0	[10, 12)	1520	16	16	20
	4	0	[12, 15)	1681		13	25
C	5	0	[15, 19.5)	1105	24	21	34
	6	0	≥ 19.5	459		23	34
D	7	1	< 8	115	23	11	23
	8	1	[8, 12)	278		13	24
E	9	1	[12, 17]	240	31	4	27
	10	1	≥ 17	98		27	35
Total							

N_s is the population size in stratum s, $n_{(1),s}$ is the number sampled from the stratum in wave 1, $n_{(2),s}$ is the number sampled from the stratum in wave 2, and n_s is the total number sampled from stratum s over both waves of the phase 2 validation sampling.

Audit Results

Variable	Phase 1 *N* = 10, 335	Phase 2 *n* = 996	Percent Error	Discrepancy
Child obesity	17.9%	42.0%	0.6	PPV=0.998, NPV=0.991
Time to event/censoring (age, yrs)	4.3 (2.9, 6.0)	4.8 (3.0, 6.0)	4.7	1.0 (range 0.04, 1.8)
Maternal weight gain (kg/wk)	0.30 (0.26, 0.38)	0.30 (0.22, 0.41)	100	– 0.02 (range – 0.66, 0.93)
Maternal BMI (kg/m^2)	25.9 (22.6, 30.5)	27.9 (23.8, 33.1)	100	0.13 (range – 6.8, 8.6)
Maternal age (yrs)	28.0 (23.5, 32.3)	27.4 (23.0, 31.8)	0	-
Maternal race			5.4	
White	61.8%	56.8		PPV=0.952, NPV=0.962
Black	23.1%	29.7		PPV=0.986, NPV=0.993
Asian	6.9%	4.0		PPV=0.904, NPV=0.998
Other/Unknown	8.2%	9.4		PPV=0.778, NPV=0.966
Maternal ethnicity, Hispanic	14.9%	14.9%	1.1	PPV=0.948, NPV=0.996
Maternal diabetes			10.9	
None	83.3%	89.4		PPV=0.991, NPV=0.553
Gestational	13.7%	6.7		PPV=0.420, NPV=0.992
Type 1 or 2	3.0%	3.9		PPV=0.472, NPV=0.977
Cesarean delivery	36.2%	38.2%	1.3	PPV=0.989, NPV=0.986
Child sex, male	52.7%	55.4%	0.4	PPV=0.995, NPV=0.998
Maternal depression	8.9%	10.9%	13.5	PPV=0.376, NPV=0.926
No private insurance	45.9%	67.6%	24.3	PPV=0.941, NPV=0.580
Singleton	98.1%	97.3%	1.2	PPV=0.992, NPV=0.826
Maternal smoking	6.3%	13.2%	11.8	PPV=0.618, NPV=0.897
Married	-	51.8%	-	-
Number prior live births	-	0.5 (0, 1)	-	-
Gestational age (wks)	-	39.1 (38.1, 40.3)	-	-
Child asthma	10.4%	13.0%	10.4	PPV=0.570, NPV=0.973
Maternal asthma	7.8%	11.0%	4.5	PPV=0.827, NPV=0.968

Regression Estimates

	Log hazard ratios for childhood obesity					
	Phase 1		IPW		$Raking_{Nv}$	
	β	SE	β	SE	β	SE
Maternal weight gain (kg/wk)	0.87	0.18	1.17	0.33	1.06	0.27
Maternal BMI (5 kg/m^2)	0.28	0.02	0.32	0.03	0.32	0.03
Maternal age (10 yrs)	-0.05	0.04	0.15	0.11	0.15	0.11
Maternal race, Black	-0.03	0.06	-0.24	0.14	-0.24	0.14
Maternal race, Asian	0.24	0.11	0.08	0.25	0.10	0.25
Maternal race, other/unknown	0.41	0.08	0.04	0.17	0.04	0.17
Maternal ethnicity, Hispanic	0.72	0.06	0.95	0.15	0.95	0.14
Maternal diabetes, gestational	0.12	0.06	-0.54	0.22	-0.54	0.22
Maternal diabetes, type 1/2	0.13	0.12	-0.19	0.27	-0.15	0.26
Cesarean delivery	0.12	0.05	0.17	0.10	0.17	0.10
Child sex, male	0.12	0.05	-0.15	0.10	-0.15	0.10
Maternal depression	0.08	0.08	-0.19	0.18	-0.17	0.18
No private insurance	0.18	0.05	0.60	0.14	0.59	0.14
Singleton	0.44	0.21	-0.00	0.33	0.03	0.32
Maternal smoking	0.32	0.10	0.48	0.17	0.46	0.17
Married			0.32	0.13	0.31	0.13
Number prior live births			-0.07	0.05	-0.08	0.05
Gestational age (wks)			0.03	0.02	0.03	0.02

Hazard Ratio

Holding all other factors constant, a child from a woman who gained 250 grams more per week during pregnancy (i.e., 10 kg in added weight over a 40 week pregnancy) had an estimated 30% increased hazard of obesity before age 6 (**HR=1.30; 95% CI 1.14-1.48**) based on the generalized raking estimator.

Unvalidated phase 1 data estimated a 24% increased hazard of obesity (**HR=1.24; 95% CI 1.14-1.36**).

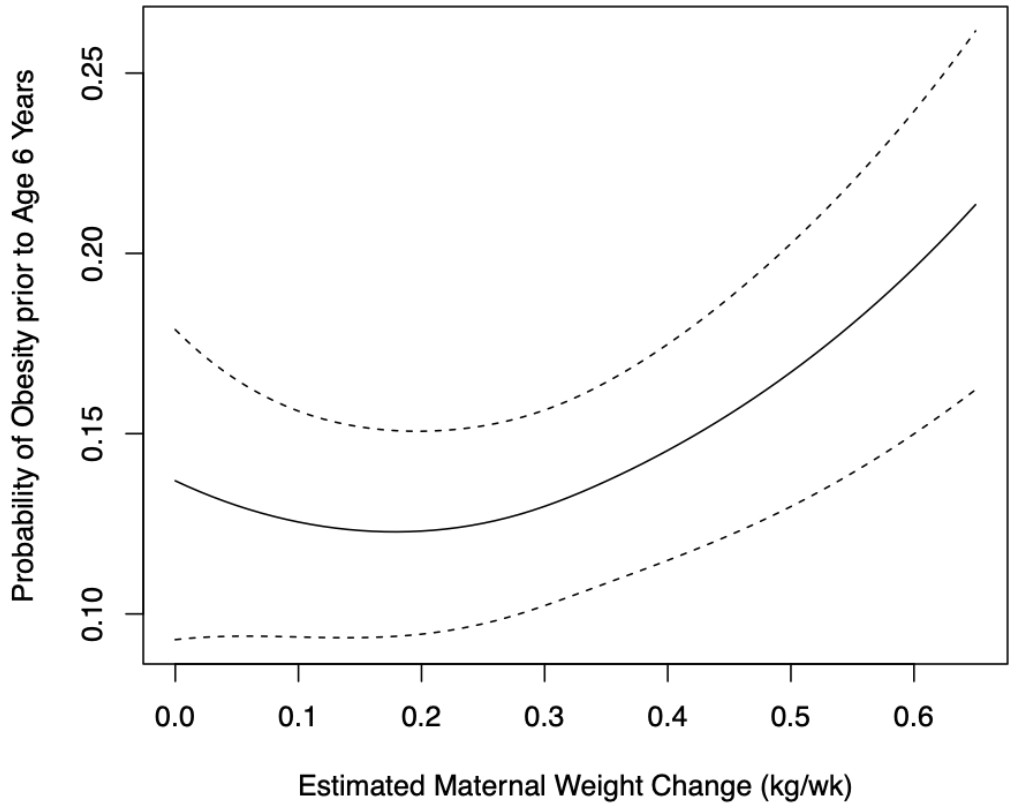
Nonlinear Association

An additional analysis raking with the naive influence function suggested that the relationship between maternal weight gain during pregnancy and childhood obesity was non-linear (p=0.007), with a fairly constant hazard of obesity for women who gained under 11-12 kg during pregnancy, but increasing hazards thereafter; no such non-linear relationship was seen using the phase 1 data alone (p=0.87).

Table: Adjusted hazard ratios for childhood obesity based on maternal weight gain per week during pregnancy. (Median weight gain was 0.28 kg/wk or about 11 kg over pregnancy.)

	Hazard Ratio	95% Confidence Interval
Average maternal weight gain per week during pregnancy (kg/wk)		
0	1.12	0.88, 1.43
0.1	1.02	0.93, 1.12
0.2 (reference)	1	
0.3	1.06	0.99, 1.14
0.4	1.20	1.03, 1.39
0.5	1.39	1.14, 1.70
0.6	1.66	1.32, 2.09

Nonlinear Association



All other covariates are set to their medians / modes. Three knots were used in restricted cubic splines.

Other Secondary Associations

Childhood obesity analysis:

	% Error	Naive HR (95% CI)	Raking HR (95% CI)
No private insurance	24.3	1.20 (1.09, 1.32)	1.80 (1.37, 2.37)

Childhood asthma analysis:

	% Error	Naive OR (95% CI)	Raking OR (95% CI)
Weight gain (250 g/wk difference)	10.4	0.88 (0.75, 1.02)	1.07 (0.74, 1.53)

Discussion

- First multiwave validation study
  - Majority of EHR studies do not validate data
  - Small subset that do validate, typically validate suboptimal records
  - Very few properly incorporate validation data into analyses
- Developed R package, `optimall` (Jasper Yang)
- Maternal weight gain during pregnancy associated with childhood obesity
- A lot of work to come to same conclusion as naive analysis
  - Don't know until you do it

Discussion (continued)

- Limitations
  - Validated data are not necessarily truth
  - Other challenges with using EHR data (e.g., confounding, erratic data capture, missing data)
- Future research
  - Other analysis approaches
    - Multiple imputation, semiparametric likelihood methods
    - Optimal validation designs with multiple parameters of interest
    - On-going validation studies in multi-cohort HIV collaborations

Thank You