A Maximum Likelihood Approach to Electronic Health Record Phenotyping Using Positive and Unlabeled Patients.

Lingjiao Zhang, Xiruo Ding, Yanyuan Ma, Naveen Muthu, Imran Ajmal, Jason H Moore, Daniel S Herman and Jinbo Chen

Nigel Petersen, Shruthi Vaidyanathan, Chelsea Murphy

December 6, 2022

Introduction: Why EHR data?

- Healthcare is important!
- EHR data allows us to use techniques from machine learning and statistical modeling.
- Potential for advancements in modern healthcare.

Motivation: Positive-only learning

- Obtaining good data is often expensive.
- Quality vs Quantity tradeoff with EHR data.
- Introduces the need for learning from positive-only data.

Prior Methods: Naive logit

- Treat all unlabelled patients as controls.
- Fit logistic regression on fully labelled data

$$\text{logit } \mathbb{P}(Y = 1 \mid \mathbf{X}, \beta) = \mathbf{X}^T \beta$$

- Possible to assign differing labels to several patients with similar covariates.

Prior Methods: Elkan-Noto Algorithm

- An anchor variable is a particular feature with high positive predictive value, typically chosen by a domain expert.
- Data consists of pairs $\{(\mathbf{X}_i, S_i)\}_{i=1}^{N}$
- Prediction depends on the relationship between

$$\mathbb{P}(S_i = 1 \mid \mathbf{X}), \quad \mathbb{P}(Y_i = 1 \mid \mathbf{X}), \quad c = \mathbb{P}(S = 1 \mid Y = 1)$$

- Requires a strong assumption on the observed labels.

Methods: Maximum Likelihood Approach

- Working logistic regression model

$$\text{logit}\mathbb{P}(Y = 1 \mid \mathbf{X}, \beta) = \mathbf{X}^T \beta$$

- Estimate β and c by maximum likelihood

$$L(\beta, c) = \prod_{i=1}^{N} p(X_i, S_i = 1)^{S_i} p(X_i, S_i = 0)^{1-S_i}$$

Methods: Model Assumptions

- Anchor has maximal positive predictive value

$$\mathbb{P}(Y = 1 \mid S = 1) = 1$$

- Sensitivity c is independent of covariates

$$c = \mathbb{P}(S = 1 \mid Y = 1) = \mathbb{P}(S = 1 \mid \mathbf{X}, Y = 1)$$

- A prior result that holds under the above assumptions

$$\mathbb{P}(Y = 1 \mid \mathbf{X}) = c\mathbb{P}(S = 1 \mid \mathbf{X})$$

Methods: Model Properties

- Possible to obtain estimators of other quantities of interest

$$\hat{h} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(S_i = 1) \qquad \hat{q} = \hat{h}/\hat{c} = \hat{\mathbb{P}}(Y = 1)$$

- Estimators of β and c are consistent.

- Does not requires strong assumptions made by EN algorithm.

Results: Datasets Used

Dataset 1: Simulated EHR data

- Sample of 10000 generated from logistic distribution with predictors mimicking EHR variables
- Anchor variable generated from Bernoulli distribution with anchor sensitivity c as success probability

Results: Datasets Used

Dataset 2: Real-world data with simulated anchor

- 10000 patients from Penn Medicine EHR with hypertension
- Testing for primary aldosteronism (PA) as the main phenotype
- Anchor artificially generated with sensitivity 0.2 by setting S=1 in 20% of cases

Results: Datasets Used

Dataset 3: Real-world data with real-world anchors

- 6193 patients with PA lab test orders
- Case set A: Diagnostic procedure for PA
- Case set B: Lab test order for procedure

Results: ML Method Performance

- The ML method generated consistent estimates of anchor sensitivity and higher PPVs and TPRs across all three data scenarios, in comparison to the EN algorithm and the naive logit modeling

Model	Simulation	Method Val.	Set A	Set B
ML Method	89%	44%	66%	78%
Ideal learning	90%	45%	None	None
EN algorithm	84%	39%	59%	61%
Naive logit	84%	38%	28%	41%

Reflection

- On the whole the paper proposes a useful alternative method to phenotype EHR data, in comparison to existing algorithms
- The motivation and the assumptions for the method proposed were all clearly outlined
- The result comparisons between algorithms also clearly outlined the efficiency of the ML method in comparison to the existing EN algorithm and naive logit model
- The proposed method can be applied to many phenotype scenarios of "positive-only" data in EHR
- Future research could involve expanding the method to include variable selection to build the model

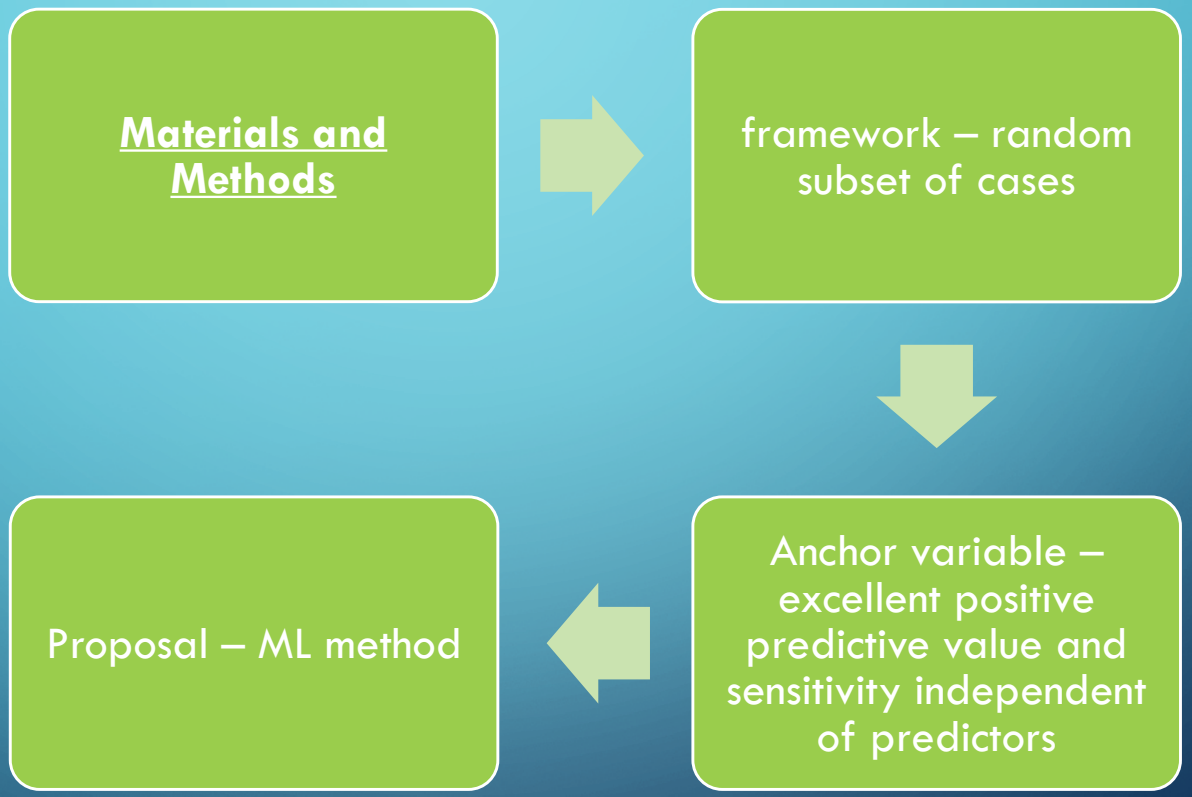
CONCLUSION

KEY SUMMARY POINTS

Objective

- Electronic Health Record (EHR) – requires labeled cases and controls
- Assigning labels requires manual chart review => labor intensive
- Identifying gold-standard controls is prohibitive for some phenotypes
- An accurate EHR phenotyping approach does not require labeled controls

WHAT ARE THE KEY SUMMARY POINTS



WHAT ARE THE KEY SUMMARY POINTS?

- **Results**
- ML method outperformed the EN algorithm and naïve logit on predictive accuracy in
- 1) DATASET1 - Monte Carlo Simulation Studies
- 2) DATASET2 - Application to identify hypertension patients with hypokalemia requiring oral supplementation using a simulated anchor
- 3) DATASET3 - Application to identify primary aldosteronism patients using real-world cases and anchor variables
- ML generated consistent estimates of 2 critical parameters

-phenotype prevalence

- Proportion of true labeled cases

Upon identification of scalable and transferable anchor variable to different practices, our approach should facilitate development of scalable, transferable and practice-specific phenotyping models.

Dataset1 Simulation Results

HOW CONCINVCING ARE THE RESULTS?

		ML	EN	Naive Logit	Ideal Learning
Simulation Results					
	Anchor sensitivity c	Identical to ideal learning	0.37 ESE (0.04)	?	Identical to ML
	Phenotype prevalence q	Identical to ideal learning	0.14 ESE (0.01)	?	Identical to ML
	AUC	0.994	0.993	0.993	0.994
	80% sensitivity (TPR)	PPV 86% FPR 0.7%	PPV 84% FPR 0.9%	PPV 83% FPR 0.9%	
	80% PPV	Higher TPR than EN and Naive Logit			
			4% EN (ESE:0.5%) with predicted probabilities > 1 (even after increase validation set size to 5000)		

Compared to the ideal logistic regression, ML method yielded identical estimates of the anchor sensitivity c and the phenotype prevalence q and a comparable predictive accuracy.

At their respective threshold for 80% sensitivity TPR (PPV), ML yielded higher specificity(TPR) compared to the EN and naïve logit.

4% of EN algorithm – predicted probabilities > 1

HOW CONVINCING ARE THE RESULTS?

Dataset 2: Method Validation Using Real-World EHR Data and a Simulated Anchor Variable

ML method achieved a PPV (70% TPR) and FPR that are identical to that of ideal learning and a comparable TPR (50% PPV).

5% of EN algorithm – predicted probabilities > 1

		ML	EN	Naive Logit	Ideal Learning
Method Validation using Real-World EHR Data and a Simulated Anchor Variable	AUC	0.85	0.83	0.85	0.86
	70% TPR	PPV 23%	PPV 20%	PPV 23%	PPV 23%
	FPRs	16%	19%	?	16%
	50% PPV	TPR 44%	TPR 39%	TPR 38%	TPR 45%
			5% EN with predicted probabilities > 1		

Dataset 3: A preliminary Phenotyping Model for PA Using Real-World Predictors and Cases

HOW CONVINCING ARE THE RESULTS?

		ML	EN	Naive Logit	Ideal Learning
A Preliminary Phenotyping Model for PA using Real-World Predictors and Cases					
Case Set A	C (sensitivity)	0.56	0.35	?	?
	q (phenotype prevalence)	4%	7%	?	?
	0.5 TPR	0.66	0.59	0.28	?
			0.6% of EN predicted probabilities > 1		
Case Set B	C	0.62	0.41	?	?
	q	5%	8%		
	TPR	Higher than EN & Naive Logit			
			0.7% of EN predicted probabilities > 1		

For both case sets, the ML fitted model seems to have high discriminatory power with anchor-positive cases having high predicted probabilities.

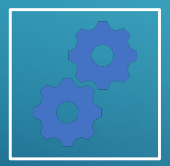
The ML method achieved consistently higher TPR than that of the EN and naïve logit.

WHAT CONCLUSIONS CAN BE DRAWN FROM THE PAPER?

- ML
- Facilitates accurate semiautomated EHR phenotyping with minimal manual labeling
- => Facilitates EHR clinical decision support and research

IS THIS PAPER SUBSTANTIALLY DIFFERENT THAN PREVIOUS WORK AND MADE A SUBSTANTIAL CONTRIBUTION TO THE LITERATURE?

Naïve Logit – fits a logistic regression to the positive-only data, where the unlabeled patients acts as controls.



EN Algorithm

Intuitive & easy to implement

Proposed estimator of c is biased

=> Biased estimation of $p(Y=1|X)$ and prevalence q

c biased toward 0 => $p(Y=1|X) > 1$

IS THIS PAPER SUBSTANTIALLY DIFFERENT THAN PREVIOUS WORK AND MADE A SUBSTANTIAL CONTRIBUTION TO THE LITERATURE?

- ML
- Fits model p(Y=1 | X; B) and estimates anchor sensitivity c at the same time
- ML method estimates are obtained by maximizing the log likelihood function and using the inverse of the information matrix to establish the large sample variance-covariance matrix of these estimates.
- The phenotype prevalence q = h/c is
- 1) estimated as q_hat = h_hat/c_hat where h_hat is estimated as the ML of the sample fraction of those with S=1 or P(S=1) divided by the estimate of anchor sensitivity
- 2) Or using the average of the estimated phenotype probabilities

IS THIS PAPER SUBSTANTIALLY DIFFERENT THAN PREVIOUS WORK AND MADE A SUBSTANTIAL CONTRIBUTION TO THE LITERATURE?

- ML method > EN algorithm and naïve logit in
- **Predictive accuracy metrics** and **estimates of anchor sensitivity and phenotype prevalence**
- **Transferability to other practices**
- The anchor concept may be more easily transferred than the full model
- Model validation regarding the calibration and predictive accuracy depend on the labels for a random set of patients. However, Anchor variable + ML method => innovative method for internally assessing model calibration and predictive accuracy using positive-only data, excluding the external model validation step
- Rather than validating a classically fit model, chart review need only be completed to verify that the anchor has high PPV for the phenotype of interest in order to generalize our method to secondary sites
- Recent work shows that phenotyping methods benefits from 1)noisy labels with ranodm error 2) anchor variable structure where we can extend our method in this regard.

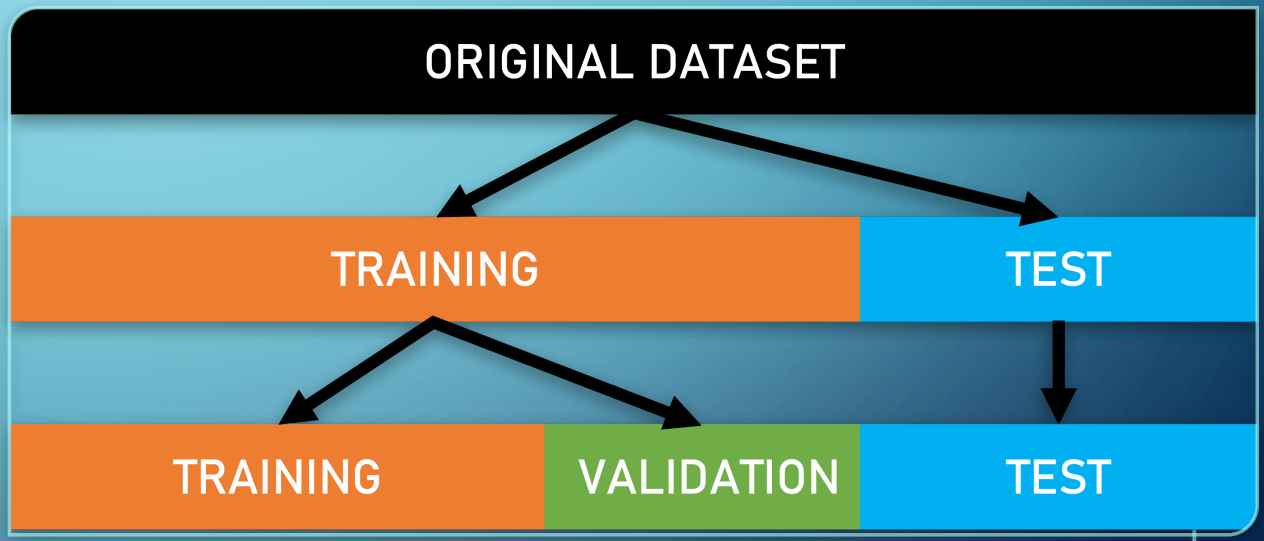
LIMITATIONS

WHAT MIGHT THE ISSUES BE IN APPLYING THE APPROACH TO ANOTHER DATA SET OR PROBLEM?

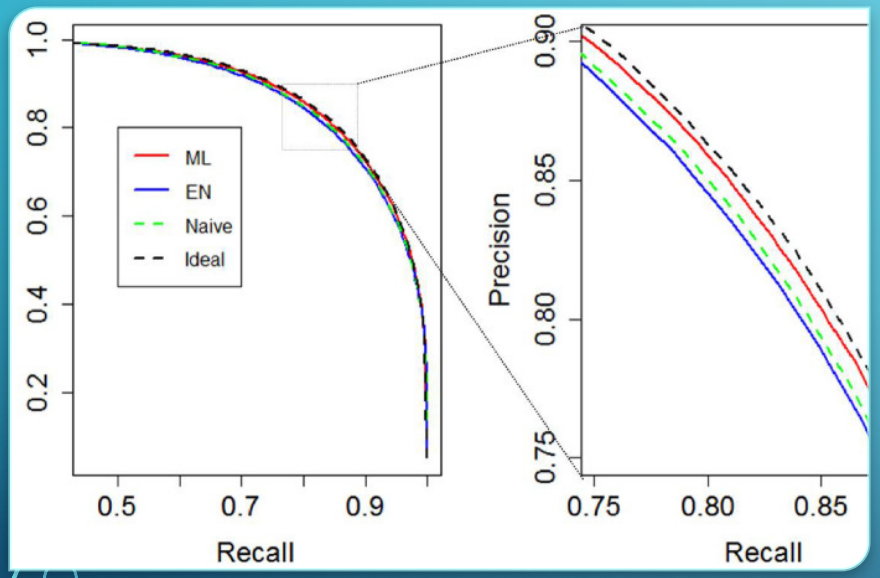
- Selecting and setting the value of an anchor variable in data
- Expert knowledge is essential
- Model performance relies on 2 assumptions of the anchor variable
- 1) has perfect positive predictive value but is not required to have high sensitivity
- 2) sensitivity is independent of all phenotype model predictors.
- Unclear => anchor approriateness could be supported by explicitly proving the estimated phenotype prevalence, model sensitivity, or the conditional independence assumption.

WHAT RESULTS ARE MISSING FROM THE PAPER?

- Validation of the actual data !!!
- The diagnosis of PA depends on specific diagnosis testing, chart review is not sufficient to identify all PA patients in a group => inplausible to attain a sufficient annotated validation set
- The paper cited a similar design called the Dutch study which does not necessarily match the data analyzed in this paper.



ARE THE AUTHOR'S CONCLUSIONS WELL-INFORMED BY THE SIMULATIONS, REAL DATA ANALYSIS, OR THEORETICAL RESULTS?



Precision vs. Recall

- **Dataset 1 Simulation Results - Not convincing**
- Simulated data is not the actual data
- Compares the model quality in a virtual environment
- May not reach the same conclusion in reality
- Model setting – referee and players
- Anchor variable is set by researcher

ARE THE AUTHOR'S CONCLUSIONS WELL-INFORMED BY THE SIMULATIONS, REAL DATA ANALYSIS, OR THEORETICAL RESULTS?

- **Dataset 2 – Method Validation using Real-World EHR Data and a Simulated Anchor Variable**
- An anchor variable S with sensitivity 0.2 by randomly setting S to 1 for 20% of all cases (Y=1), and to 0 for the remaining 80% of cases and for all controls
- Anchor is artificially created, only solves the problem in this particular dataset
- Although the method achieved a satisfied result, for datasets with few controls or no controls, it is hard to determine another sensitivity value and know whether the value is optimal based on the characteristics of another dataset

ARE THE AUTHOR'S CONCLUSIONS WELL-INFORMED BY THE SIMULATIONS, REAL DATA ANALYSIS, OR THEORETICAL RESULTS?

- **Dataset 3 – A Preliminary Phenotyping Model for PA using Real-World Predictors and Cases**
- No control variable => cannot create an artificial anchor variable and set its sensitivity value
- The researcher later supplemented this set using an anchor variable strategy
- We need to measure P(Y=1) but cannot observe P(Y=0)
- Only able to compare the model results to other algorithm results BUT unable to compare the model results to the actual results
- Need a plausible validation step for conclusion to be more convincing
- This can be achieved by choosing a dataset with adequate number of controls, using the methods mentioned above to select a disease-related anchor variable, make predictions and compare it with the actual results

HOW COULD THE AUTHOR'S CONCLUSIONS BE STRENGTHENED?

ML method achieved good sensitivity and PPV for identifying patients with aldosteronism

The proof-of-principle analysis has substantial room for improvement that focuses on specific predictors chosen by domain experts and has yet to thoroughly delve into feature selection and engineering

Applicable for developing phenotyping models when # of predictors < # of records

Explore additional predictors across high dimensional EHR data => improve accuracy and obtain precise estimates of anchor sensitivity

Extend the current ML method to promote variable selection when constructing the prediction model

Expect further improvements from more extensive modeling, including exploration of alternative missing data approaches