# A Omitted Proofs

**Proof of Proposition 1.** We use the Triangle Inequality to bound the distance between the vector of predictions of the baseline model and the predictions of a competing model in the $\epsilon$-level set. Let $y = \{y_i\}_{i=1}^n$ be the vector of labels, let $\hat{y} = \{h_0(\boldsymbol{x}_i)\}_{i=1}^n$ be the vector of predictions of the baseline model, and let $y' = \{h'(\boldsymbol{x}_i)\}_{i=1}^n$ be the predictions of a competing model $h'$ in the $\epsilon$-level set. Note that $y, y', \hat{y} \in \{+1, -1\}^n$. Now, we can express the risk of the baseline model $\hat{R}(h_0)$, the risk of the competing model $\hat{R}(h')$, and the discrepancy between $h$ and $h'$ ,denoted $\delta(h_0, h')$, in terms of these three vectors by

$$\hat{R}(h_0) = \frac{1}{4}\|y - \hat{y}\|$$
$$\hat{R}(h') = \frac{1}{4}\|y - y'\|$$
$$\delta(h_0, h') = \frac{1}{4}\|y' - \hat{y}\|$$

Next, consider the triangle formed in $\mathbb{R}^n$ by the points $y, y'$ and $\hat{y}$, with side lengths $\|y - \hat{y}\|$, $\|y' - \hat{y}\|$ and $\|y - y'\|$. The Triangle Inequality gives us that

$$\|y' - \hat{y}\| \leq \|y - y'\| + \|y - \hat{y}\|.$$

Substituting using the three equations above, we have

$$\delta(h_0, h') \leq \hat{R}(h_0) + \hat{R}(h').$$

Since $h' \in S_\epsilon(h_0)$, we have by the definition of the $\epsilon$-level set that $\hat{R}(h') \leq \hat{R}(h_0) + \epsilon$. We can then rewrite the above expression to yield

$$\delta(h_0, h') \leq 2\hat{R}(h_0) + \epsilon$$

Recall that $\delta_\epsilon(h_0) := \max_{h' \in S_\epsilon(h_0)} \delta(h_0, h')$. Since each $h' \in S_\epsilon(h_0)$ satisfies $\delta(h_0, h') \leq 2\hat{R}(h_0) + \epsilon$, we have the result that $\delta_\epsilon(h_0) \leq 2\hat{R}(h_0) + \epsilon$. $\square$

## B MIP Formulation for Training the Best Linear Classifier

We fit a classifier that minimizes the training error by solving an optimization problem of the form:

$$\min_{h \in \mathcal{H}} \quad \sum_{i=1}^{n} \mathbb{1}[h(\boldsymbol{x}_i) \neq y_i] \tag{5}$$

We solve this optimization problem via the following MIP formulation:

$$
\begin{aligned}
\min \quad & \sum_{i=0}^{n} l_i & & \\
\text{s.t.} \quad & M_i l_i \geq y_i(\gamma - \sum_{j=0}^{d} w_j x_{ij}) & i = 1, ..., n & \quad \text{(6a)} \\
& \gamma \leq -h_0(\boldsymbol{x}_j) \sum_{j=0}^{d} w_j x_{ij} & & \quad \text{(6b)} \\
& 1 = l_i + l_{i'} & (i, i') \in K & \quad \text{(6c)} \\
& w_j = w_j^+ + w_j^- & j = 0, ..., d & \quad \text{(6d)} \\
& 1 = \sum_{j=0}^{d} (w_j^+ - w_j^-) & & \quad \text{(6e)} \\
& l_i \in \{0, 1\} & i = 1, ..., n & \\
& w_j \in [-1, 1] & j = 0, ..., d & \\
& w_j^+ \in [0, 1] & j = 0, ..., d & \\
& w_j^- \in [-1, 0] & j = 0, ..., d &
\end{aligned}
$$

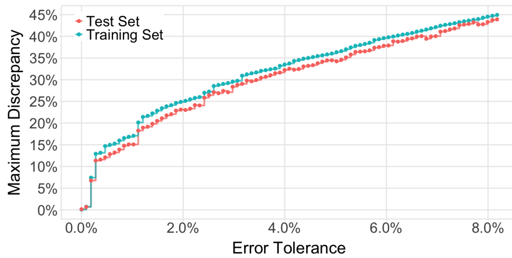
Here, constraints (6a) set the mistake indicators $l_i \leftarrow \mathbb{1}[h(\boldsymbol{x}_i) \neq y_i]$. These constraints depend on: (i) a margin parameter $\gamma > 0$, which should be set to a small positive number (e.g., $\gamma = 10^{-4}$); and (ii) the "Big-M" parameters $M_i$ which can be set as $M_i = \gamma + \max_{\boldsymbol{x}_i \in X} \|\boldsymbol{x}_i\|_\infty$ since we have fixed $\|\boldsymbol{w}\|_1 = 1$ in constraint (6e). Constraint (6c) produces an improved lower bound by encoding the necessary condition that any classifier must make exactly one mistake between any two points $(i, i') \in K$ with identical features $\boldsymbol{x}_i = \boldsymbol{x}_{i'}$ and conflicting labels. Here, $K = \{(i, i') : \boldsymbol{x}_i = \boldsymbol{x}_{i'}, y_i = +1, y_{i'} = -1\}$ is the set of points with conflicting labels.

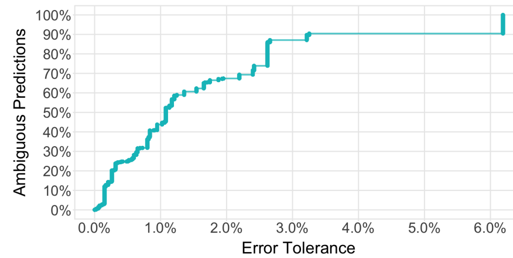
# C Additional Experimental Results

## EXACT MEASUREMENT

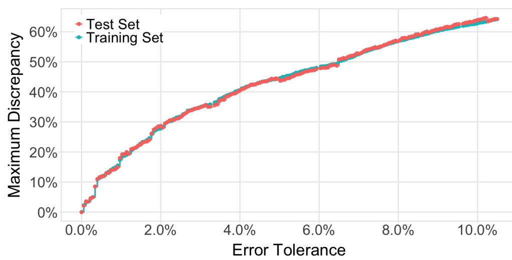
DATASET | DISCREPANCY | AMBIGUITY

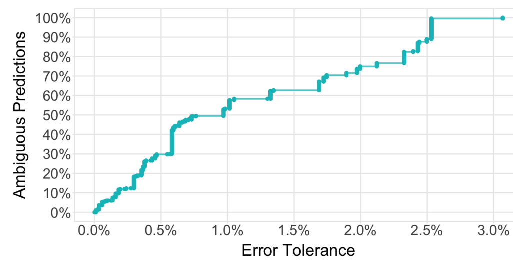
`compas arrest`



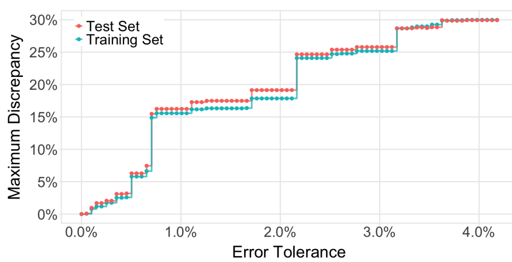



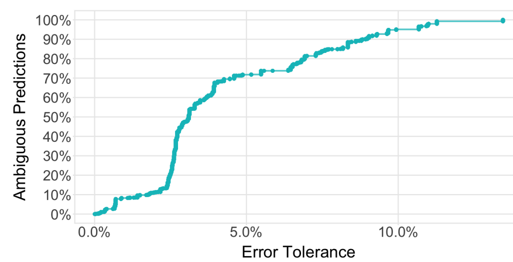

`compas violent`



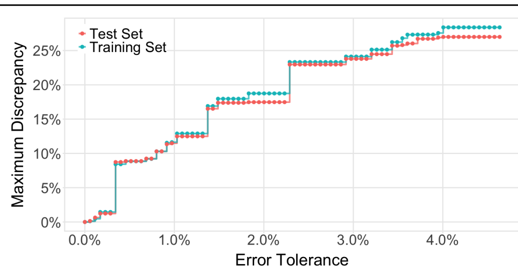




`pretrial CA arrest`




`pretrial CA fta`



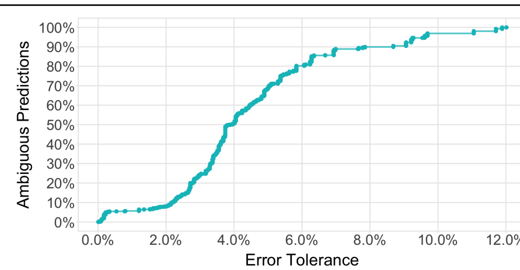




*Figure 4.* Multiplicity profiles for the `compas` and `pretrial` datasets.

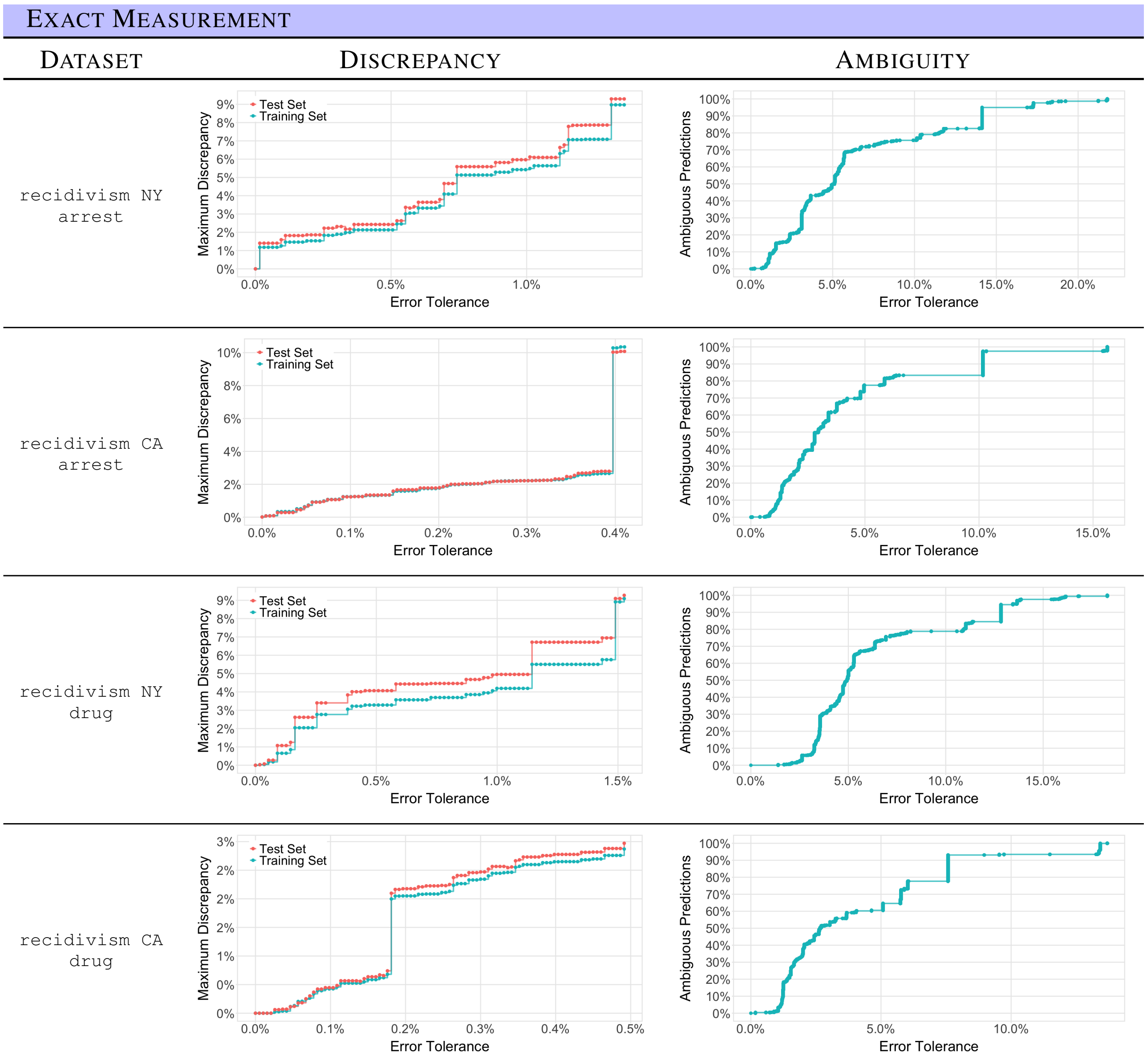



*Figure 5.* Multiplicity profiles for the `recidivism` datasets.

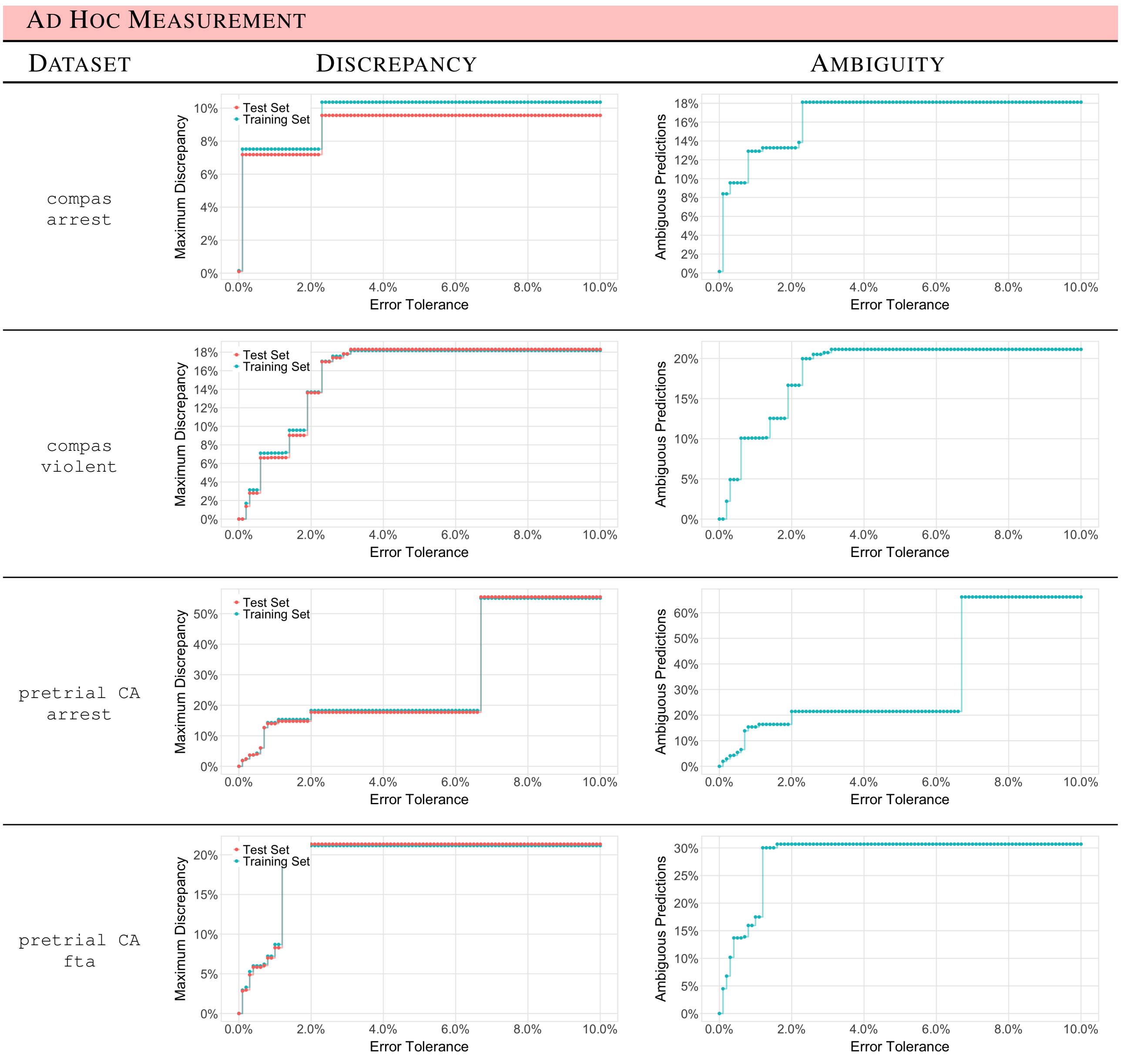



*Figure 6.* Multiplicity profiles for the `compas` and `pretrial` datasets produced via pools of logistic regression models.

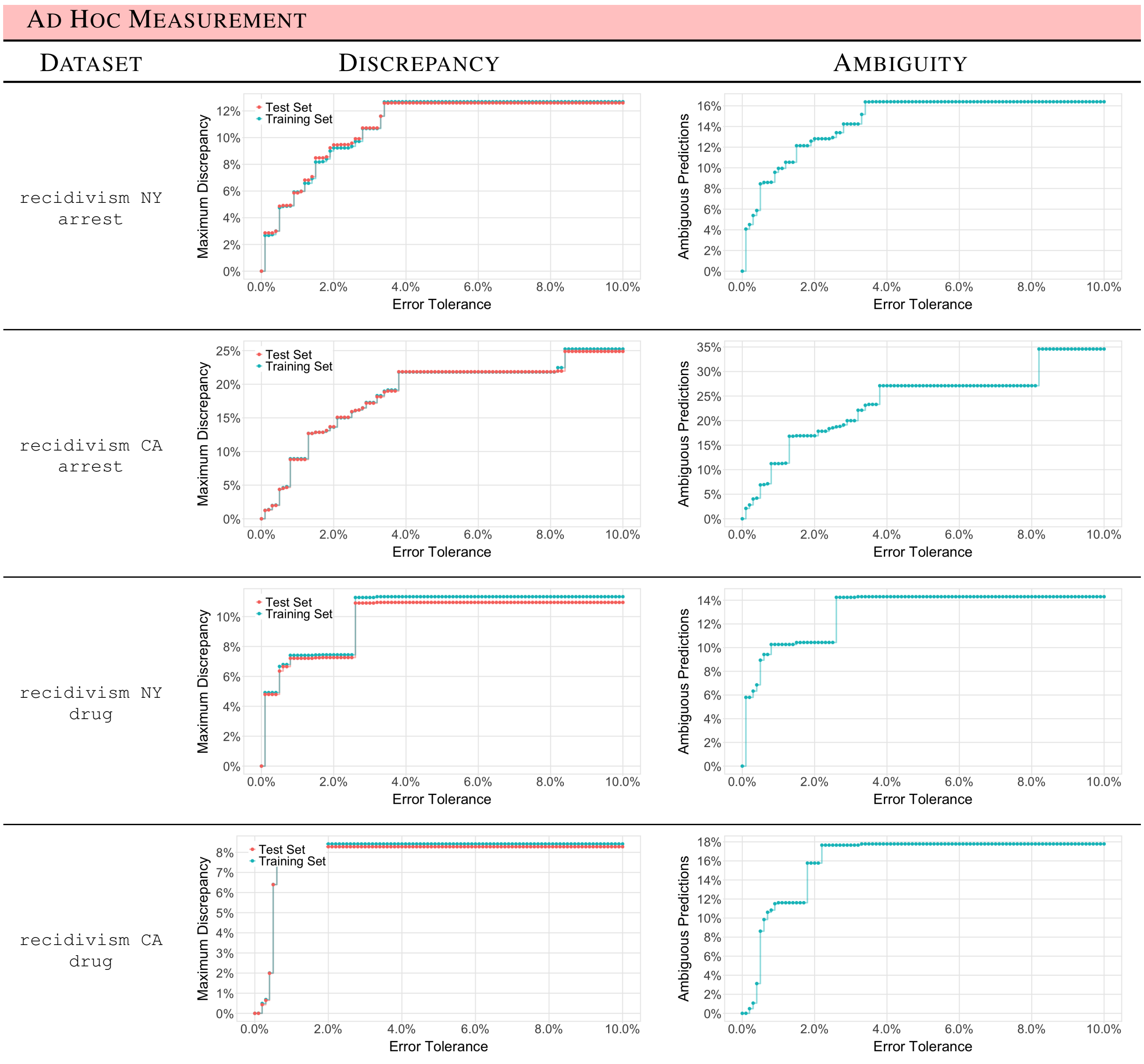



*Figure 7.* Multiplicity profiles for the `recidivism` datasets produced via pools of logistic regression models.