# 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.

# Contents

# 1 Report

## 1.1 Introduction and Motivation

As Electronic Health Record (EHR) data becomes increasing more prevalent in modern health care, the need for efficient and accurate analysis of EHR data continues to grow. As with most applications of machine learning and statistical modelling, obtaining an ideal set of training data is often infeasible. In the EHR setting particularly, obtaining a well labelled sample of data often comes at the cost of sample size, potentially compromising the accuracy of the model, which can lead to further issues as a result. This trade-off introduces the need to develop an approach to learning that holds in a more general setting, and a particular setting of interest for us will be predicting phenotypes from EHR data with positive-only labelling. At the time the paper was published, there were two existing approaches to the problem at hand, the naive logit model and the Elkan-Noto (EN) algorithm. The naive logit model, the standard approach at the time, assigns all of the unlabelled patients the "control" label and fits a standard logistic regression model on the now fully labelled data. This approach can lead to performance issues with the model, particularly in the case where unlabelled patients have very similar covariates to case-labelled patients. One possible remedy for this issue comes through the use of anchor variables. An anchor variable is a particular feature that has a strong relationship to the phenotype of interest. Typically chosen by a domain expert, the presence of a well chosen anchor variable can indicate a positive phenotype with high certainty. The other existing approach makes use of binary anchors variables, and under particular assumptions on the data, performs with higher accuracy. The EN algorithm includes an intermediate step in predicting phenotypes, which the proposed method will generalize. The approach takes several steps, first fitting a logistic regression model on a smaller training set to predict anchor status from the covariates, and using the remaining data as a validation set to estimate the sensitivity of the anchor variable. Finally, the predicted phenotype is a function of the two quantities estimated in the first two steps. The main issue with the EN algorithm is that its performance relies on the complete separability of the distributions of the values of the phenotype (case and control). When this assumption is not met, it can lead to a biased estimation of the anchor sensitivity, and hence biased estimation of the response. The short-comings of the naive logit model and EN algorithm further introduce the need for a more general approach to predicting phenotypes in positve-only EHR data.

## 1.2 Methods

The primary method introduced in this paper is a Maximum Likelihood (ML) method used to learn from positive-only EHR data. Before introducing the details of the ML method, we introduce necessary notation and terminology. A triple $(\mathbf{X}, Y, S)$ is considered an EHR random variable, where $\mathbf{X}$ is a vector of covariates, $Y$ is the phenotype, and $S$ is a binary anchor variable. They denote the anchor sensitivity by $c = \mathbb{P}(S = 1 \mid Y = 1)$, the phenotype prevalence by $q = \mathbb{P}(Y = 1)$, and the anchor prevalence by $h = \mathbb{P}(S = 1)$. As the ML method builds on some of the ideas introduced in the EN algorithm, we first introduce the latter. Both approaches rely on the common use of binary anchor variables, which must satisfy the following assumptions:

1. $c = \mathbb{P}(S = 1 \mid Y = 1) = \mathbb{P}(S = 1 \mid Y = 1, \mathbf{X})$, namely the sensitivity is independent of the covariates.
2. The anchor variable $S$ has the highest possible Positive Predictive Value (PPV)

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

Under the above assumptions, given $N$ pairs $\{(\mathbf{X}_i, S_i)\}_{i=1}^{N}$, the EN algorithm works as follows:

1. Randomly partition the observed data into training and validation sets, of sizes $n_t$ and $n_v$, respectively, where $N = n_t + n_v$.
2. Fit a logistic regression model on the training set.
3. Use the validation set to estimate the sensitivity $c$ by

$$\hat{c} = \sum_{i=1}^{n_v} \mathbb{P}(S_i = 1 \mid \mathbf{X}_i)\mathbb{I}(S_i = 1) \Big/ \sum_{i=1}^{n_v} \mathbb{I}(S_i = 1)$$

4. Use the results from parts 2 and 3 to predict $\mathbb{P}(Y = 1 \mid \mathbf{X}) = \mathbb{P}(S = 1 \mid \mathbf{X})/\hat{c}$

As mentioned in the previous section, the main problem with the EN algorithm was the additional need for complete separability of the distributions of the phenotype labels. This is a rather strong assumption that is not often met in practice, and the performance of the EN algorithm can be compromised when this assumption is not met. Particularly, a violation of this assumption can lead to bias in the estimator of $c$, which can potentially cause the predicted values to fall outside of $[0, 1]$. The ML method relies on a similar approach using anchor variables, but no longer requires such an assumption on the labels. It uses a working logistic regression model $\text{logit}\mathbb{P}(Y = 1 \mid \mathbf{X}, \beta) = \mathbf{X}^T\beta$ for prediction, and uses maximum

likelihood to find estimators $\hat{\beta}$ and $\hat{c}$ simultaneously. Under the two assumptions on the sensitivity and anchor variable mentioned prior, the ML approach maximizes the likelihood function of the sample $\{(\mathbf{X}_i, S_i)\}_{i=1}^N$ given by

$$\begin{aligned} L(\beta, c) &= \prod_{i=1}^{N} p(X_i, S_i = 1)^{S_i} p(X_i, S_i = 0)^{1-S_i} \\ &\propto \prod_{i=1}^{N} \{c\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i, \beta)\}^{S_i} \{1 - c\mathbb{P}(Y_i = 1 \mid \mathbf{X}_i, \beta)\}^{1-S_i} \end{aligned}$$

Both the EN and ML approaches rely on the common result that

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

linking the phenotype to the anchor prevalence and covariates $\mathbf{X}$ using $c$. Further quantities of interest can be estimated directly, or using $c$, namely the anchor prevalence $h$ is estimated by the proportion of samples with the anchor present, namely $\hat{h} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(S_i = 1)$. Using the relationship $q = h/c$, they obtain a plug-in estimator of $q$ using the estimators $\hat{c}$ and $\hat{h}$, namely $\hat{q} = \hat{h}/\hat{c}$. While the EN algorithm and ML method share similarities, like their use of binary anchor variables and dependence on (1), they differ in complexity, and their approach to estimation. The EN algorithm uses a validation set to estimate sensitivity using the logistic model fit on the training set, and uses (1) directly for prediction in the last step, whereas the ML model obtains the necessary estimators in a single step when maximizing the likelihood of the entire sample. The relationship between phenotypes and anchors is baked into the likelihood function directly, allowing for the estimation of sensitivity and regression coefficients simultaneously. The main difference between the naive logit model and the ML method is the use of anchor variables, the ML method makes use of anchor variables and sensitivity to link the phenotypes to the anchor and covariates, rather than assigning control to unlabelled patients, which can negatively influence the generalizability of the model. Lastly, the ML method is a rather desirable one because of its simplicity, weaker assumptions on the data (as compared to the EN algorithm) and because of the consistency of the ML estimators $\hat{\beta}$ and $\hat{c}$, shown through simulation methods.

## 1.3 Results

The proposed ML method was initially tested on a data set that was simulated to imitate the data structure of an EHR. Following this, the ML method was validated using real-world EHR data with a simulated anchor variable. Finally, the method was applied to develop a model to phenotype for Primary Aldosteronism (PA) using a real-world data set with real-world anchor variables. During this process, comparisons were also made with the existing EN algorithm and the naive logit model, as well as an ideal learning method using the actual phenotype label Y as the outcome which was used as the benchmark.

The simulated data set was generated to imitate the data structure of typical EHR systems. The outcome variable $Y$ is generated from a logistic regression model with 9 different independently distributed predictors of varying strength determined by the coefficients, generated from three types of distributions. The formula and predictor distributions for the model is provided in Figure 1 in the Figures section. The anchor sensitivity $c$ was set at 0.5, with anchor variable $S$ generated from a Bernoulli distribution Bernoulli($c$) for each case $Y = 1$. For each control $Y = 0$, the anchor variable is set to 0. The generation of the anchor variable $S$ only for the cases ensures that it has high PPV, which fulfills the first main assumption of the method in relation to the anchor variable. The second anchor variable assumption of independence is also fulfilled, since the generation of $S$ using the Bernoulli distribution is also not dependent on any of the covariates. For the EN algorithm, each of the simulations obtained a random sample of 10000 as the training set, and a testing set of 5000. Additionally, 20% of the training set (2000) were put aside as a validation set to estimate anchor sensitivity. The simulation was repeated 1000 times for each combination of parameters. The results of the simulation data provided in the paper were primarily focused on phenotype prevalence $q = 0.1$, with results for the other prevalence values being quite similar.

Two different real-world data sets were used from the Penn Medicine primary care EHR for the ML method validation and the phenotyping model respectively. In the data set for the ML method validation, 10000 patients with hypertension were selected for PA screening. All patients were adults and were randomly selected between 2007 and 2017, restricted to patients with hypertension by only selecting those with 2 or more outpatient encounters where they had a hypertension diagnosis. The phenotype variable $Y$ was defined as the patient having 3 or more outpatient orders for oral potassium supplements, with the final dataset containing 798 cases and 9204 controls. The predictors were selected from diagnosis codes, lab results, prescriptions, test results, vitals and other EHR information by field experts. The anchor variable was simulated with anchor sensitivity 0.2 by setting S to 1 for

20% of cases and 0 for 80% of cases and all controls. This anchor variable generation satisfies both the high PPV assumption ($\mathbb{P}(Y = 1 \mid S = 1) = 1$) as it is only generated for cases, and the anchor sensitivity independence assumption it is also generated independent of the other covariates. For the EN algorithm, 20% of the data set was put aside as the validation set for anchor sensitivity estimation.

The final data set from the Penn Medicine primary care EHR was used to develop a model to phenotype for PA. The data set contained 6319 patients who had a laboratory test order to screen for PA. Similar to the method validation data set, the predictors were selected by clinical experts from various EHR variables. Since the ordering of laboratory tests can somewhat identify for a phenotype and is not random, a binary variable was also created to indicate the presence of the lab test result and included in the model with the actual test results. For the cases, an existing PA research registry was used, where patients were identified as cases since they had a diagnostic procedure called adrenal vein sampling, which was only performed for those who were diagnosed with PA. This was described as case set A, with 149 patients. This set was further added to with another set of patients using an anchor variable strategy. This case set B included 47 patients with a lab test order for adrenal vein cortisol, which was only performed as part of the above diagnostic procedure. Thus, there were two anchor variables identified, one with the actual diagnostic procedure, and another with the lab test that is included in the diagnostic procedure. Both anchor variables satisfy the assumption of high PPV as they are only conducted in confirmed cases, thus $\mathbb{P}(Y = 1 \mid S = 1) = 1$. Both anchors also satisfy the anchor sensitivity assumption as they seem to be generated independent of other covariates and only using the phenotype status. Based on this, the final number of cases in the dataset was 196. The True Positive Rate (TPR) was determined using the anchor-labeled case set and the PPV was estimated by chart review of 185 patients in longitudinal care where 132 were assigned positive for PA, 5 were unknown, and 48 were negative.

In all three situations, the proposed ML method not only provided consistent and accurate estimates, but performed significantly better than the existing EN algorithm and the naive logit methods across all data scenarios, both simulated and real-world. This indicates that it sufficiently solved the problem of phenotyping data with incomplete labels, and is applicable to real-world EHR data with more complex variable structures as well. In the simulation data, as demonstrated in Table 1 in the Figures section, the ML method had consistent estimates of anchor sensitivity and phenotype prevalence that were closer to the true values set during the simulation process and also to the ideal learning comparison benchmark. In comparison, the EN algorithm underestimated anchor sensitivity at 0.37 and thus overestimated phenotype prevalence at 0.14. At the decision thresholds for 80% sensitivity (TPR), the

ML method had higher specificity and higher PPV and a lower False Positive Rate (FPR). The decision thresholds for 80% PPV also demonstrated that the ML method had a higher sensitivity and was also the closest to the ideal learning method, as demonstrated in Table 2 in the Figures section. The estimates for $\hat{\beta}$, $\hat{c}$, and $\hat{q}$ by the ML method were also much closer to the true values, and the variances for these estimates were simpler to obtain.

In the method validation data, the ML method again showed the more efficient performance with the closest estimates of anchor sensitivity and phenotype prevalence to the true values. Additionally, the accuracy of the phenotyping was also demonstrated by the ML method and naive logit having similar "area under the ROC curve" (AUC) values to the ideal learning method, with the AUC values of the EN algorithm being the lowest. At the respective thresholds for 70% sensitivity and 50% PPV, the ML method performed better than the EN algorithm with higher PPV and higher TPR, and lower FPR as demonstrated in Table 2 in the Figures section.

Finally, in the preliminary model to phenotype for PA, the ML method performed better than the EN algorithm in both case sets, with more accurate estimates of anchor sensitivity and a closer estimate of phenotype prevalence. The TPR in the ML method was also consistently higher across thresholds compared to the EN algorithm and the naive logit methods, as shown in Table 2 in the Figures section. Most notably, the ML method identified 7 unlabeled patients as positive for PA, who did not meet the diagnostic code criteria as well as the diagnostic criteria for the PA lab test results.

## 1.4 Limitations

The issues in applying the approach to another data set or problem are selecting and setting the value of an anchor variable. Expert knowledge is essential to meticulously select a variable or a competitive variable and the model performance relies on the following assumptions of the anchor variable mentioned in 1.2. If unclear, the anchor appropriateness could be supported by explicitly proving the estimated phenotype prevalence, model sensitivity, or the conditions independence assumption.

The missing results from the paper should be the validation of the actual data. They applied the ML method to create models that identify patients with PA. When selecting as cases PA patients who went through a sub-typing diagnostic process, the focus is on patients with more serious and actionable disease rather than all PA patients. According to the conditional independence assumption, the prevalence of PA qualified for adrenal vein sampling is around 5% among patients who screened for PA. It would be ideal for a random sample of patients to have expert annotation. The prevalence of PA is low, meaning that lots of patients

would need to be annotated. However, the big issue is that the diagnosis of PA depends on specific diagnosis testing, chart review is not sufficient to identify all PA patients in a group. Consequently, it is implausible to attain a sufficient annotated validation set. Based on the literature, 5% of the prevalence of all PA in primary care populations and 10% for prevalence of PA in tertiary care settings. A recent similar design called the Dutch study show a 3% (95% CI 1.4%-4.9%) prevalence of PA varied by provocative testing, among patients newly diagnosed with hypertension and screened for PA. To conclude, there is a lack of data for validation, but the paper cited other data which does not necessarily match the data analyzed in this paper.

The authors' conclusions is not well informed by the simulations, real data analysis, or theoretical results. The simulation result is not very convincing because the simulated data is not the actual data, but compares the model quality in a virtual environment which means that we may not reach the same conclusion in reality. The model setting makes it both a referee and players where the anchor variable is set by the researcher, causing the simulation results to be unconvincing. From the precision vs. recall plot, there is not a significant difference among the results of ML, EN, Logistic and Ideal model and the slight difference can only be seen as we magnify the plot.

For the second dataset case, the anchor variable $S$ is artificially created 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. The anchor variable seems to be a major issue here as it is artificially created that only solves the problem in this particular dataset. Even though 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.

The third dataset has no control variable, we cannot create an artificial anchor variable and set its sensitivity value. The researcher later supplemented this set using an anchor variable strategy to include patients with a laboratory test order for adrenal vein cortisol which is only performed as part of this the adrenal vein sampling procedure. We need to measure $\mathbb{P}(Y = 1)$, but cannot observe $\mathbb{P}(Y = 0)$, therefore we are only able to compare the model results to other algorithm results but unable to compare the model results to the actual results. To make the conclusion more convincing, we need a plausible validation step to confirm the model results with the actual results. This can be achieved by choosing a dataset with adequate number of controls, using the methods aforementioned select a disease-related anchor variable, make predictions and finally compare it with the actual results.

The authors' conclusions can be strengthened in the following ways. The maximum likelihood method achieved good sensitivity and PPV for identifying patients with aldosteronism. This

proof-of-principle analysis has substantial room for improvement. The analysis focused on specific predictors chosen by domain experts, and has yet to thoroughly delve into feature selection and engineering. The method is applicable for developing phenotyping models when the number of potential predictors is far less than the number of records. Improved accuracy and precise estimates of anchor sensitivity can be achieved by exploring additional predictors across high dimensional EHR data. The author aims to extend the current ML method to promote variable selection when constructing the prediction model. Furthermore, we would expect huge further improvements from more extensive modeling, including exploration of alternative missing data approaches.

## 1.5 Reflection

As a whole, the paper effectively proposes a useful and efficient method to phenotype EHR data, especially in comparison to existing algorithms. Many aspects of the paper were clearly written out, such as the creation of the ML method. Each motivation to propose such a model and its structure and efficiency in comparison with existing methods are properly explained and justified. Additionally, the mathematical formulas and assumptions made for the method proposed were also clearly described with proper justification, such as the clear definition and assumptions for the anchor variable. Finally, the comparison of the model accuracy parameters in the results clearly outlined the efficiency of the ML method in comparison to existing methods in phenotyping incompletely labeled EHR data. All in all, the paper outlines a clear argument for the use of the proposed ML method and clearly justifies it using multiple data scenarios and results with clear comparisons. Some aspects of the paper, such as certain details about the simulation data were a bit unclear and could have included more justification. For example, a clarification on why certain distribution and coefficient values were used for some of the predictors would have made the method of simulation much clearer. Additionally, some of the methods described were included in the results section rather than the methods section, such as the application of the ML method to real world data. It would have been helpful to explain this application as well in the methods section and justify it before explaining the results of such an application in the results section along with the simulation results.

The ML method proposed in the paper can be applied to many different scenarios of phenotypes with partially labeled or "positive-only" data, specifically in the use of EHR data to predict the occurrence of particular diagnoses or diseases. An example of this could be neurological diseases such as strokes or epilepsy or even different types of cancers (Nogues et al, 2022). Additionally, as mentioned in the paper, future research could involve expanding the

ML method to include variable selection to build the model, since the method as it currently stands involves a previously selected set of variables to include. Another possible avenue to explore would be to identify different possible anchor variables for other phenotypes that also suffer from "positive-only" data, and evaluate the efficiency of the ML methods to these alternate scenarios.

## 1.6 Conclusion

**Key Summary Points**

The process of determining, analyzing or predicting patients' phenotypes using the typical EHR data depends on expert-annotated cases and controls which involves time-consuming standard medical chart review, at the same time determining benchmark controls is impossible in some phenotypes. Therefore, they developed an accurate EHR phenotyping approach that does not depend upon labeled controls. The foundation is based on a random sample of cases described in the form of an anchor variable with outstanding PPV and sensitivity independent of predictors. A ML approach that effectively manipulates available data using the specified cases and unlabeled patients to build logistic regression phenotyping models. The predictive accuracy of the ML method exceeds that of the existing algorithms in the three sets of cases in the paper. Phenotype prevalence and the fraction of labeled true cases are the two critical parameters of the ML method. As they identified an assignable anchor variable to different practices, the ML approach should promote the phenotyping models that is flexible, transferable and applicable.

**Conclusion**

To solve the problems of limiting use of EHR data in clinical support and research due to its incompleteness and asymmetry, we proposed a new likelihood-based approach that uses both the labeled cases and unlabeled patients to facilitate accurate semiautomated EHR phenotyping with the least possible standard labeling which in turn promote phenotype model development and transferability of a broad range of EHR clinical decision support and research applications.

The results are convincing. Refer to Table A1 in page 12 for detail!
The phenotyping models for PA based on the 2 case sets were similar, to a certain degree implies the robustness of the ML method in regard to anchor selection. The current implementation is extended so that anchor sensitivity can vary across a fixed number of discrete strata that are predefined by patient EHR data. This approach is then applied in PA modeling to exclude major predictor-anchor dependence. The proposed ML method that develops

a logistic regression prediction model using positive-only EHR data using cases identified by anchor variables has demonstrated, via three sets of cases, that this method develops models that accurately identify unlabeled cases and yields consistent estimate of phenotype prevalence.

The paper is rather different than previous work and made a substantial contribution to the literature. EN algorithm is intuitive and has simple implementation. However, the proposed estimator of $c$ is often biased unless the predictor distributions for cases and controls are completely separable, which in turn leads to biased estimation of $\mathbb{P}(Y = 1 \mid X)$ and prevalence $q$. As the estimated $c$ is biased toward 0, the estimate of $\mathbb{P}(Y = 1 \mid X)$ could exceed 1. In order to solve this problem, the proposed ML method fits model and estimates anchor sensitivity $c$ at the same time. We obtain the ML estimates by maximizing the log-likelihood function and use the inverse of the information matrix to establish the large sample variance-covariance matrix of these estimates. Finally, the phenotype prevalence $q$ is estimated as either

$$\hat{q} = \hat{h}/\hat{c} \qquad \text{or} \qquad \frac{1}{N}\sum_{i=1}^{n} \mathbb{P}(Y_i = 1 \mid \mathbf{X}_i, \hat{\beta})$$

According to the predictive accuracy metrics and estimates of anchor sensitivity and phenotype prevalence, the ML model seemed to consistently outperform the EN algorithm and naive logic. In addition, a remarkable feature of the ML method is its 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. With the anchor variable framework and ML method, innovative method for internally assessing model calibration and predictive accuracy using positive-only data, excludes the external model validation step. Comparing the work of validating a classically fit or transferred model, in order to generalize the method to secondary sites, chart review need only be completed to verify that the anchor has high PPV for the phenotype of interest. Phenotyping methods benefit from both noisy labels with random error and anchor variable structure.

| | ML | EN | Naive Logit | Ideal Learning |
|---|---|---|---|---|
| **Simulation Results** | | | | |
| Anchor sensitivity C | Identical to ideal learning | 0.37<br>ESE (0.04) | ? | Identical to ML |
| Phenotype prevalence q | Identical to ideal learning | 0.14<br>ESE (0.01) | ? | Identical to ML |
| AUC | 0.994 | 0.993 | 0.993 | 0.994 |
| 80% sensitivity (TPR) | PPV 86%<br>FPR 0.7% | PPV 84%<br>FPR 0.9% | PPV 83%<br>FPR 0.9% | |
| 80% PPV | Higher TPR than EN and Naive Logit | | | |
| | | 4% EN (ESE:0.5%) with PP > 1 (even after increase validation set size to 5000) | | |
| **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 PP > 1 | | |
| **A Preliminary Phenotyping Model for PA using Real-World Predictors and Cases** | | | | |
| **Case Set A** | | | | |
| C | 0.56 | 0.35 | ? | ? |
| q | 4% | 7% | ? | ? |
| 0.5 TPR | 0.66 | 0.59 | 0.28 | ? |
| | | 0.6% of EN PP > 1 | | |
| **Case Set B** | | | | |
| C | 0.62 | 0.41 | ? | ? |
| q | 5% | 8% | | |
| TPR | Higher than EN & Naive Logit | | | |
| | | 0.7% of EN PP > 1 | | |

Table A1: The results are convincing, as we can see in from the table. The ML method outperformed EN and naive logit based on predictive accuracy and estimates of c and q. In the first dataset, 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 naive logit. In the second dataset, ML method achieved a PPV (70% TPR) and FPRs that are identical to that of ideal learning and TPR (50% PPV) that is comparable to that of ideal learning. In the third dataset, for both case sets, the ML fitted model seems to have high discriminatory power with anchor-positive cases having high predicted probabilities (PP). The ML method achieved consistently higher TPR than that of the EN and naive logit.

# 2 References

Nogues, I.-E., Wen, J., Lin, Y., Liu, M., Tedeschi, S. K., Geva, A., Cai, T., & Hong, C. (2022). Weakly Semi-supervised phenotyping using Electronic Health records. Journal of Biomedical Informatics, 134, 104175–104175. https://doi.org/10.1016/j.jbi.2022.104175

Zhang, L., Ding, X., Ma, Y., Muthu, N., Ajmal, I., Moore, J. H., Herman, D. S., & Chen, J. (2020). A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients. Journal of the American Medical Informatics Association: JAMIA, 27(1), 119–126. https://doi.org/10.1093/jamia/ocz170

# 3 Figures

Logistic regression (logit $\mathbb{P}(Y = 1|X; \beta) = \beta_0 + \sum_{i=1}^{9}(\beta_k X_k)$)

- $X_1$, $X_2$, $X_3$: weak predictors - ($\beta_1$, $\beta_2$, $\beta_3$) = (0.2, 0.4, 0.6)
- $X_4$, $X_5$, $X_6$: moderate predictors - ($\beta_4$, $\beta_5$, $\beta_6$) = (-1.0, -1.4, 1.8)
- $X_7$, $X_8$, $X_9$: strong predictors - ($\beta_7$, $\beta_8$, $\beta_9$) = (-2.0, 2.4, 2.8)
- $X_1$, $X_4$, $X_7$ generated from normal distribution $N(5, 10)$
- $X_2$, $X_5$, $X_8$ generated from Bernoulli distribution Bernoulli(0.5)
- $X_3$, $X_6$, $X_9$ from logit transformed standard uniform distribution
- $\beta_0$ was changed to achieve phenotype prevalence at 5%, 10%, 15%, 20%

Figure 1: The logistic regression model developed to generate simulation data to test the maximum likelihood (ML) method.

| Model | Anchor Sensitivity | Phenotype Prevalence |
|---|---|---|
| True simulation values | 0.50 | 0.10 |
| ML Method | 0.50 (0.021) | 0.10 (0.004) |
| Ideal learning | No anchor | 0.10 (0.0003) |
| EN algorithm | 0.37 (0.04) | 0.14 (0.01) |
| Naive logit model | No anchor | 0.05 (0.002) |

Table 1: Comparison of mean anchor sensitivity and phenotype prevalence estimates with standard errors for the ML method, EN algorithm, naive logit and ideal learning models fitted to 1000 iterations of the simulated EHR data.

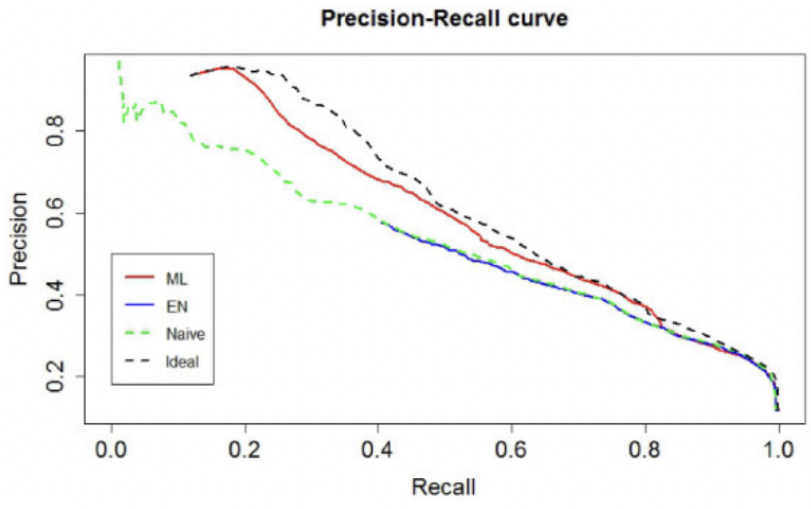


Figure 2: Precision-recall curve of the method validation data set results for the ML method, EN algorithm, naive logit model and ideal learning models from the original paper (Zhang et al, 2020).

| Model | Simulation | Method Validation | Case set A | Case 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% |

Table 2: Comparison of true positive rates (TPR) at the respective thresholds for the ML method, EN algorithm, naive logit and ideal learning models fitted to simulation (80% PPV), method validation (50% PPV), and real-world anchors (50% PPV) data sets.

# 4 Contributions

- Introduction and Motivation — Nigel Petersen
- Methods — Nigel Petersen
- Results — Shruthi Vaidyanathan
- Limitations — Chelsea Murphy
- Reflection — Shruthi Vaidyanathan
- Conclusion — Chelsea Murphy