Maximum Likelihood with Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation

Amr M. Alexandari*, Anshul Kundaje†, **Avanti Shrikumar***†

*co-first authors †co-corresponding authors

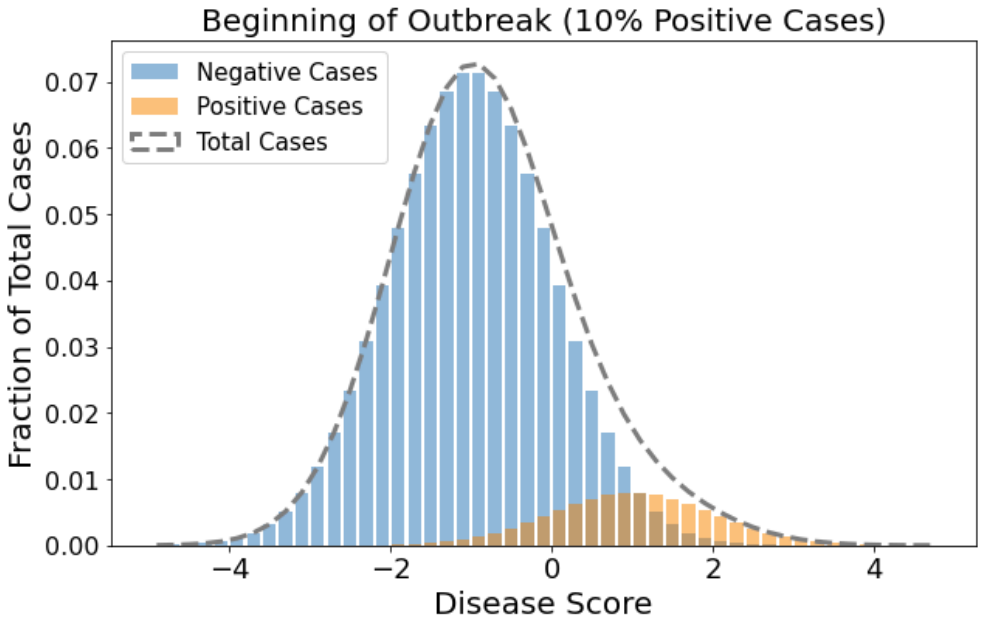
Amr Alexandari
PhD Student
Dept. of Computer Science

Anshul Kundaje
Assistant Professor
Depts. of CS & Genetics

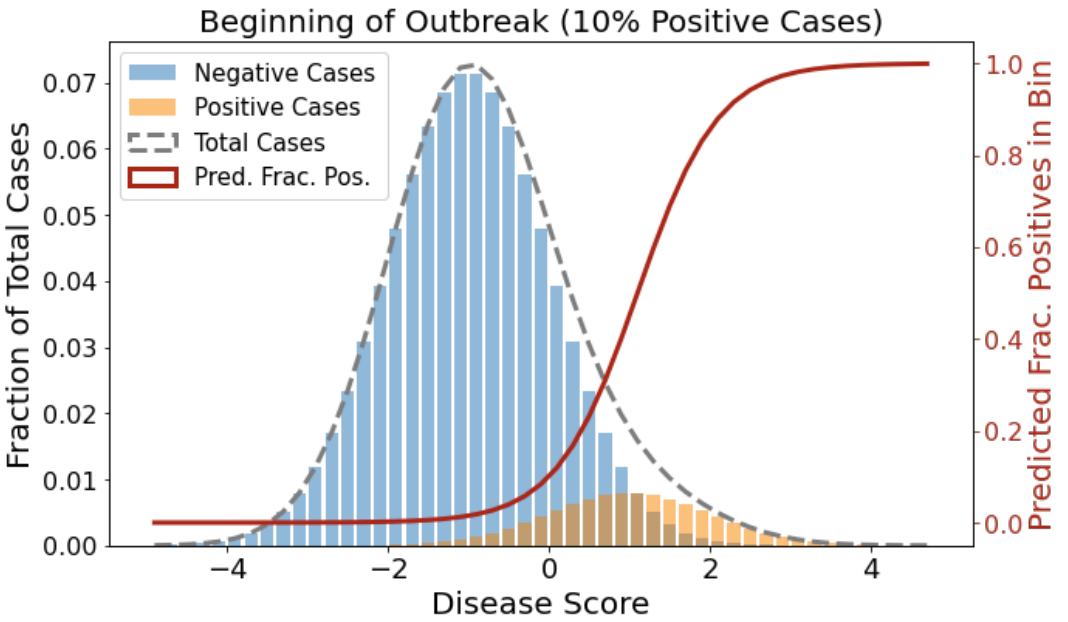
Stanford University

ICML
International Conference On Machine Learning

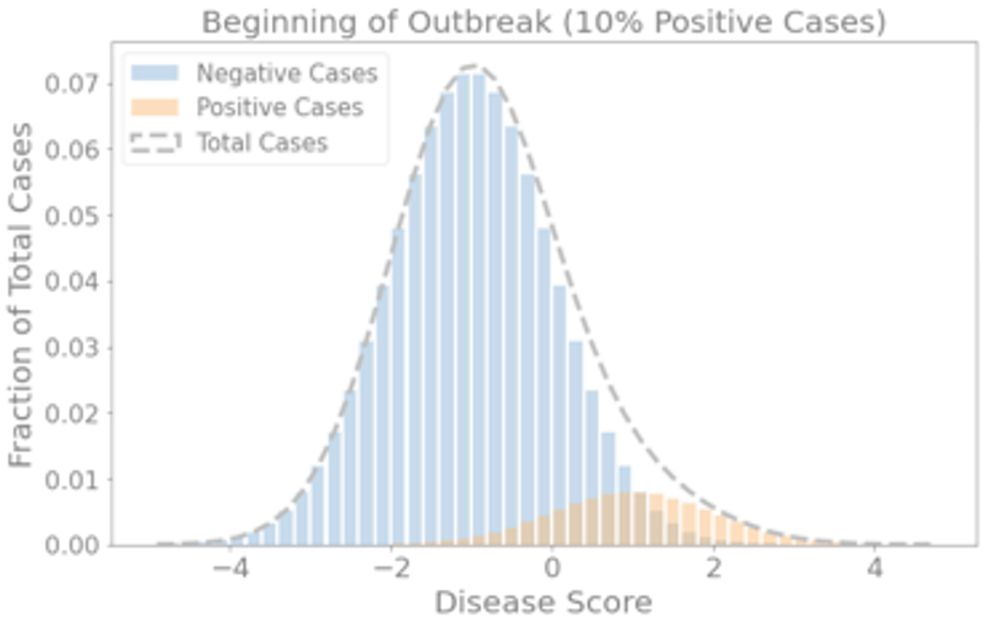
Label Shift Illustrated



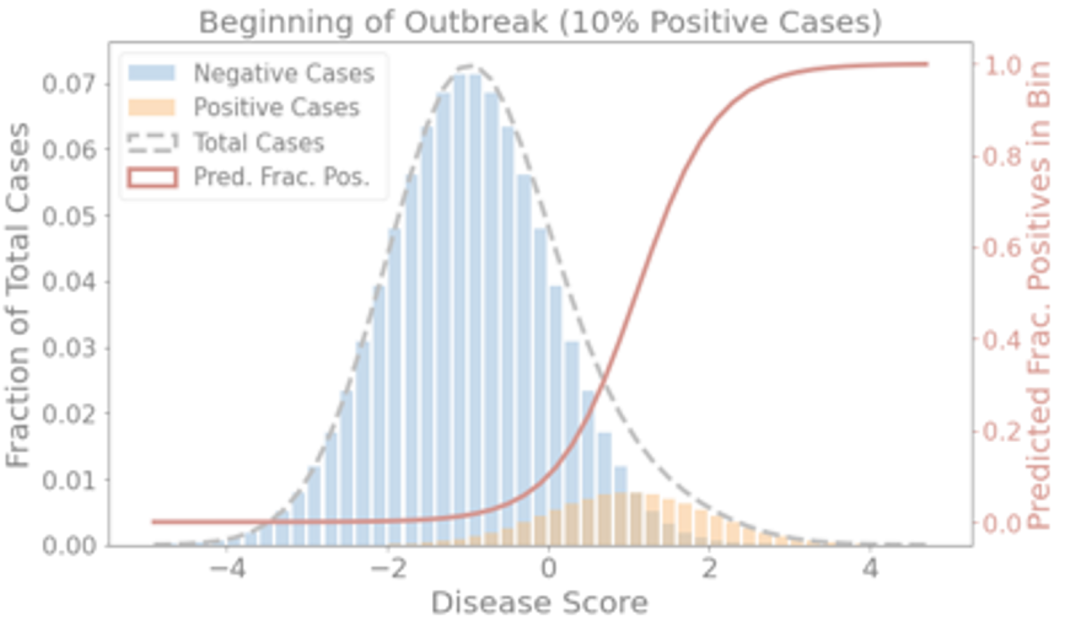
Train Model

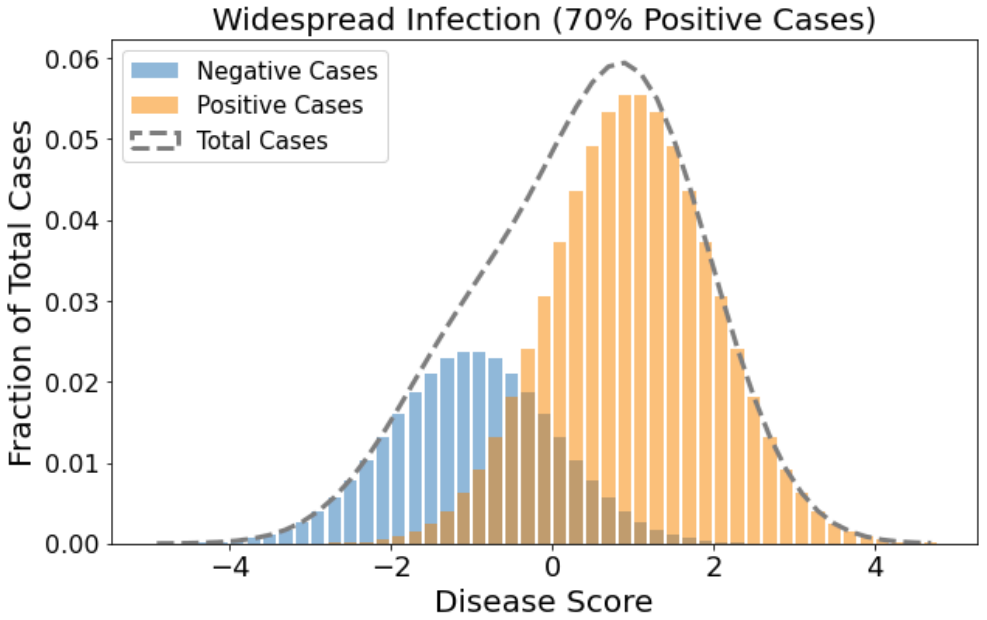


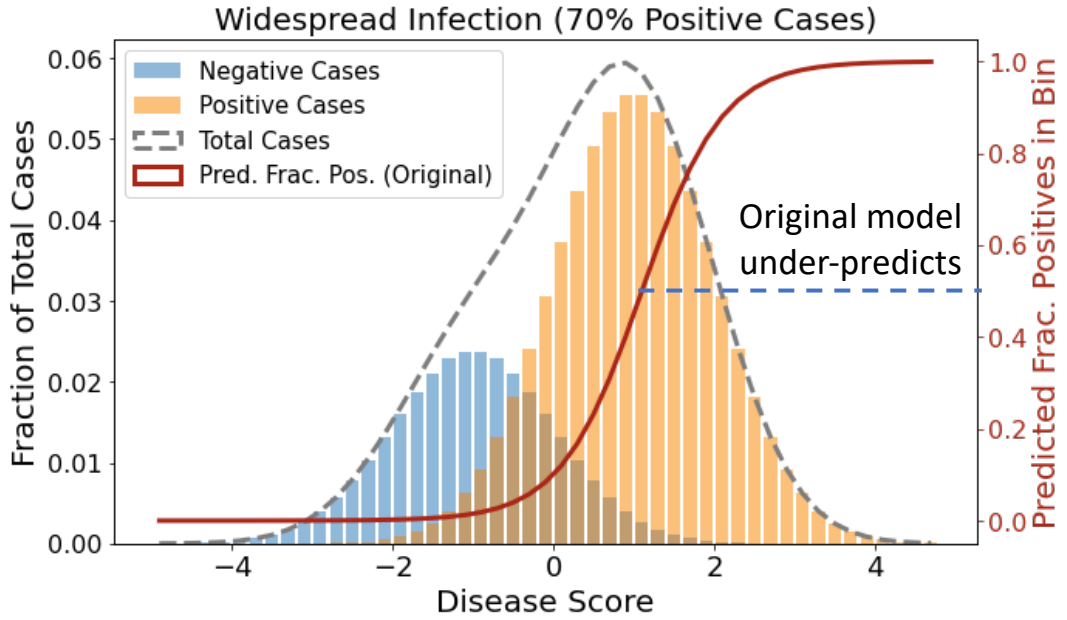
Label Shift Illustrated



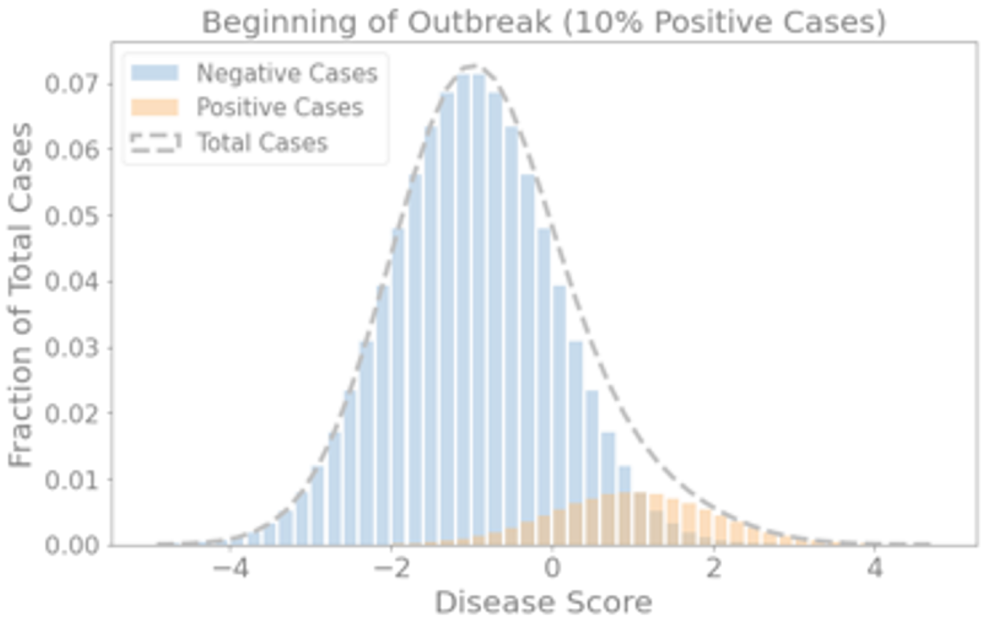
Train Model



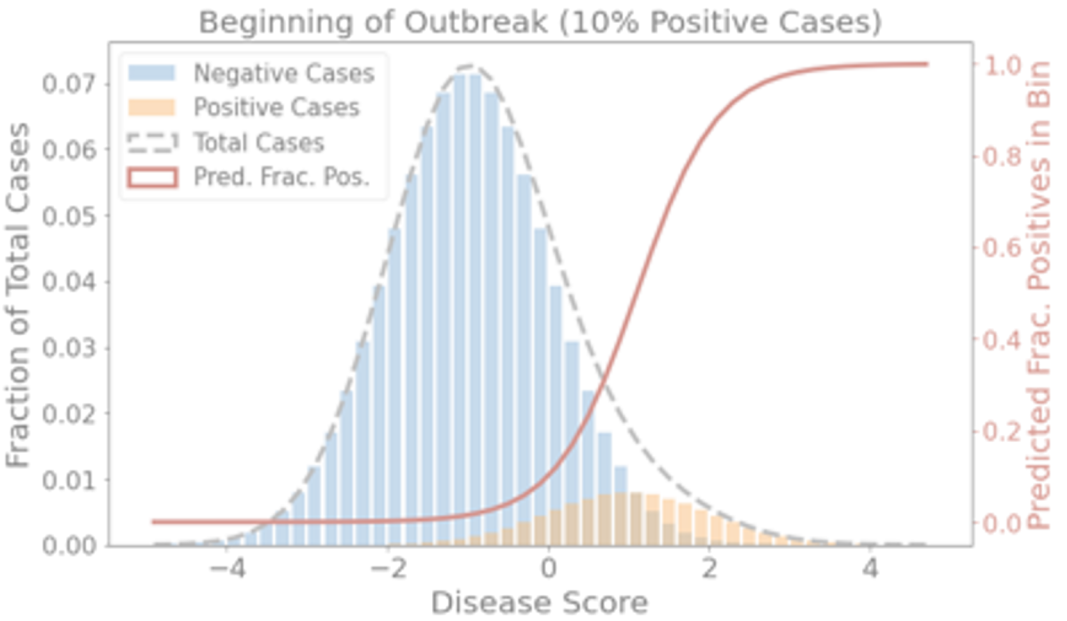


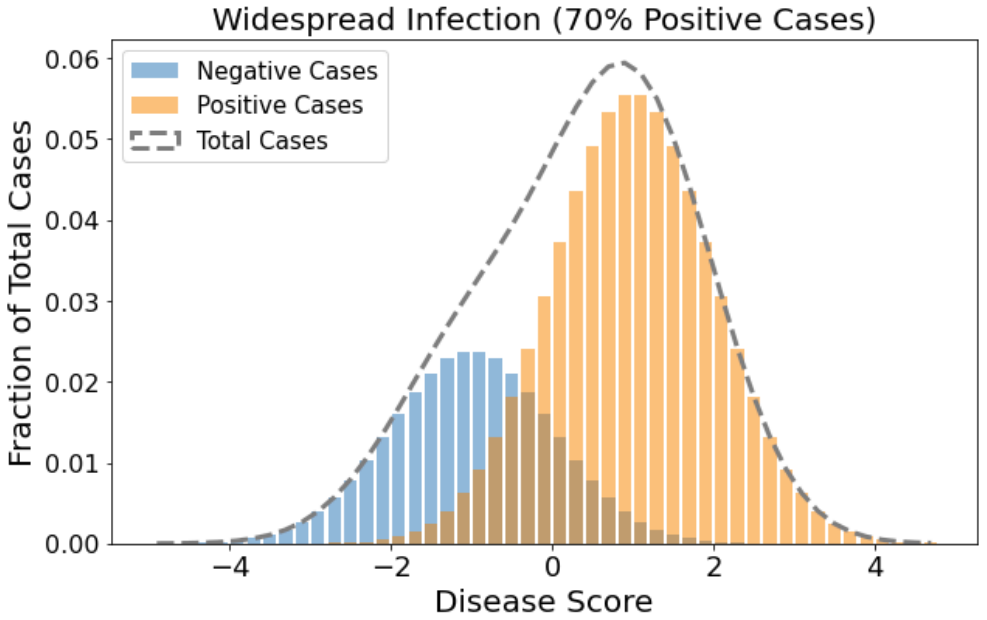


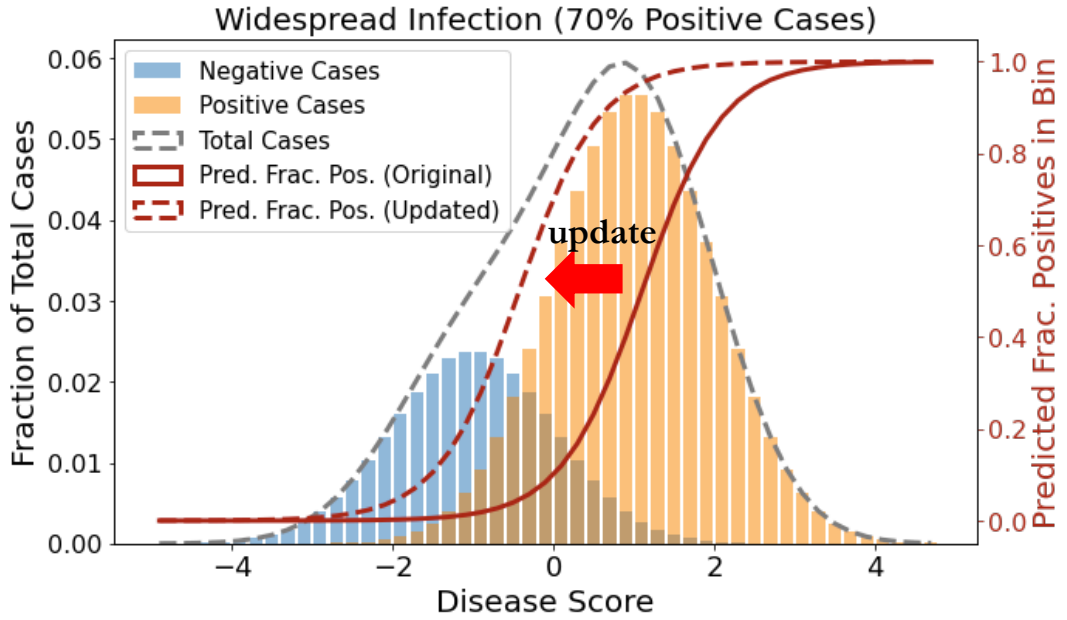
Label Shift Illustrated



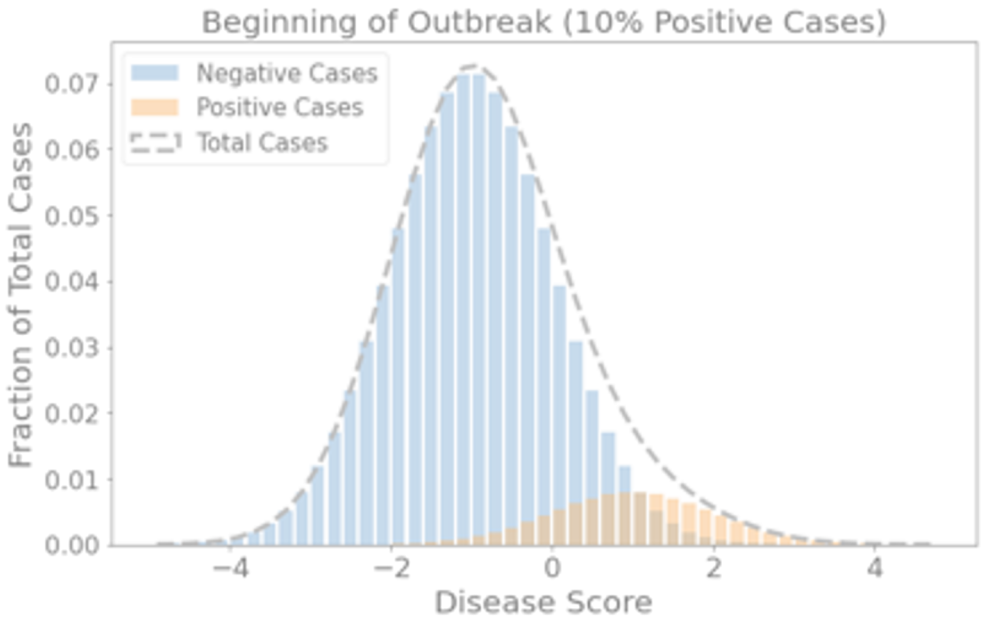
Train Model

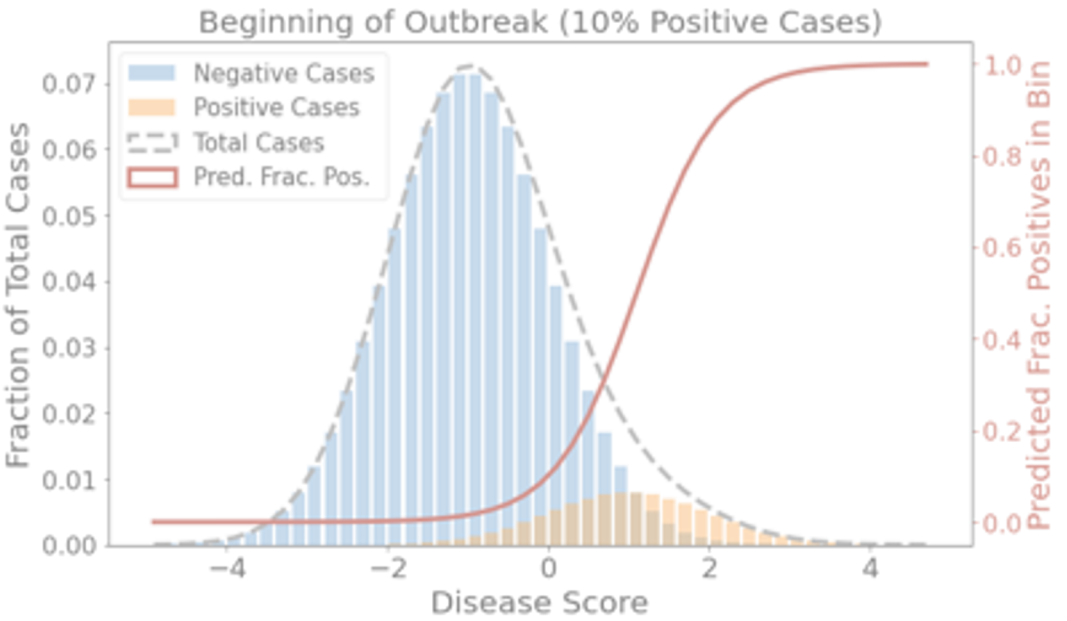


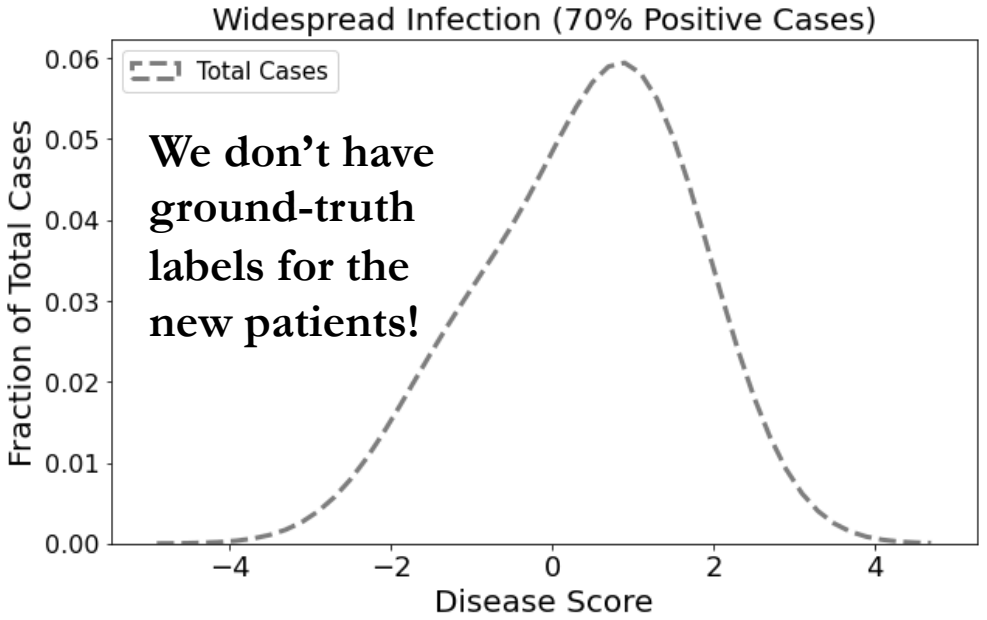


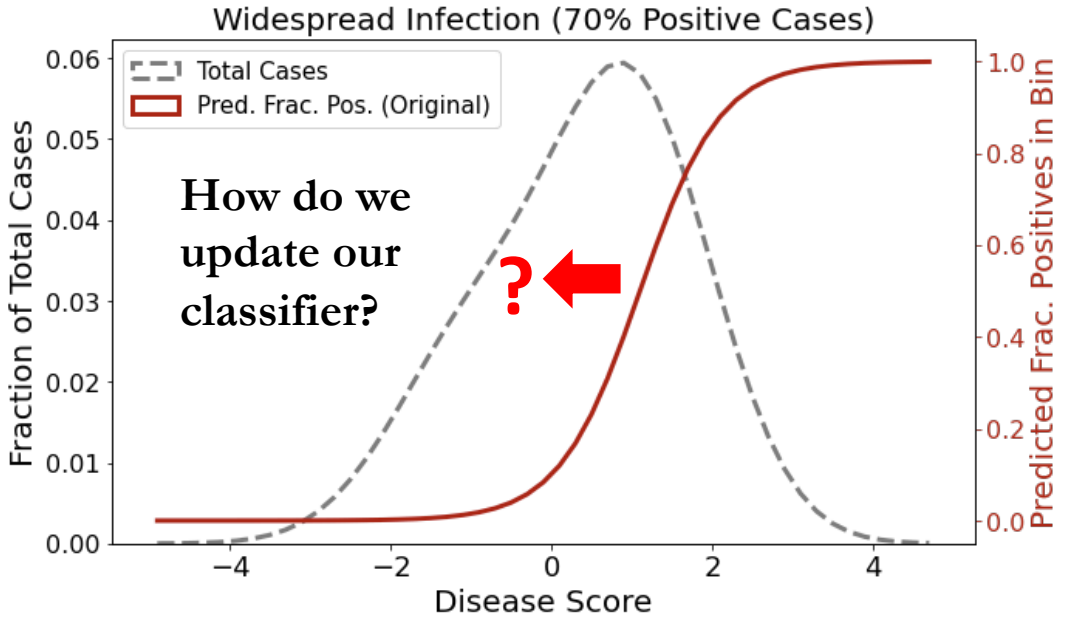


Label Shift Illustrated









Main Contributions

- An approach that achieves **state-of-the-art** on label shift adaptation
  - Scales to datasets with high-dimensional inputs
  - Does not require model retraining
- Combines Max Likelihood with **specific** types of calibration.
  - Calibration with Temp. Scaling (TS) was insufficient (& sometimes harmful!)
  - Achieved state-of-the-art with extensions of TS (one of which we propose) that correct for systematic bias

Formal Definition of Label Shift

Let:

- y denote our labels (whether or not person has disease)
- $\boldsymbol{x}$ denote the observed symptoms
- $p(\boldsymbol{x}, y)$ denote joint distribution $(\boldsymbol{x}, y)$ at beginning of outbreak ("source domain")
- $q(\boldsymbol{x}, y)$ denote joint distribution at widespread stage ("target domain"), when we don't know labels
- Goal: adapt source-domain classifier that predicts $p(y|\boldsymbol{x})$ to instead predict $q(y|\boldsymbol{x})$ for target domain

Core assumption: disease has same symptoms irrespective of outbreak stage, i.e. $p(\boldsymbol{x}|y) = q(\boldsymbol{x}|y)$.

- Thus, difference between source & target domain is **exclusively** caused by shift in label proportions $p(y)$ and $q(y)$**.** Formally, $q(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)q(y)$
- Also called **prior probability shift** (Amos, 2008), corresponds to "anti-causal learning" i.e. predicting cause y from effects $\boldsymbol{x}$ (Schloelkopf, 2012).
- Anti-causal learning is appropriate here because diseases status y cause the symptoms $\boldsymbol{x}$.

Estimating $q(y|\boldsymbol{x})$ with Bayes' Rule

- Although $p(\boldsymbol{x}|y)$ is preserved, computing it **is hard when $\boldsymbol{x}$ is high-dimensional.**
- Much easier to estimate $p(y|\boldsymbol{x})$ and $p(y)$ from the source domain, as y is lower-dimensional.
- If we know $q(y)$, we can retrieve $q(y|x)$ **without ever estimating $p(\boldsymbol{x}|y)$** using Bayes' Rule (first shown in Saerens et al., 2002):

We first write $q(y|\boldsymbol{x}) = \frac{q(y,\boldsymbol{x})}{q(\boldsymbol{x})} = \frac{q(\boldsymbol{x}|y)q(y)}{\sum_{y^*} q(\boldsymbol{x}|y^*)q(y^*)}$ (terms in red are not explicitly known)

Substituting $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$ (label shift assumption), we have $q(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)q(y)}{\sum_{y^*} p(\boldsymbol{x}|y^*)q(y^*)}$

Through Bayes' rule, observe that $p(\boldsymbol{x}|y) = \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}$

Substituting, we get $q(y|\boldsymbol{x}) = \frac{\frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}q(y)}{\sum_y \frac{p(y|\boldsymbol{x})p(\boldsymbol{x})}{p(y)}q(y)}$

$p(\boldsymbol{x})$ cancels out, giving $q(y|\boldsymbol{x}) = \frac{\frac{p(y|\boldsymbol{x})}{p(y)}q(y)}{\sum_y \frac{p(y|\boldsymbol{x})}{p(y)}q(y)}$

Reminders:
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$

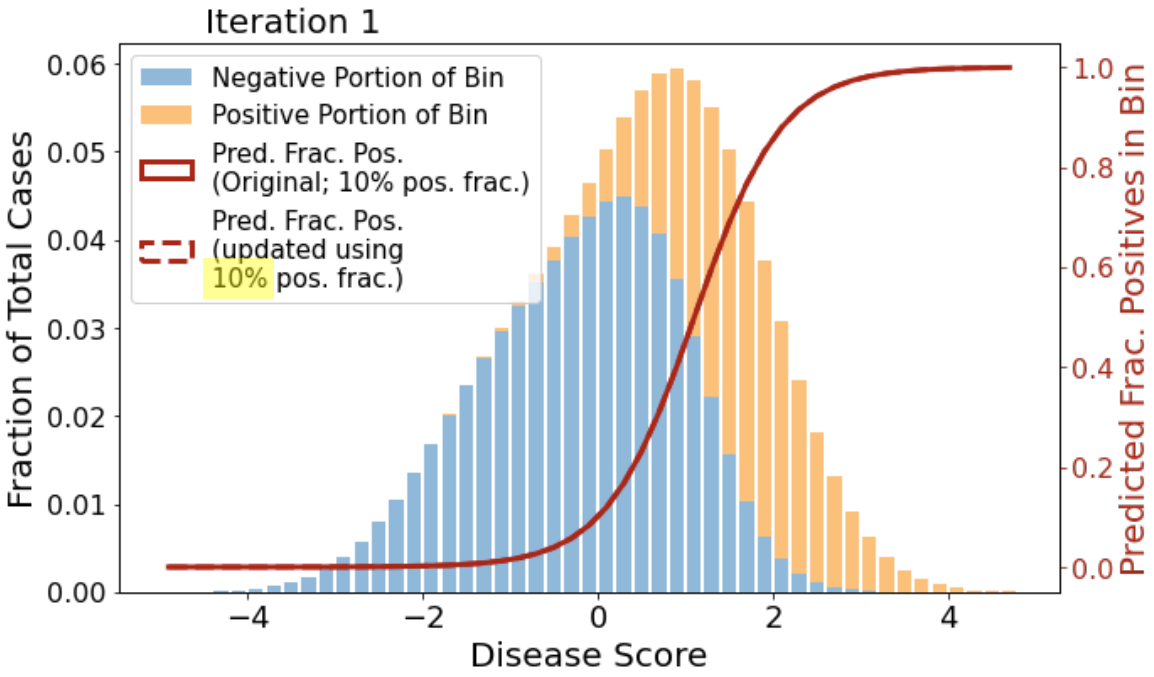
Reminders:

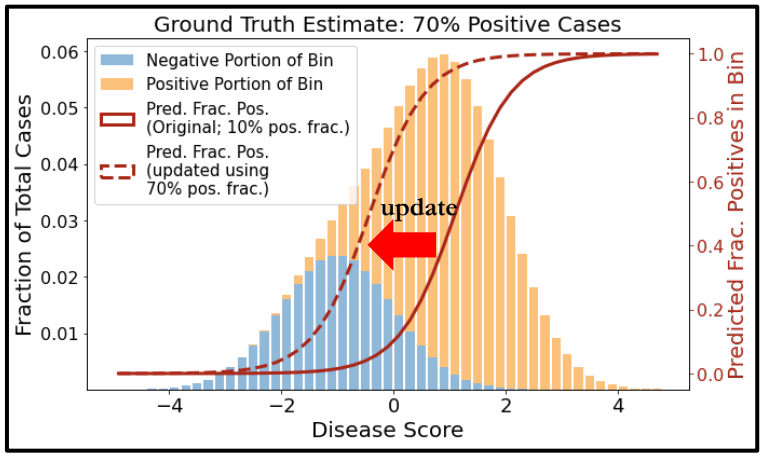
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





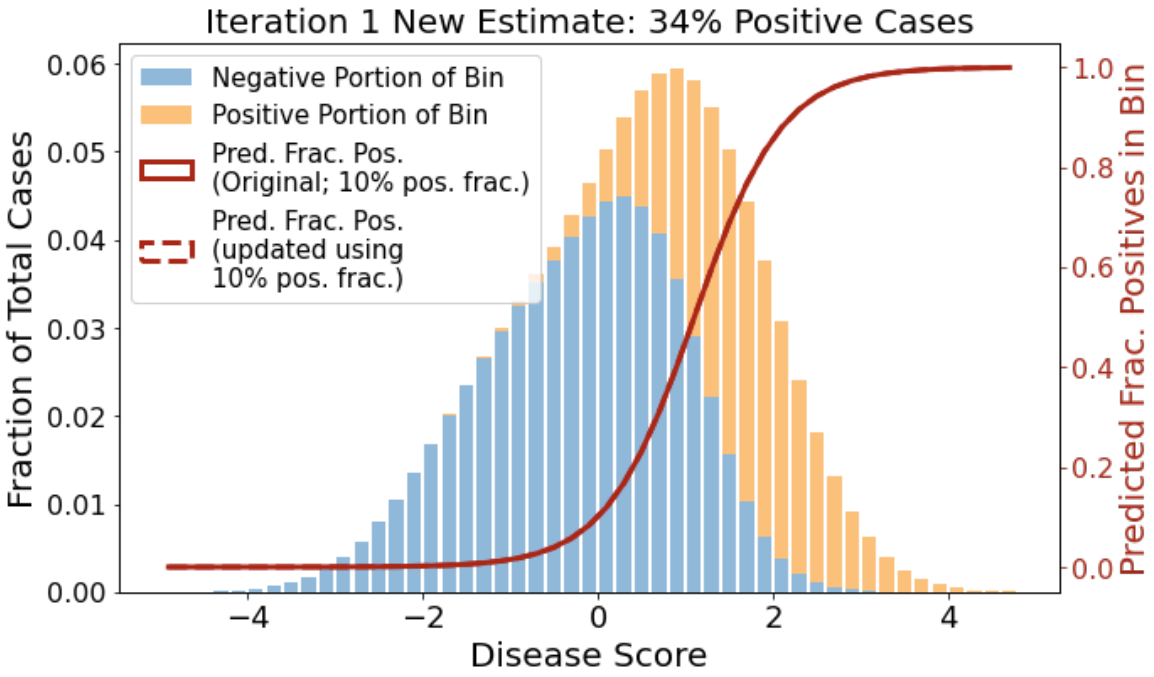
Reminders:

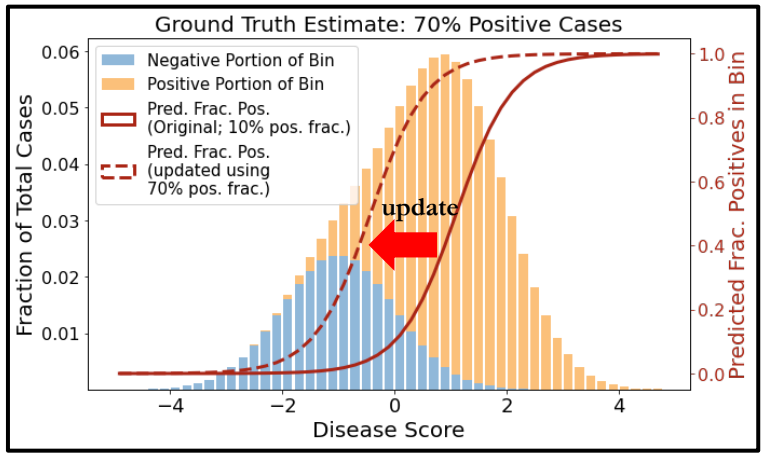
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





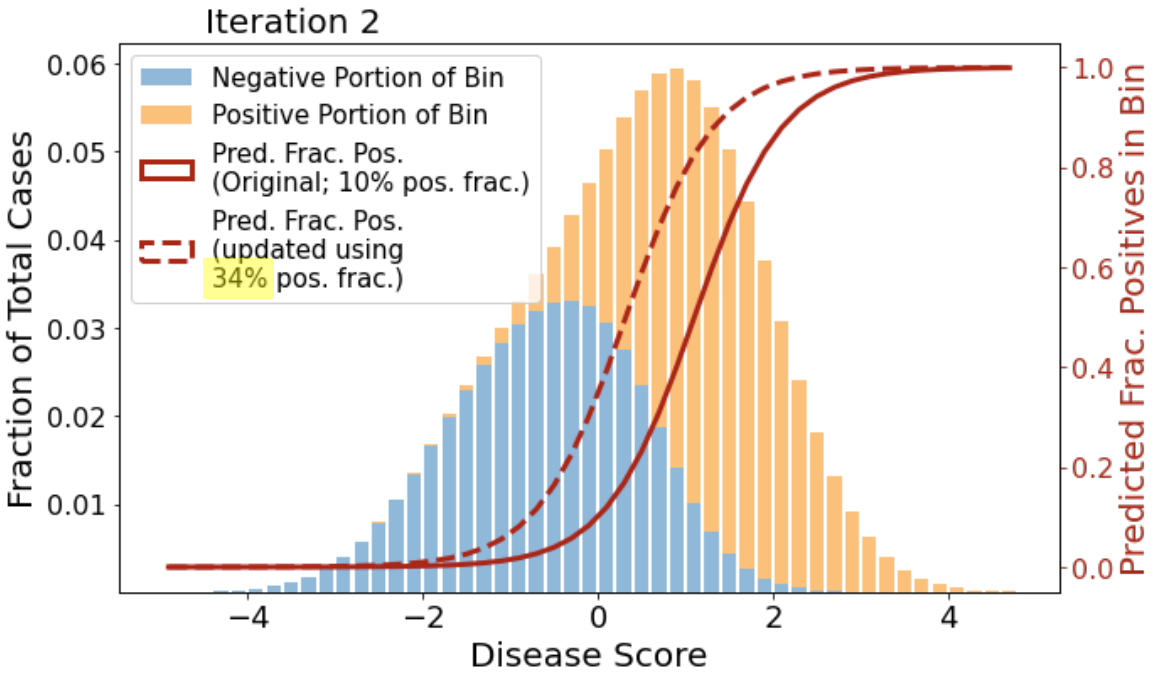
Reminders:

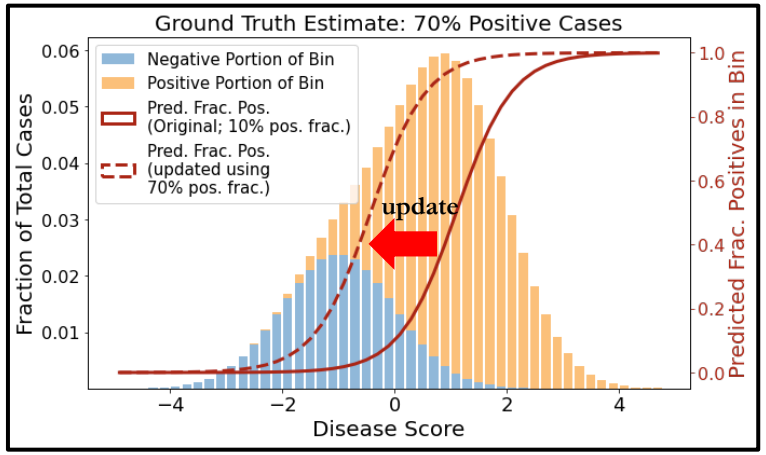
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





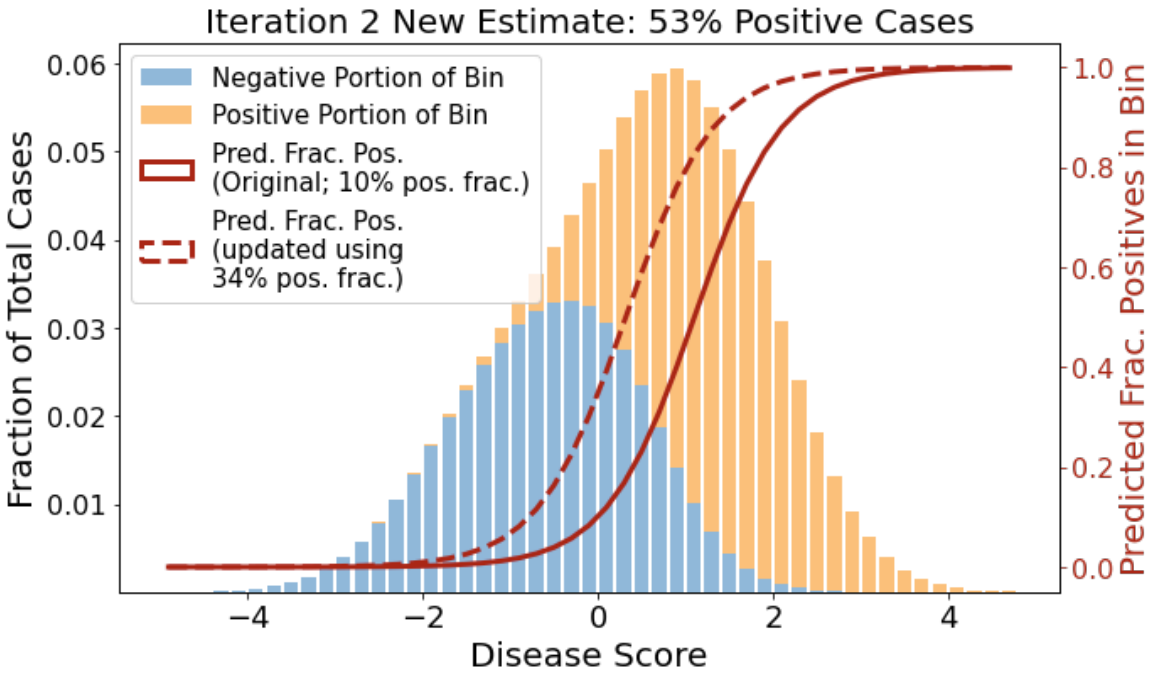
Reminders:

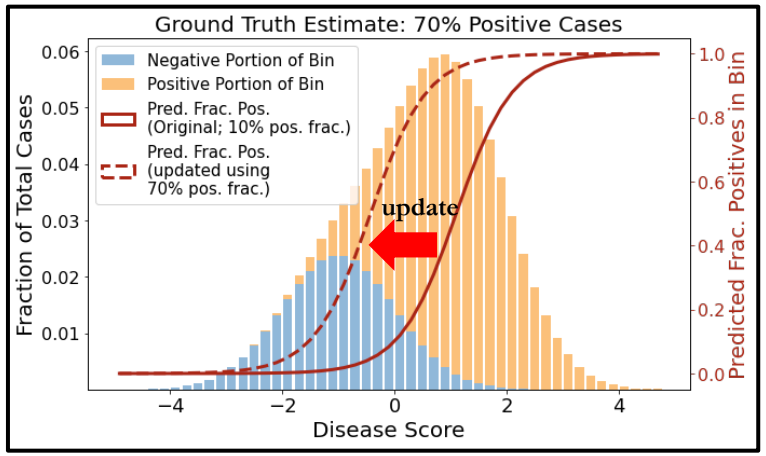
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





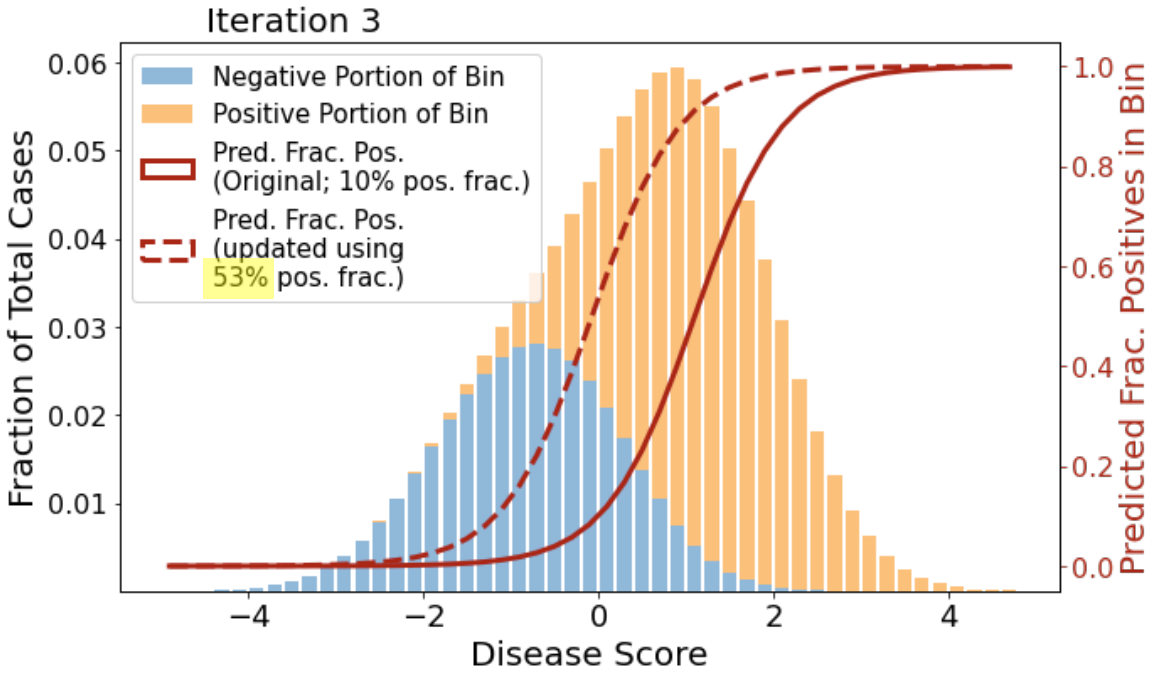
Reminders:

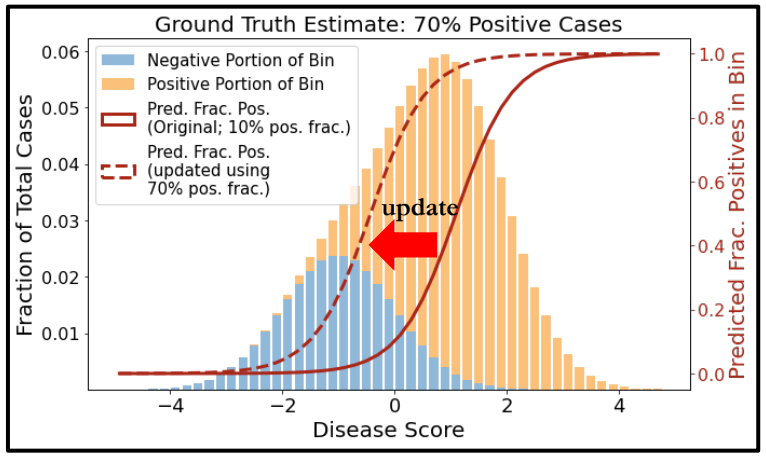
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





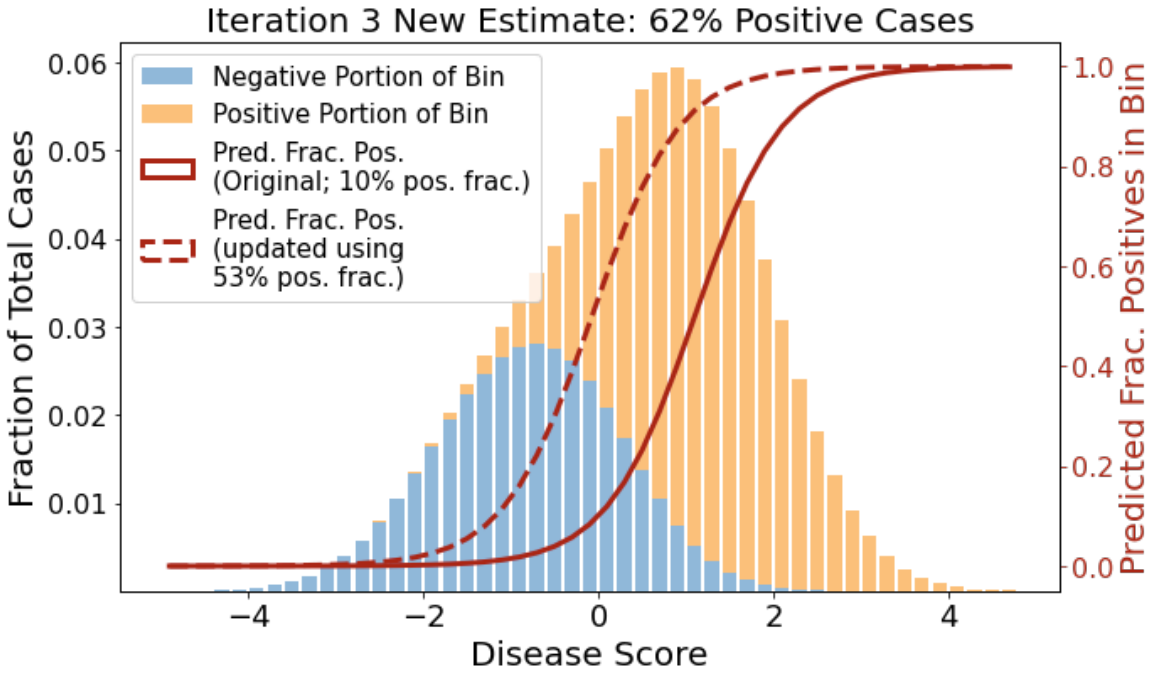
Reminders:

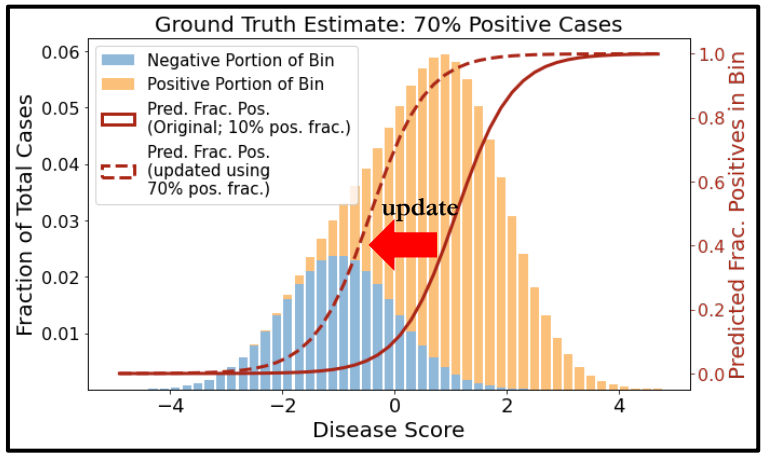
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





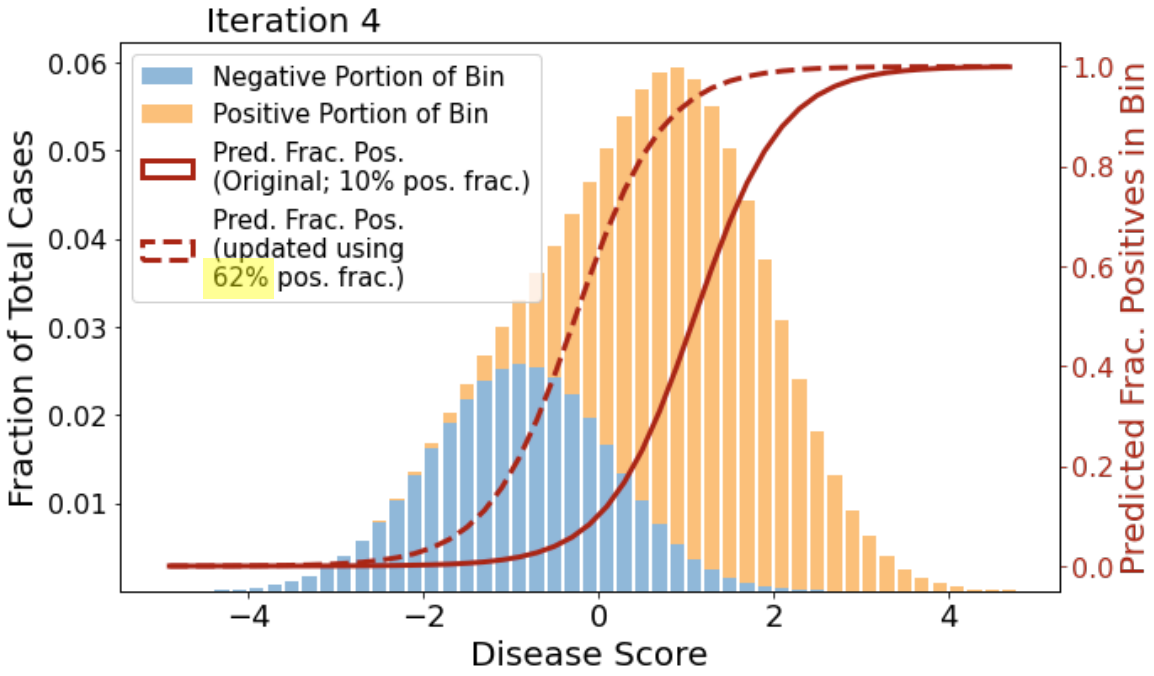
Reminders:

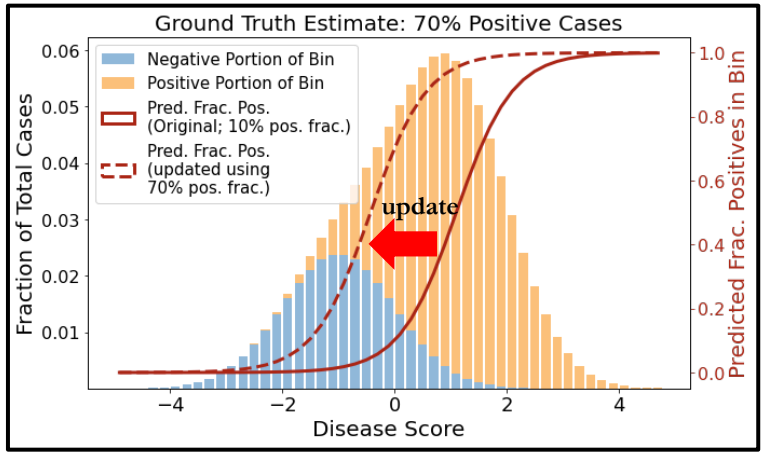
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





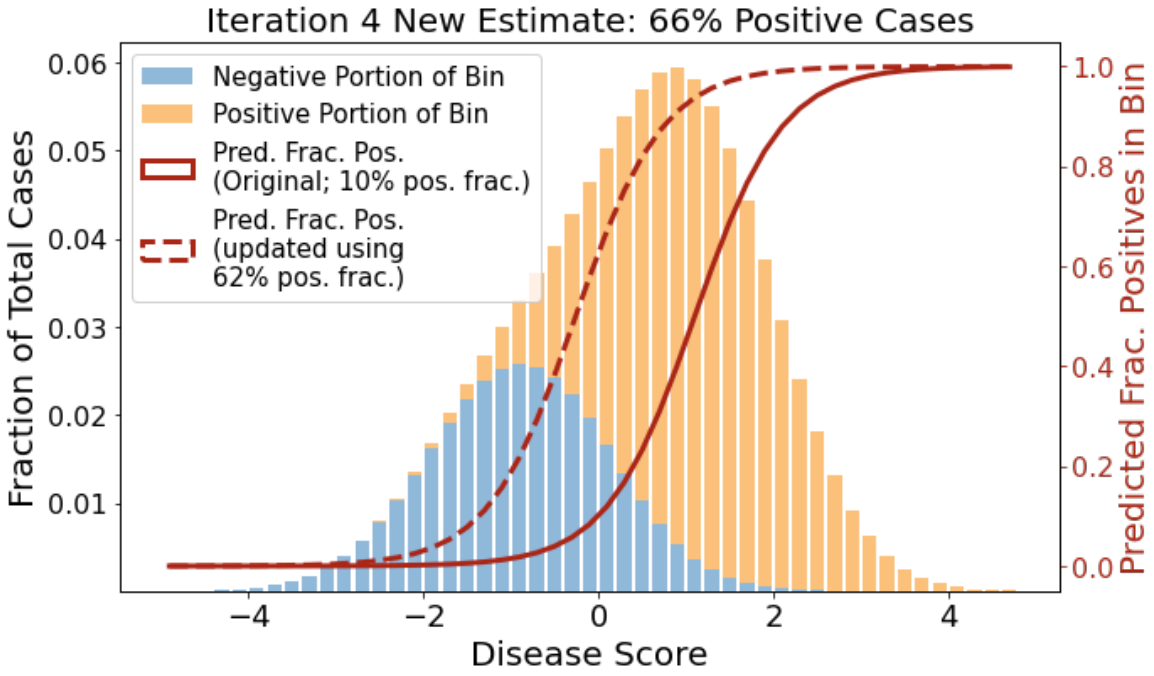
Reminders:

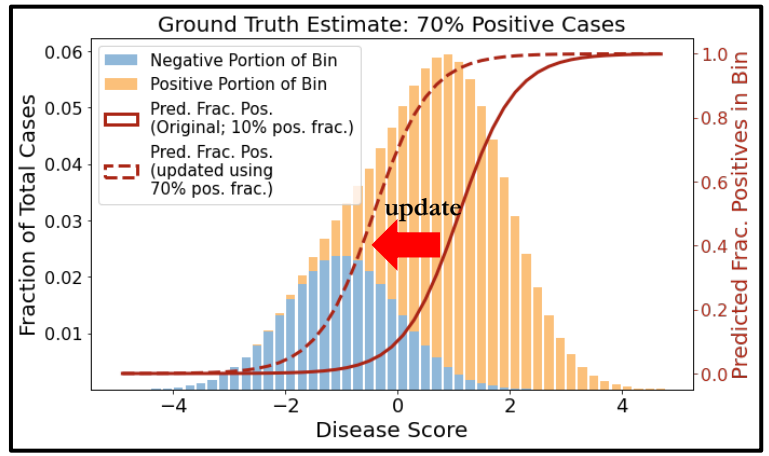
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





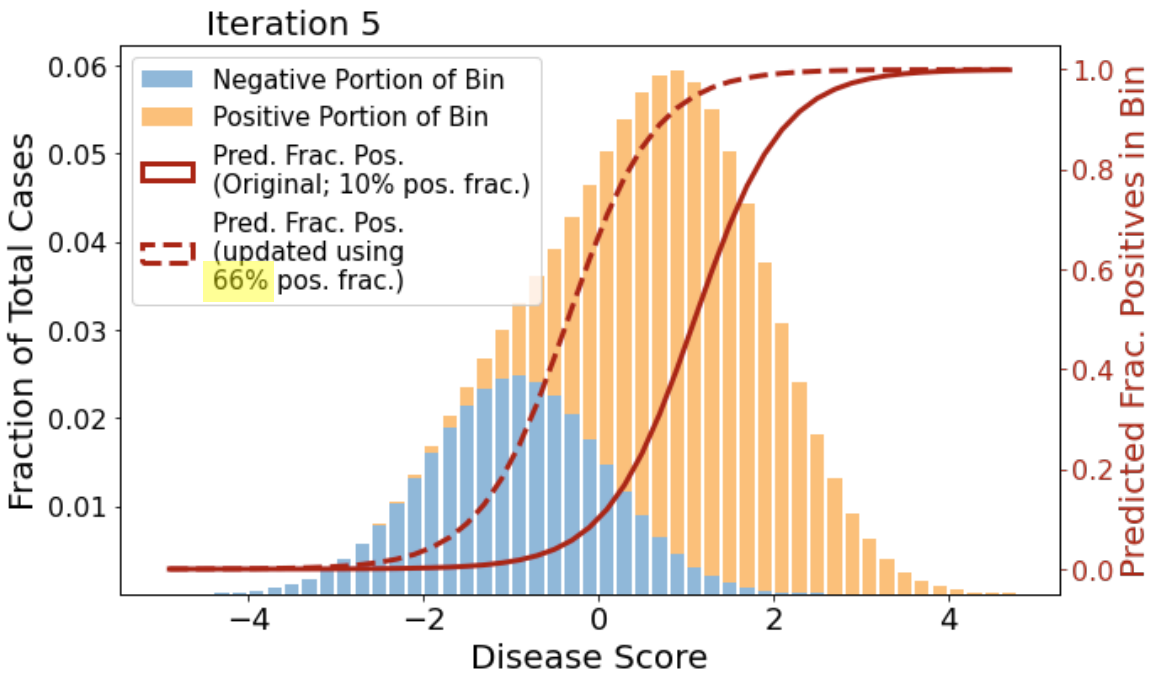
Reminders:

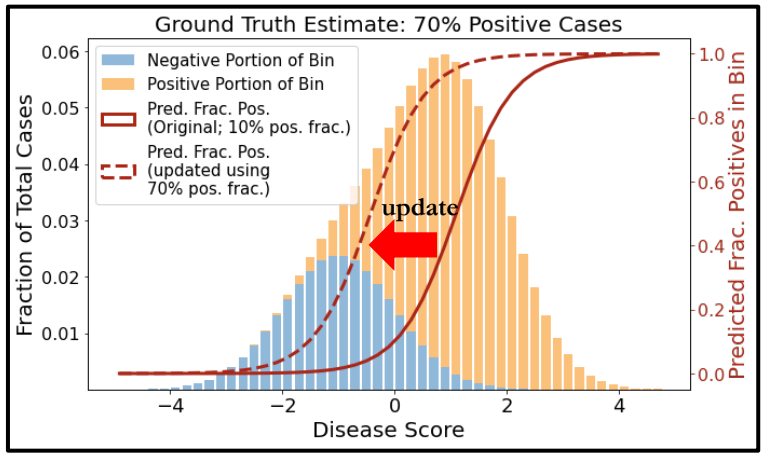
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





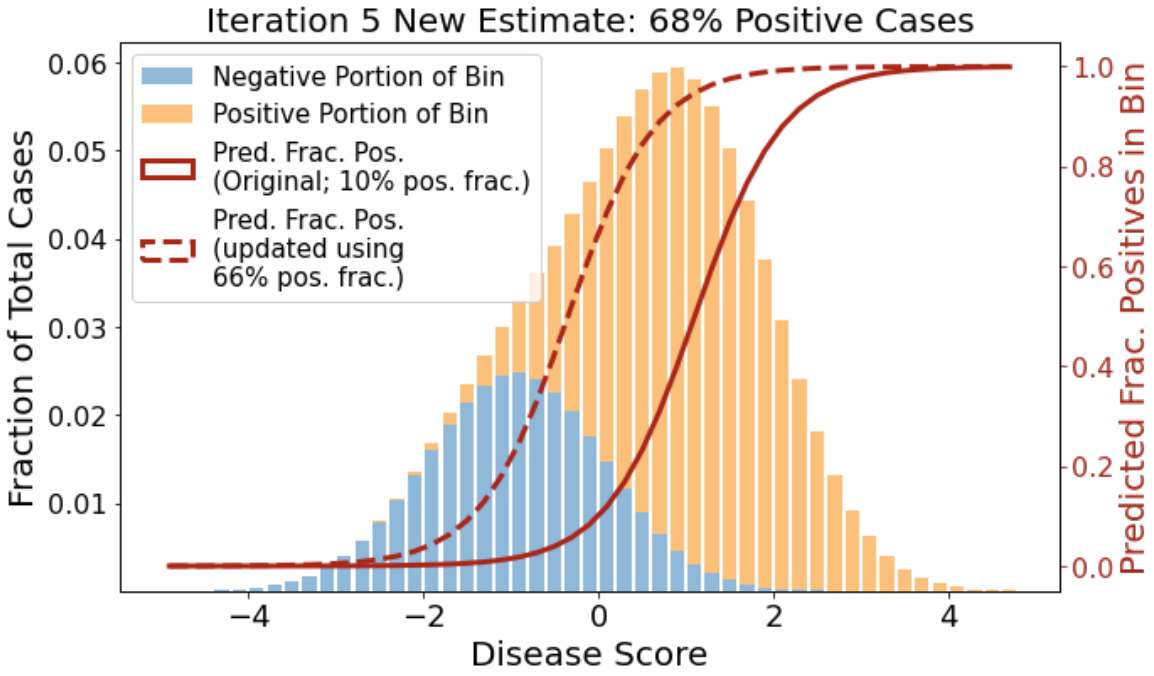
Reminders:

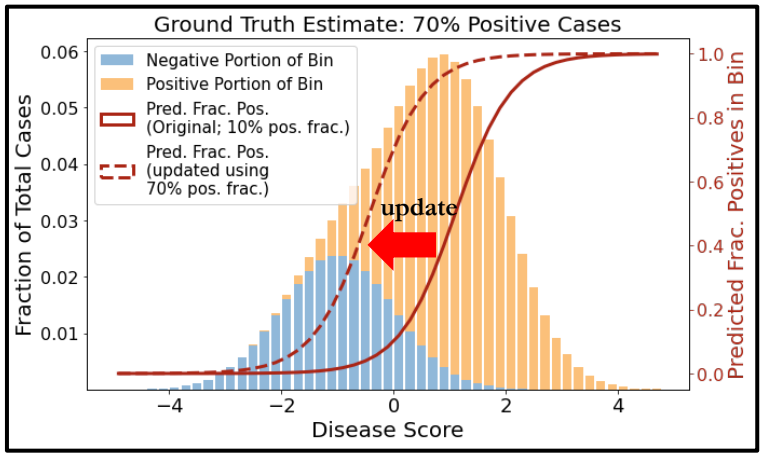
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





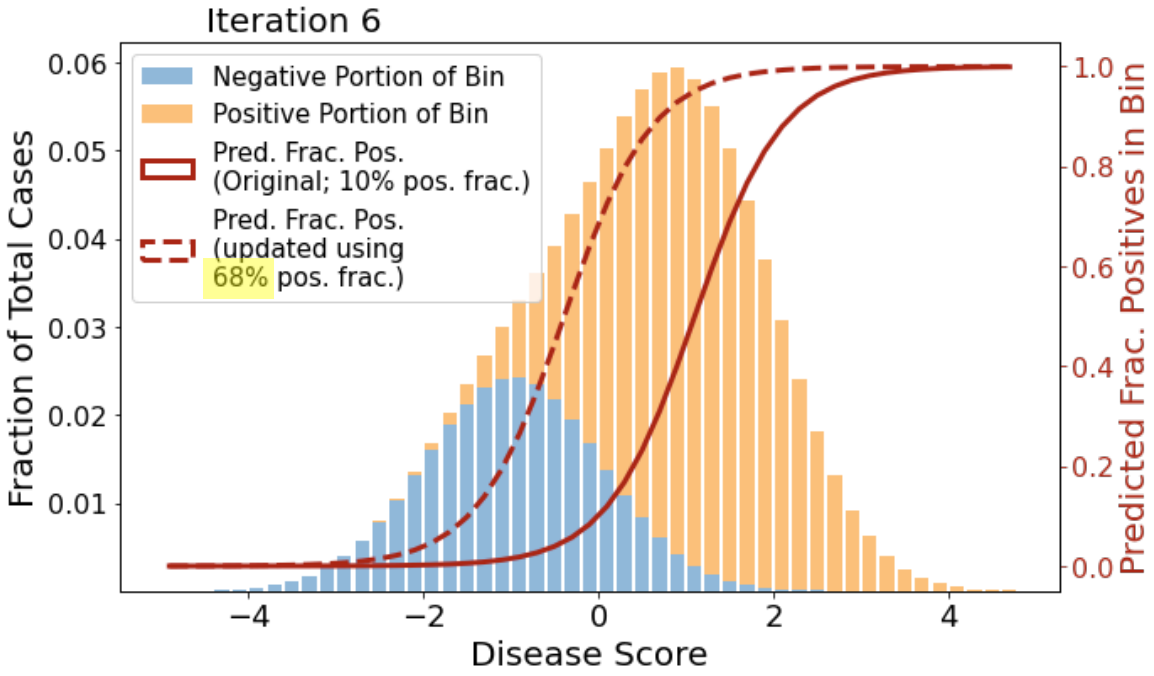
Reminders:

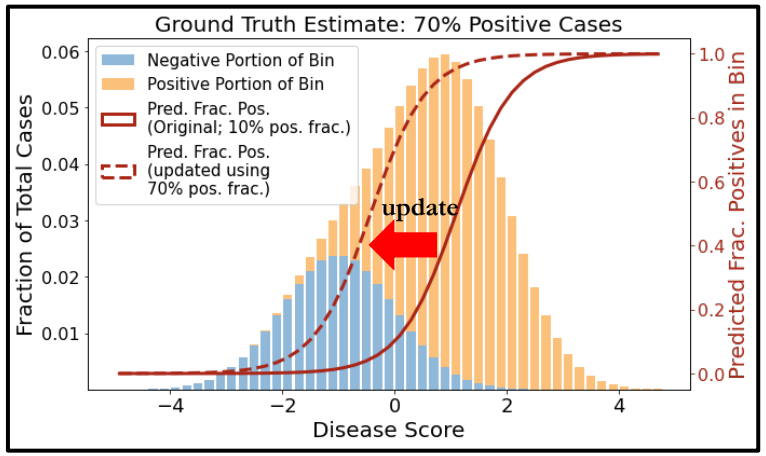
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





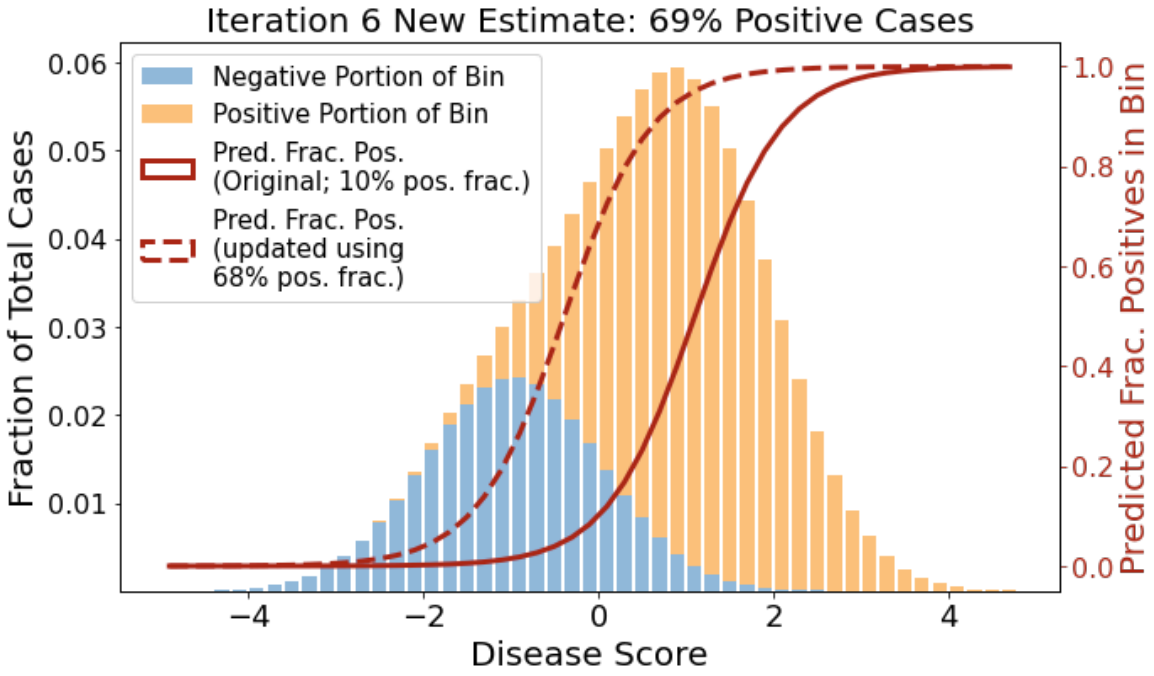
Reminders:

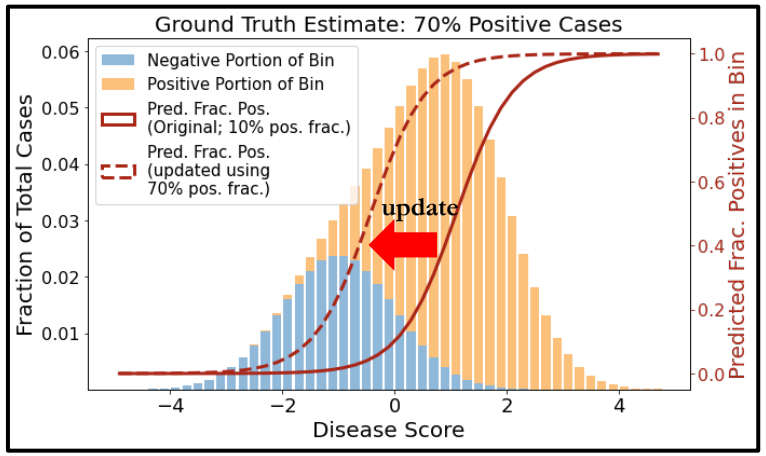
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





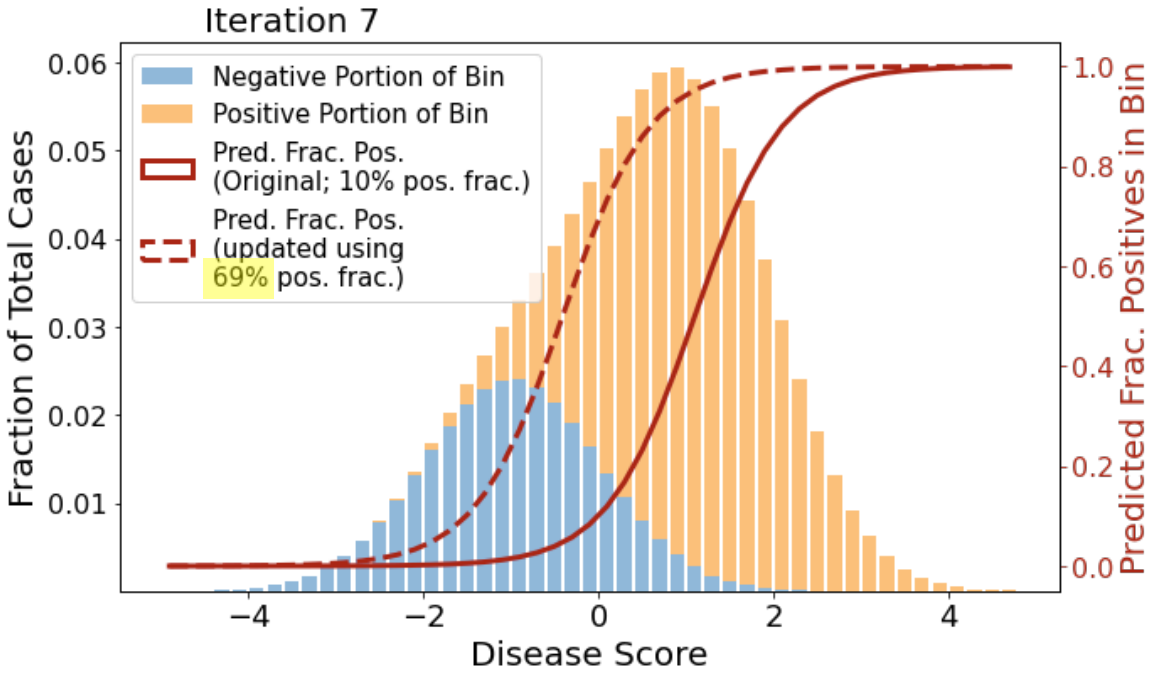
Reminders:

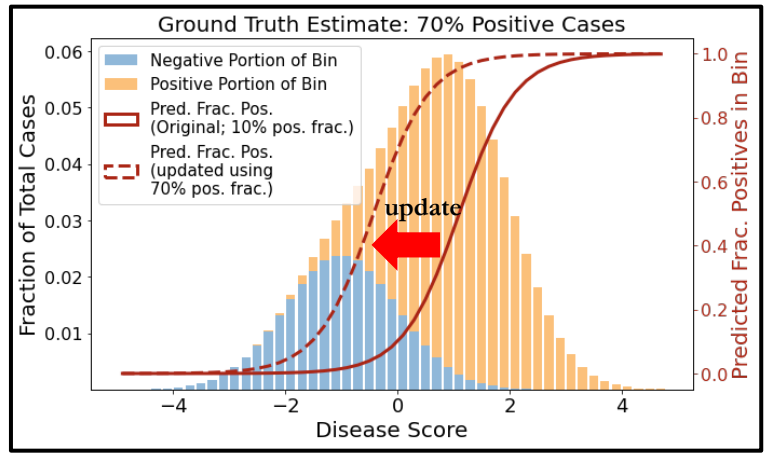
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





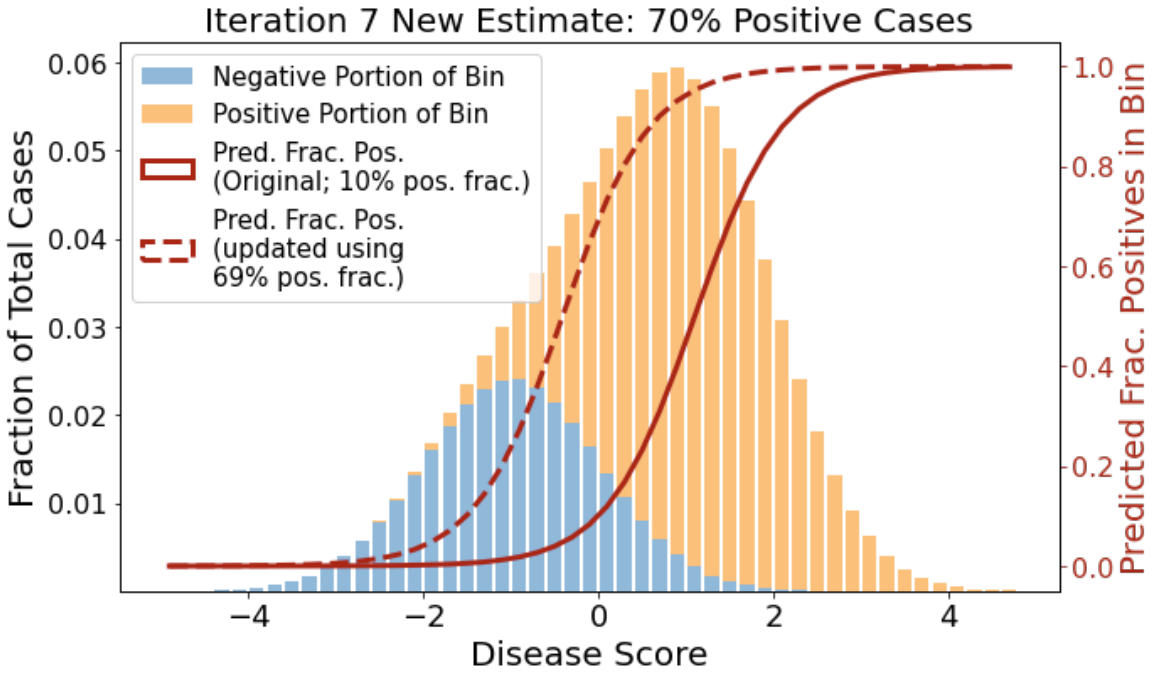
Reminders:

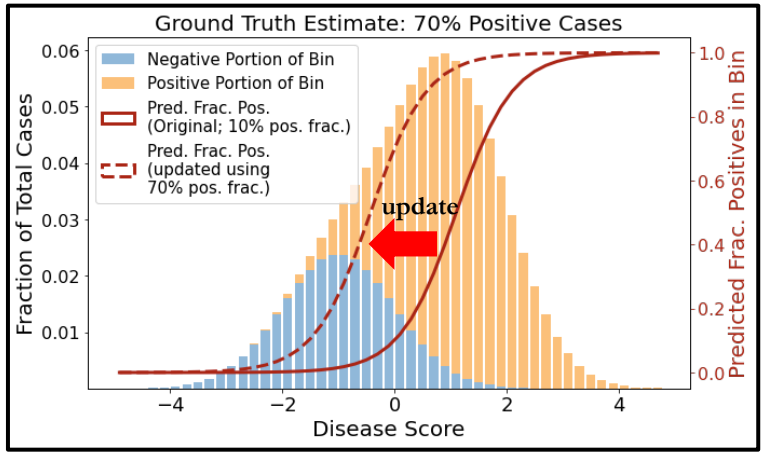
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

A Simple Iterative Approach to Label Shift…

In practice, we are not told $q(y)$ – how can we estimate it?

- Could use $p(y|\boldsymbol{x})$ to predict on test set & average predictions to estimate $q(y)$
- Could then use $q(y)$ to update $p(y|\boldsymbol{x})$, and repeat the process until convergence!





Reminders:

- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

Iterative approach ←→ Maximum Likelihood

- The simple iterative approach is a valid EM algorithm that optimizes the log likelihood $\sum_k \log \sum_y q(\boldsymbol{x}_k|y)q(y)$ w.r.t. parameters $\mathrm{q(y)}$. First shown in Saerens et al. (2002).
- Note: Saerens et al. (2002) has been **<u>incorrectly described</u>** in several recent papers as being unable to scale to high-dimensional $\boldsymbol{x}$ because it requires estimating $p(\boldsymbol{x}|y)$. The algorithm **only requires $p(y|\boldsymbol{x})$ and $p(y)$**, and thus scales to high-dimensional $\boldsymbol{x}$.
- In our paper, we further showed the **optimization is concave**; thus, EM converges to the global optimum , and one can use **any convex optimizer** for Max. Likelihood

Reminders:
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule

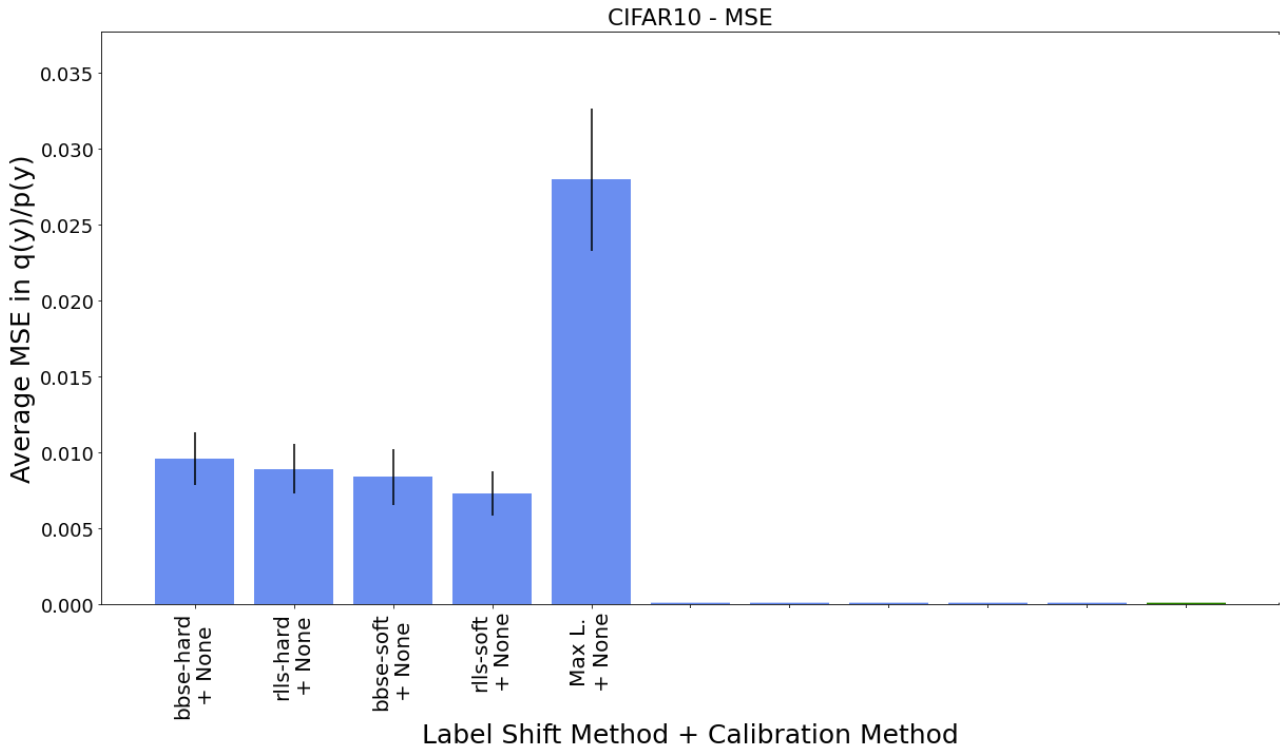
Recent Work on Label Shift Adaptation

- Prior work (Lipton et al., ICML 2018) proposed Black Box Shift Estimation (BBSE) to estimate $q(y)/p(y)$. BBSE builds a confusion matrix using held-out data & does not assume the predicted $p(y|\boldsymbol{x})$ are calibrated.
- Azizzadenesheli et al., ICLR 2019 improved on BBSE with Regularized Learning under Label Shifts (RLLS). Also leverages a confusion matrix built on held-out data.
- **Major drawback**: both BBSE and RLLS require **model retraining** using $q(y)/p(y)$ as the importance weights. Importance weighting does not work as well as expected with deep neural networks (Byrd & Lipton, 2019)
- Neither BBSE nor RLLS were benchmarked against Max Likelihood (which does not require retraining)

Reminders:
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule
- Given accurate $p(y|\boldsymbol{x})$, $p(y)$, we can find $q(y)$ through Maximum Likelihood (including EM)

CIFAR10 benchmarking

- Evaluation metric: mean squared error in estimate of $q(y)/p(y)$
- Dirichlet shift ($\alpha = 0.1$) simulated over 10 trials for each of 10 different trained models (100 trials in total). N=2000 samples were used in validation & test sets (results are qualitatively similar for different α and N as well).



Problem: Miscalibration

- Bayes' rule for deriving $q(y|\boldsymbol{x})$ given $q(y)$ assumes we have accurate $p(y|\boldsymbol{x})$**. In practice, this is often not the case because** $p(y|\boldsymbol{x})$ **from modern neural network is typically mis-calibrated** (Guo et al., 2017)
- (Loosely) calibration means: if model says $p(disease|\boldsymbol{x}) = 0.5$, then there is actually a 50% chance that the person has the disease
- Even when modern neural networks rank the predictions correctly, the probabilities themselves may be very inaccurate (e.g. $p(disease|\boldsymbol{x})$ may be 0.9 when it should be 0.5)

Reminders:
- $\boldsymbol{x}$ denotes features (e.g. symptoms)
- y denotes labels (e.g. disease status)
- p indicates source-domain (labels known)
- q indicates target domain (labels unknown)
- Label shift assumes $q(\boldsymbol{x}|y) = p(\boldsymbol{x}|y)$
- If we estimate $p(y|\boldsymbol{x})$, $p(y)$ from source data & are told $q(y)$, we can find $q(y|\boldsymbol{x})$ using Bayes' rule
- Given accurate $p(y|\boldsymbol{x})$, $p(y)$, we can find $q(y)$ through Maximum Likelihood (including EM)

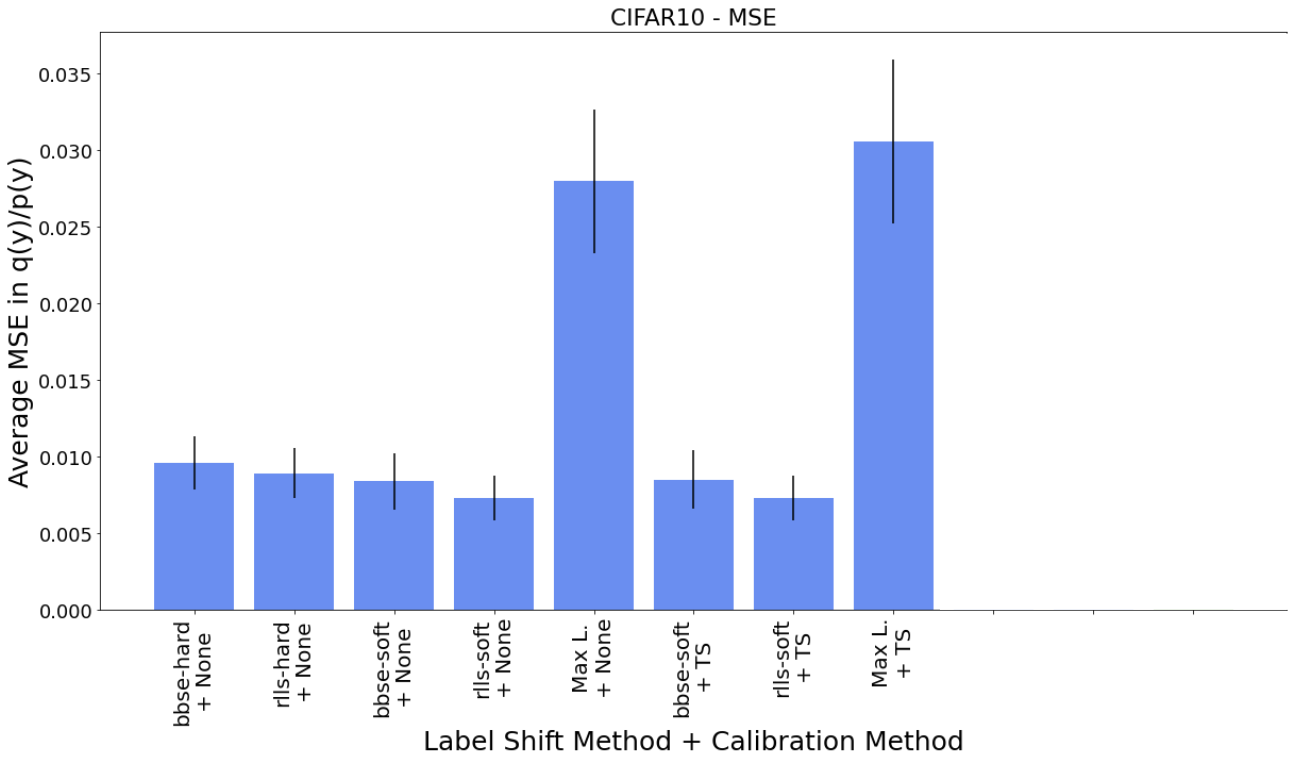
Getting Max. Likelihood Estimation to Work…

- Both BBSE and RLLS require a held-out set on which to find the confusion matrix
- We reasoned: if major barrier to Max. Likelihood is calibration requirement, why not use the held-out set to calibrate the predictions prior to doing the optimization?
- Guo et al. (ICML 2017) recommended Temperature Scaling (TS), where softmax logits $z(\boldsymbol{x}^k)$ are scaled by "temperature" T to optimize cross-entropy on validation set:

$$p(y_i|\boldsymbol{x}^k) = \frac{e^{z(\boldsymbol{x}^k)_i/T}}{\sum_j e^{z(\boldsymbol{x}^k)_j/T}}$$

Trying Temperature Scaling…

- Evaluation metric: mean squared error in estimate of $q(y)/p(y)$
- Dirichlet shift ($\alpha = 0.1$) simulated over 10 trials for each of 10 different trained models (100 trials in total). N=2000 samples were used in validation & test sets (results are qualitatively similar for different α and N as well).



Getting Max. Likelihood Estimation to Work…

- Both BBSE and RLLS require a held-out set on which to find the confusion matrix
- We reasoned: if major barrier to Max. Likelihood is calibration requirement, why not use the held-out set to calibrate the predictions prior to doing the optimization?
- Guo et al. (ICML 2017) recommended Temperature Scaling (TS), where softmax logits $z(\boldsymbol{x}^k)$ are scaled by "temperature" T to optimize cross-entropy on validation set:

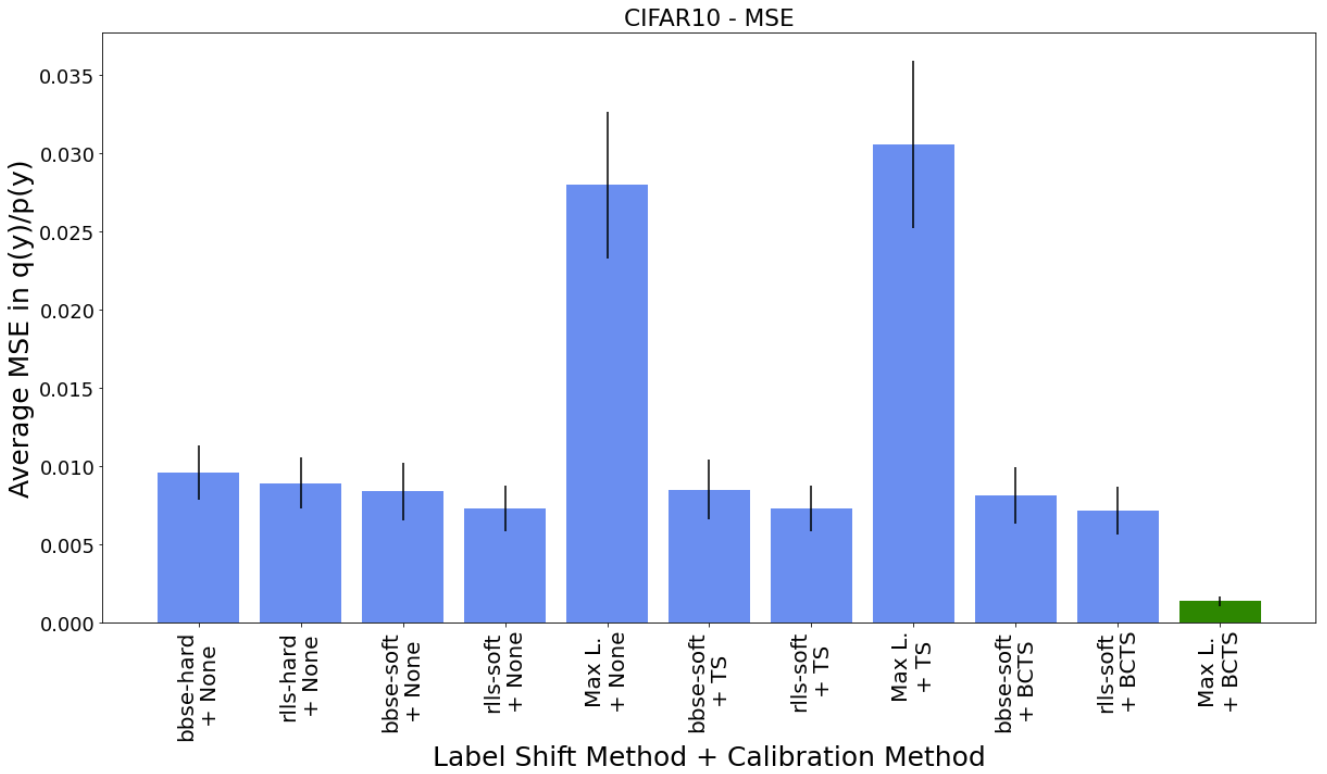
$$p(y_i|\boldsymbol{x}^k) = \frac{e^{z(\boldsymbol{x}^k)_i/T}}{\sum_j e^{z(\boldsymbol{x}^k)_j/T}}$$

- We observed **systematic bias in $\boldsymbol{p(y)}$** from Temperature Scaling. To fix, we devised a variant that included bias correction terms, called Bias-Corrected Temperature Scaling (BCTS):

$$\text{BCTS: } p(y_i|\boldsymbol{x}^k) = \frac{e^{\frac{z(\boldsymbol{x}^k)_i}{T}+b_i}}{\sum_j e^{\frac{z(\boldsymbol{x}^k)_j}{T}+b_j}}$$

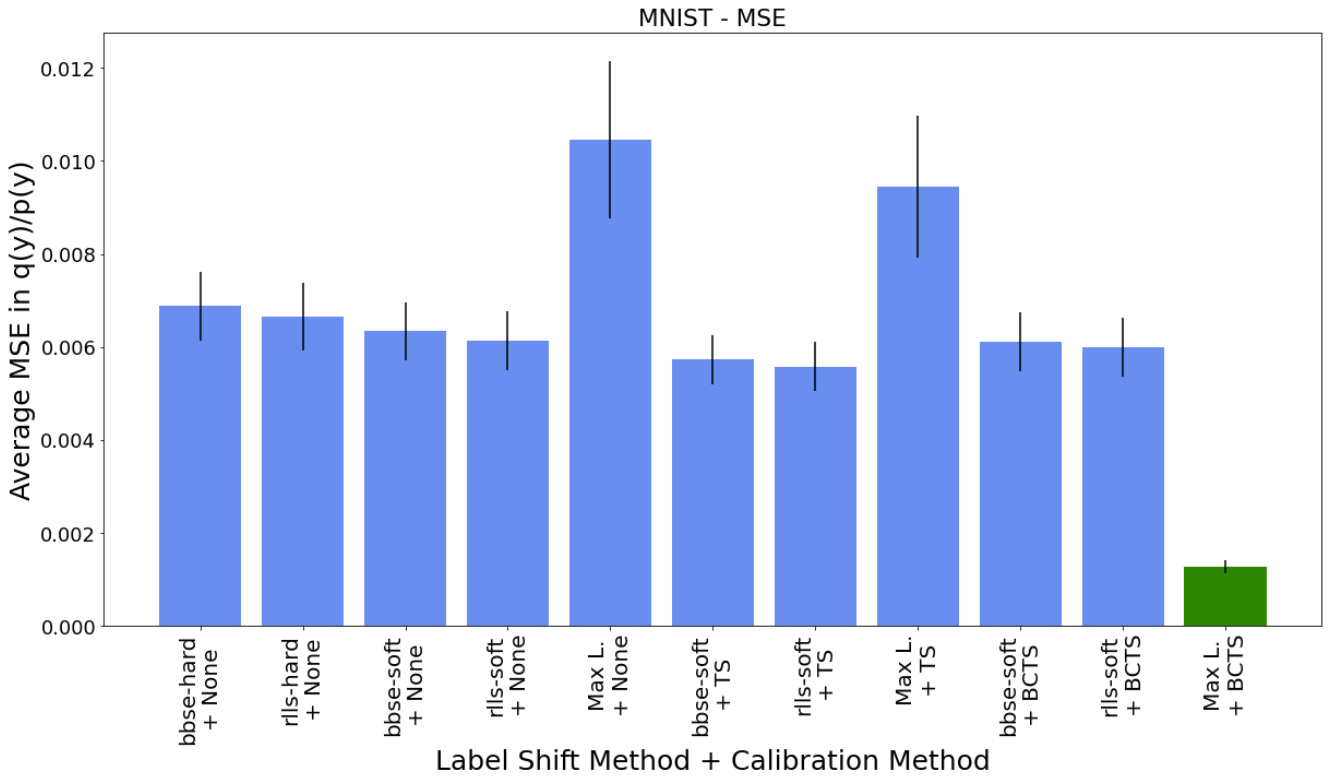
CIFAR10 benchmarking

- Evaluation metric: mean squared error in estimate of $q(y)/p(y)$
- Dirichlet shift ($\alpha = 0.1$) simulated over 10 trials for each of 10 different trained models (100 trials in total). N=2000 samples were used in validation & test sets (results are qualitatively similar for different α and N as well).



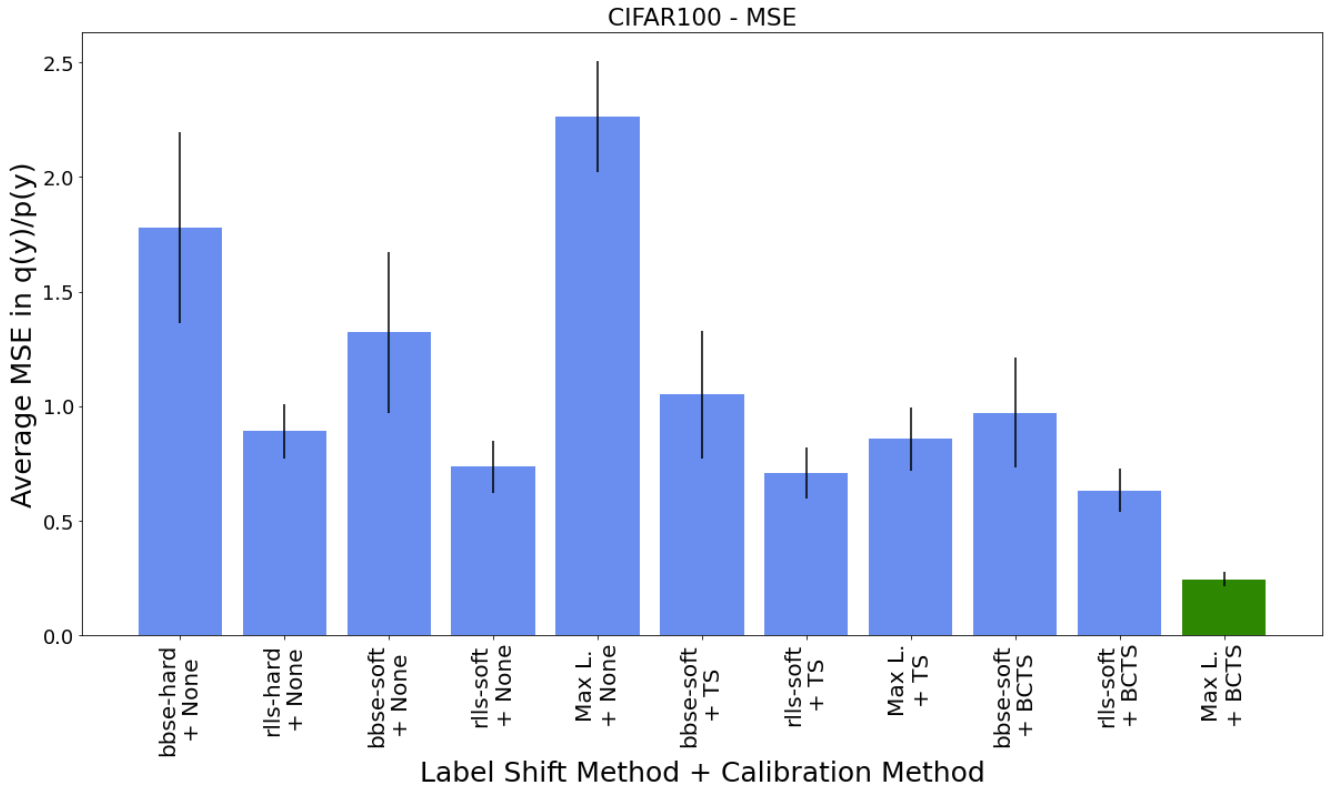
MNIST results

- Evaluation metric: mean squared error in estimate of $q(y)/p(y)$
- Dirichlet shift ($\alpha = 0.1$) simulated over 10 trials for each of 10 different trained models (100 trials in total). N=2000 samples were used in validation & test sets (results are qualitatively similar for different α and N as well).



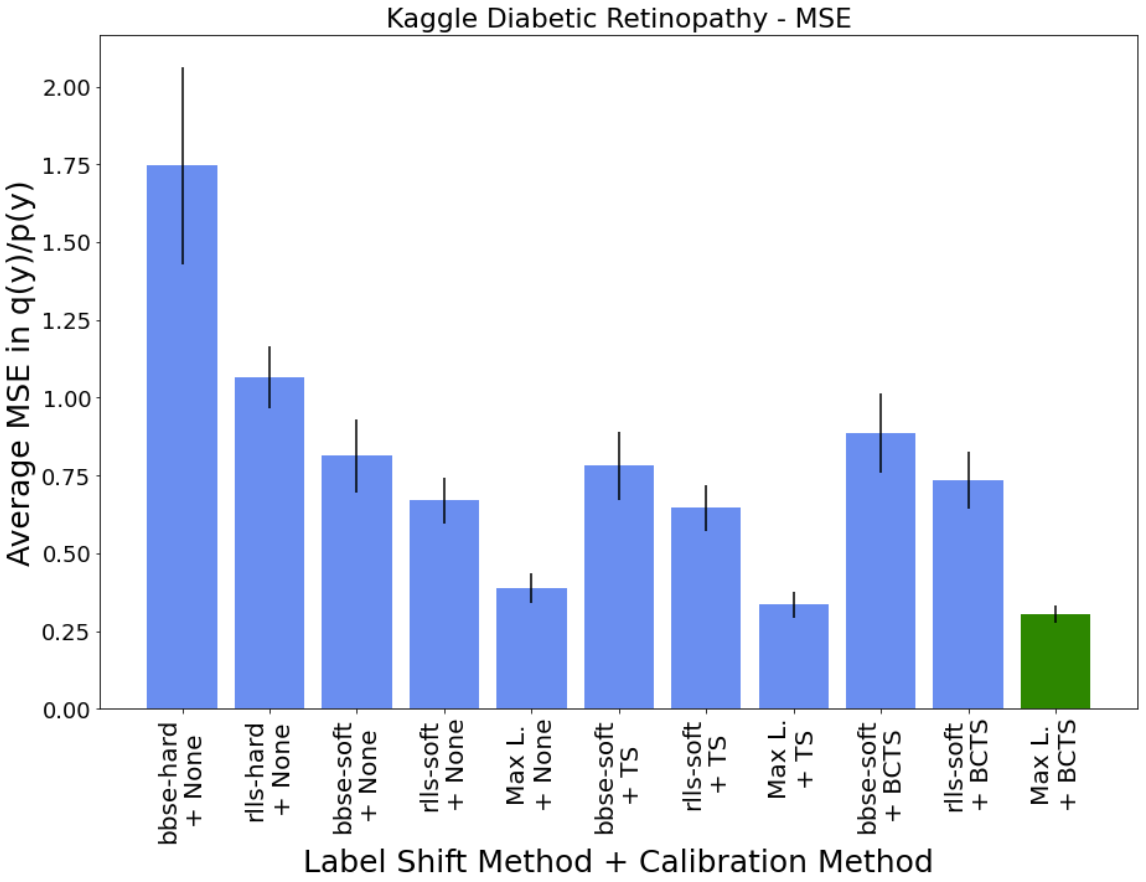
CIFAR100 results

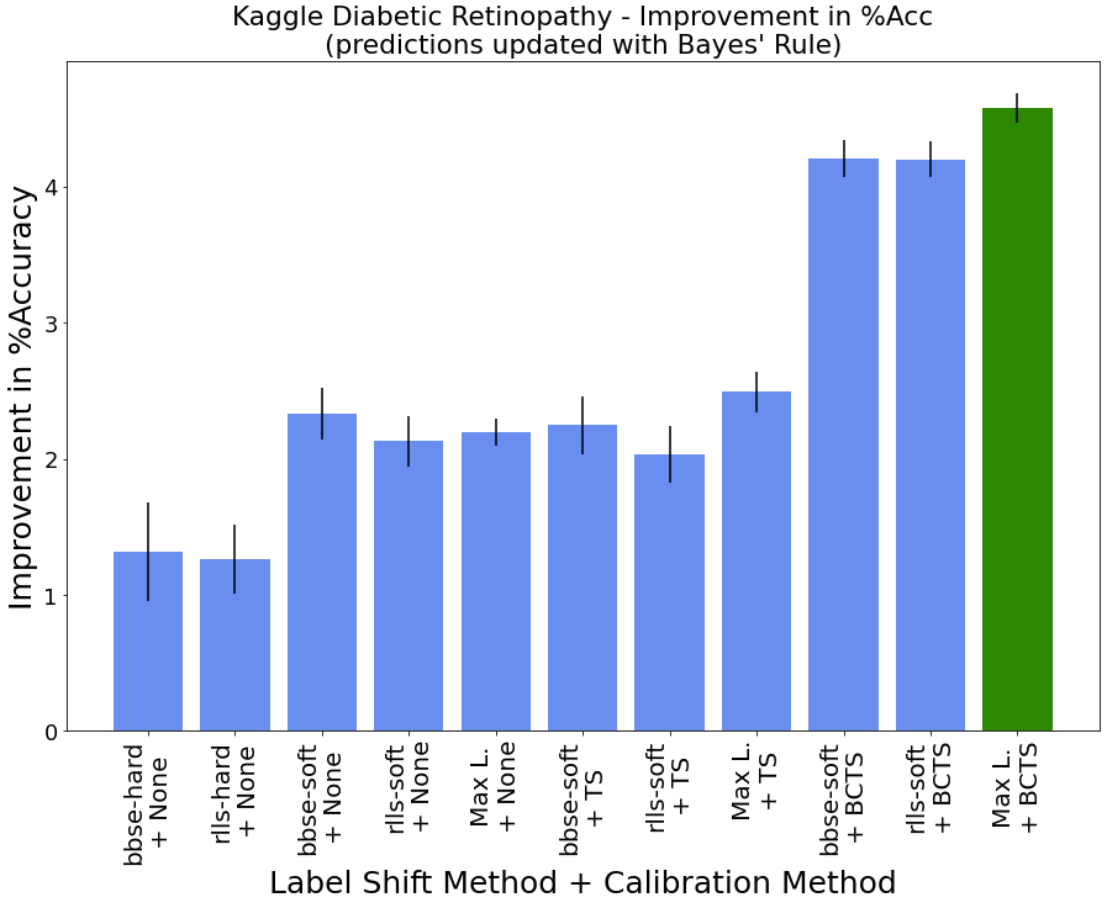
- Evaluation metric: mean squared error in estimate of $q(y)/p(y)$
- Dirichlet shift ($\alpha = 0.1$) simulated over 10 trials for each of 10 different trained models (100 trials in total). N=7000 samples were used in validation & test sets (results are qualitatively similar for different α and N as well).



Diabetic Retinopathy Detection

- Class proportion shift; target domain set to have 50% healthy instead of original 73% healthy. N=1500 samples were used in validation & test sets (results are qualitatively similar for different % and N as well).





Conclusion

- Maximum Likelihood + specific types of calibration gives **state-of-the-art** performance at domain adaptation to label shift
- Popular calibration approach of Temperature Scaling (TS) was not good enough
  - Adding terms to **minimize systematic bias** was important.
  - Alongside BCTS, we found Vector Scaling (VS), which also has bias-correction, works well.
  - VS was introduced alongside TS in Guo et al. 2017, but did not outperform TS according to the ECE metric they used. Consistent with arguments that the ECE metric used in Guo et al. (which considers only the most confidently-predicted class) may not be best metric (Vaicenavicius et al., 2019).
  - Other calibration forms like Matrix-ODIR (Kull et al., NeurIPS 2019) may also work well
- Main results **independently confirmed** by Garg, Wu, Balakrishnan & Lipton (2020) https://arxiv.org/abs/2003.07554, who studied **why** our ML+BCTS works well. Quote:

> We examine the performance of various estimators across all three datasets for various target dataset sizes and shift magnitudes (Figure 1). Across all shifts, MLLS (with BCTS-calibrated classifiers) *uniformly dominates* BBSE, RLLS, and MLLS-CM in terms of MSE (Figure 1). Observe for severe shifts, MLLS is

Garg et al. paper also includes theoretical analysis of impact of miscalibration error.