Methodological aspects of small area estimation from the National Electronic Health Records Survey (NEHRS). [1]

Vladislav Beresovsky (hvy4@cdc.gov),
Janey Hsiao

National Center for Health Statistics, CDC

[1] Disclaimer: The findings and conclusions in this presentation are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention.

What is National Electronic Health Records Survey?

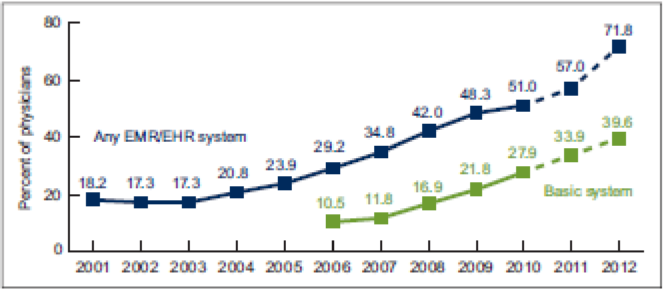
NEHRS is a mail survey of office-based physicians designed to collect data about adoption of electronic medical and health records (EMR/EHR) systems.

Survey frame combines databases of the American Medical Association (AMA) and American Osteopathic Association (AOA).

Starting from 2010 sample is selected within strata defined by states to facilitate state-level estimates.

Motivation for state-level estimates of adoption of EMR/EHR systems by office-based physicians.

Adoption of EMR is rapidly increasing *nationally* every year:



Affordable Care Act (ACA) provides incentives for physicians to adopt EMR systems. States' policies may be different in this process.

Question: How national growth of EMR adoption is reflected locally in states?

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Another motivation: small area estimates from the National Ambulatory Medical Care Survey (NAMCS).

NAMCS is annually conducted national survey of office-based physicians and providers practicing in community health centers (CHC) (`http://www.cdc.gov/nchs/ahcd.htm`).

Prior to 2012 NAMCS had multilevel design with PSUs spread around the US for efficient *national* estimates.

After 2012 NAMCS design was changed to become similar to NEHRS with the same purpose- to facilitate estimates in larger states and smaller states grouped within Census Divisions.

Methodology for model-based SAE developed for NEHRS could be applied to the post-2012 NAMCS data.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Research summary.

- Estimate state-level proportion of adoption of Any and Basic EMR by applying area-level Fay-Herriot model (Fay and Herriot, 1979) to binomial outcome variable.
- Conduct non-parametric simulations to evaluate results of model-based estimation:
  - Importance of using Area Resource File (ARF) covariates for model fitting;
  - Importance of correcting model-based predictions for bias;
  - Relative efficiency of model-based and survey estimates;
  - Coverage properties of the estimated confidence intervals;

CENTERS FOR DISEASE CONTROL AND PREVENTION

Methodology for model-based SAE

Area-level Fay-Herriot models applied to *Log* of in-state proportions of Any and Basic components of EMR systems, see the Bureau of Census technical report "Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties", Bell *& all* (2007) :

$$\log(y_i) = \log(Y_i) + e_i \quad \text{where} \quad e_i \sim ind.\, N(0, v_i)$$
$$\log(Y_i) = x_i\beta + u_i \quad \text{where} \quad u_i \sim i.i.d.\, N\left(0, \sigma_u^2\right)$$

where, for state *i*,

y_i = survey estimate of proportion of Any or Basic adoption of EMR

Y_i = true proportion

e_i = sampling error of estimate $\log(y_i)$

v_i = estimated variance of sampling error

x_i = vector of in-state regression coefficients

β = vector of regression parameters

u_i = state random effect

σ_u^2 = variance of state random effect

CDC CENTERS FOR DISEASE CONTROL AND PREVENTION

Model covariates

County-level covariates from the Area Resource File (ARF) comprise demographic, economic and healthcare information from various surveys and administartive databases.

However, the universe of studied data are physicians, while universe of covariates are population in counties.

How to aggregate county information into state-level covariates?

$$X_{\mathrm{i}}^{st,norm} = \frac{X_{\mathrm{i}}^{st,w}}{N_{\mathrm{i}}^{st,phys}} = \frac{\sum\limits_{\mathrm{j}\in\mathrm{i}} \left(X_{\mathrm{j}}^{c}/N_{\mathrm{j}}^{c,pop}\right) N_{\mathrm{j}}^{c,phys}}{\sum\limits_{\mathrm{j}\in\mathrm{i}} N_{\mathrm{j}}^{c,phys}}$$ covariates proportional to population

$$P_{\mathrm{i}}^{st,norm} = \frac{P_{\mathrm{i}}^{st,w}}{N_{\mathrm{i}}^{st,phys}} = \frac{\sum\limits_{\mathrm{j}\in\mathrm{i}} P_{\mathrm{j}}^{c} N_{\mathrm{j}}^{c,phys}}{\sum\limits_{\mathrm{j}\in\mathrm{i}} N_{\mathrm{j}}^{c,phys}}$$ covariates as percent of population

Expressions for predictions and variances, Bell & *all* (2007)

Log scale:

In-state estimates of logarithm of predicted proportions:

$$\widehat{\log(Y_i)} = (1 - w_i)\log(y_i) + w_i\left(x_i'\hat{\beta}\right), \text{ where } w_i = \frac{v_i}{v_i + \hat{\sigma}_u^2}$$

Variances:

$$\mathrm{var}\left(\widehat{\log(Y_i)} - \log(Y_i)\right) = w_i\hat{\sigma}_u^2 + w_i^2\left(x_i'\mathrm{Var}\left(\hat{\beta}\right)x_i\right)$$

Back to the original scale:

Estimates of predicted proportion in states:

$$\hat{Y}_i = exp\left(\widehat{\log(Y_i)}\right) exp\left[\mathrm{var}\left(\widehat{\log(Y_i)} - \log(Y_i)\right)/2\right]$$

Variances:

$$\mathrm{var}\left(\hat{Y}_i\right) = \hat{Y}_i^2\left[exp\left(\mathrm{var}\left(\widehat{\log(Y_i)} - \log(Y_i)\right) - 1\right)\right]$$

CENTERS FOR DISEASE CONTROL AND PREVENTION

Importance of ARF covariates for fixed-effects models.

log of proportion of adoption of **Any EMR**

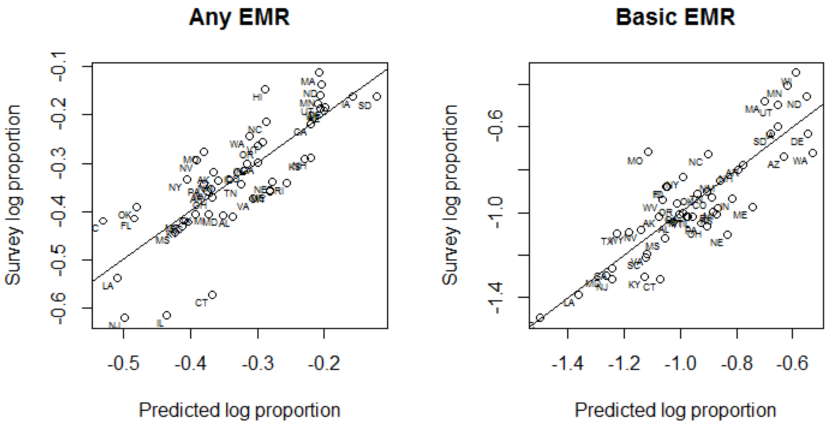
Model	**# of covariates**	R^2	AIC	$\hat{\sigma}_e^2$
Model_0	0		-217.4	0.0138
Model_A	2	0.38	-238.1	0.0088
Model_B	4	0.6	-256.5	0.006
Model_C	5	0.65	-261.1	0.0053

log of proportion of adoption of **Basic EMR**

Model	**# of covariates**	R^2	AIC	$\hat{\sigma}_e^2$
Model_0	0		-136.9	0.067
Model_A	2	0.47	-165.0	0.037
Model_B	4	0.56	-170.0	0.032
Model_C	8	0.72	-187.0	0.022

CENTERS FOR DISEASE CONTROL AND PREVENTION

Fit by fixed-effects model



CDC
Centers for Disease Control and Prevention

Fit by Fay-Herriot models depending on covariates.

log of proportion of adoption of **Any EMR**

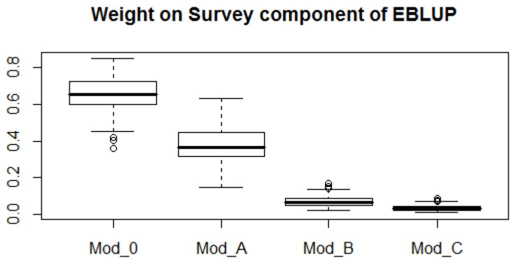
Model	Log-likelihood	AIC	$\hat{\sigma}_u^2$
Model_0	39.04	-74.08	$8.29 * 10^{-3}$
Model_A	55.62	-103.25	$2.55 * 10^{-3}$
Model_B	67.41	-122.83	$3.02 * 10^{-4}$
Model_C	68.92	-123.83	$1.4 * 10^{-4}$

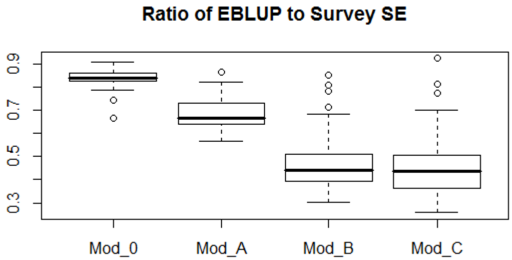
log of proportion of adoption of **Basic EMR**

Model	Log-likelihood	AIC	$\hat{\sigma}_u^2$
Model_0	-2.91	9.83	$4.62 * 10^{-2}$
Model_A	11.15	-14.30	$2.00 * 10^{-2}$
Model_B	15.20	-18.39	$1.57 * 10^{-2}$
Model_C	24.55	-29.11	$7.34 * 10^{-3}$

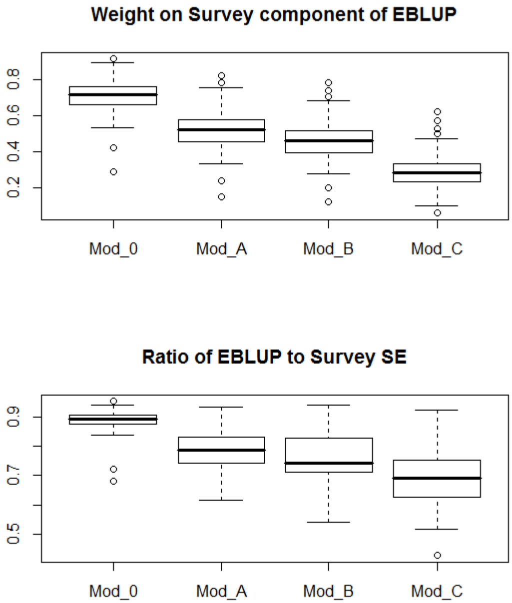
CENTERS FOR DISEASE CONTROL AND PREVENTION

EBLUP estimator of proportion of Any EMR





EBLUP estimator of proportion of Basic EMR



Simulation of NEHRS population and sample data.

Within each state applied 'inverse sampling' to generate NEHRS population from sample data, see §6 in Sverchkov and Pfeffermann, Survey Methodology (2004).

- Step 1. Generate a single 'pseudo population' by selecting *with replacement* $N = \sum w_i$ units from the original sample with probabilities $\pi_i = w_i / N$.
- Step 2. Select *without replacement* a single sample of size n with probabilities $\pi_i = 1/w_i$.
- Step 3. Model simulated samples with different number of ARF covariates selected with stepwise regression.
  Mod0 – no covariates;
  Mod1 – same 2 covariates as Model_A;
  Mod2 $\sim$ 5 covariates;
  Mod3 $\sim$ 10 covariates

CENTERS FOR DISEASE CONTROL AND PREVENTION

Average fit by Fay-Herriot models on simulated data.

log of proportion of adoption of **Any EMR**

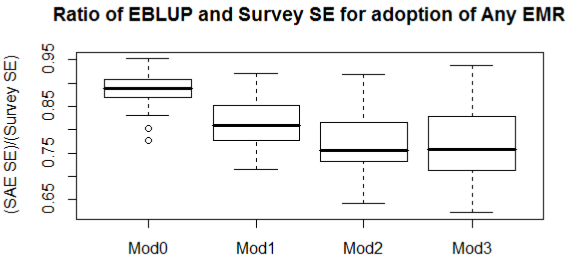
Model	Log-likelihood	AIC	$\hat{\sigma}_u^2$
Mod0	30.32	-56.64	$1.36 * 10^{-2}$
Mod1	42.36	-76.72	$7.41 * 10^{-3}$
Mod2	50.24	-85.97	$4.94 * 10^{-3}$
Mod3	56.43	-87.52	$3.79 * 10^{-3}$

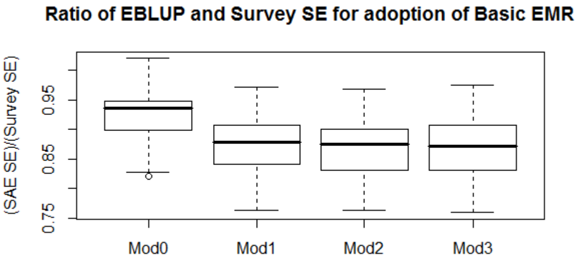
log of proportion of adoption of **Basic EMR**

Model	Log-likelihood	AIC	$\hat{\sigma}_u^2$
Mod0	-11.49	26.98	$7.35 * 10^{-2}$
Mod1	-0.95	9.89	$4.43 * 10^{-2}$
Mod2	1.81	9.82	$4.10 * 10^{-2}$
Mod3	6.42	9.95	$3.61 * 10^{-2}$

CDC Centers for Disease Control and Prevention

Average SE of EBLUP on simulated data





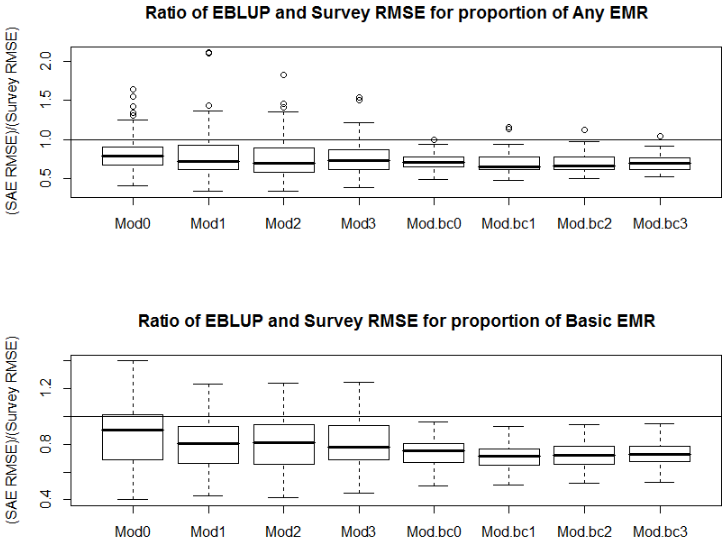
Bias-corrected EBLUP.

Efron and Morris (1972) remarked that EBLUP while performing well overall, may perform poorly for individual small areas. They proposed estimator, which compromises between limiting the maximum possible risk to any component and preserving the average gain of EBLUP by restricting the amount by which estimator differs from survey estimator y_{i} by some multiple of its standard error:

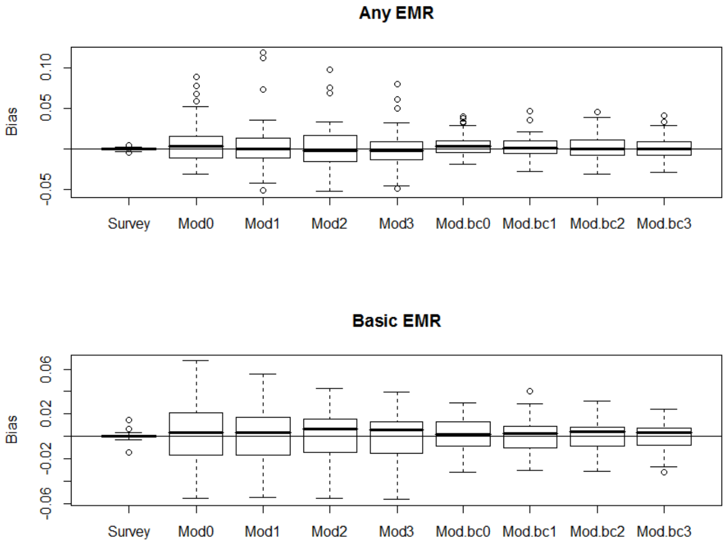
$$\widehat{Y_{\mathrm{i}}^{\mathrm{bc}}} = \begin{cases} \hat{Y}_{\mathrm{i}} & \text{if} \quad y_{\mathrm{i}} - c\sqrt{v_{\mathrm{i}}} < \hat{Y}_{\mathrm{i}} < y_{\mathrm{i}} + c\sqrt{v_{\mathrm{i}}} \\ y_{\mathrm{i}} - c\sqrt{v_{\mathrm{i}}} & \text{if} \quad \hat{Y}_{\mathrm{i}} < y_{\mathrm{i}} - c\sqrt{v_{\mathrm{i}}} \\ y_{\mathrm{i}} + c\sqrt{v_{\mathrm{i}}} & \text{if} \quad \hat{Y}_{\mathrm{i}} > y_{\mathrm{i}} + c\sqrt{v_{\mathrm{i}}} \end{cases} .$$

Regular choice considered by most authors is $c = 1$.

Root Mean Square Error (RMSE) of regular and bias-corrected EBLUP compared to Survey estimator

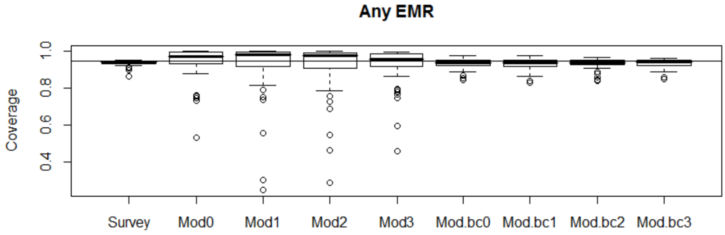


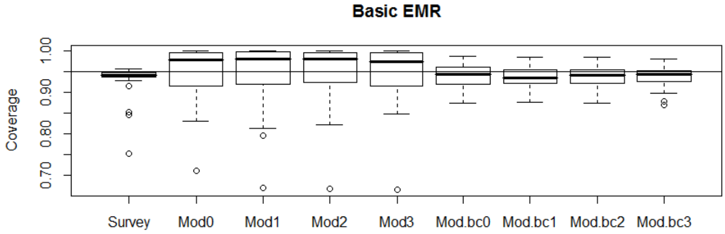
Bias of regular and bias-corrected EBLUP, and Survey estimators for proportion



CENTERS FOR DISEASE CONTROL AND PREVENTION

Coverage of regular and bias-corrected EBLUP.





Summarized comparison of RMSE of Survey and EBLUP estimators.

Table : RMSE of estimates of proportion averaged over states.

Method	Survey	Mod0	Mod1	Mod2	Mod3
	Adoption of Any EMR				
EBLUP	0.048	0.040	0.039	0.037	0.037
Bias-corrected EBLUP	...	0.034	0.033	0.033	0.034
	Adoption of Basic EMR				
EBLUP	0.054	0.046	0.043	0.043	0.043
Bias-corrected EBLUP	...	0.040	0.038	0.039	0.040

CDC
Centers for Disease Control and Prevention

Hypotheses testing using Survey and bias-corrected EBLUP for Any EMR.

Table : Number of states per simulation for which one-sided null-hypotheses were rejected with p=95%. In brackets- average power of tests.

Hypotheses	$< 25\%$	$< 50\%$	$> 50\%$	$> 75\%$
Method				
Survey	3.16 (0.23)	8.78 (0.33)	12.51 (0.47)	5.03 (0.38)
Mod0	3.80 (0.28)	11.21 (0.42)	13.94 (0.53)	6.73 (0.51)
Mod1	3.84 (0.28)	12.7 (0.48)	**14.5 (0.55)**	7.29 (0.55)
Mod2	3.94 (0.29)	**12.86 (0.49)**	14.28 (0.54)	7.52 (0.57)
Mod3	**4.15 (0.30)**	12.7 (0.48)	14.30 (0.54)	**7.59 (0.58)**

CENTERS FOR DISEASE CONTROL AND PREVENTION

Hypotheses testing using Survey and bias-corrected EBLUP for Basic EMR.

Table : Number of states per simulation for which one-sided null-hypotheses were rejected with p=95%. In brackets- average power of tests.

Hypotheses	$< 25\%$	$< 50\%$	$> 50\%$	$> 75\%$
Method				
Survey	2.89 (0.21)	9.64 (0.36)	11.96 (0.44)	6.18 (0.47)
Mod0	2.89 (0.21)	11.08 (0.41)	13.83 (0.52)	7.29 (0.56)
Mod1	3.61 (0.27)	11.12 (0.42)	**14.33 (0.54)**	**7.58 (0.58)**
Mod2	3.79 (0.29)	11.11 (0.42)	14.27 (0.54)	7.54 (0.57)
Mod3	**3.92 (0.29)**	**11.26 (0.42)**	14.03 (0.53)	7.53 (0.57)

CENTERS FOR DISEASE CONTROL AND PREVENTION

Conclusions and future work

- In many cases relation of numerous ARF covariates to healthcare indicators is very *indirect*. Despite significant improvement of fitting *sample* data by area-level models, their effect on fitting *population* data may be very limited.
- EBLUP has smaller MSE than Survey estimator. *Most of this gain is achieved due to its shrinkage nature.*
- Bias-corrected EBLUP was demonstrated to be more robust to outliers, had smaller bias and better coverage of population value.
- EBLUP provided for higher power of testing hypotheses comparing to Survey estimators.
- Proportions of adoption of EMR systems on Any and Basic levels are correlated, see adoption of EMR plot, where both are growing annually in similar way. Accounting for obvious correlations between these two variables may reduce errors of estimation. It seems reasonable to model joint distribution of these variables.

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION