A Appendix

A.1 Comparison of technical approaches for data quality issues

In this section, we provide more detail on technical approaches discussed in Section 3.

To provide more context on the space of technical interventions in view of data bias, we overview classical approaches to measurement error as well as recent methodological proposals. In the classical inferential lmeasurement error literature, noise in a regression outcome (Y) such as time to recidivism is generally less of an issue than noise in covariates, or a conditioning variable such as the protected attribute. Still, different approaches for measurement error incur different informational costs. Categories of technical strategies include those based on inferential approaches to measurement error from econometrics [114], classification-specialized approaches modifying the (surrogate) loss function [115], or agnostic robustness/sensitivity analysis (sensitivity analysis relative to adversarial perturbationos, or e.g. adapting robust logistic regression such as [116]). [37] modify the surrogate loss function to ensure unbiased estimation of average misclassification cost, assuming group-dependent label noise is known. Approaches based on agnostic robustness pursue distributional robustness in terms of, what are the ranges of accuracies achieved under feasible values of the label, under a bounded number of perturbations? These approaches do face barriers in translation for use in CJ settings. The costs of robustness may introduce harms in accuracy to other groups. The "underlying distribution of crime" is unknown, and so calibration of label noise from validation data is not immediate. [43] consider an approach that uses adversarial robustness over potential values of the protected attribute. (However, in the COMPAS example, the approach, which depends on "computationally identifiable protected attribute", suggests that the protected attribute information was not strongly computationally identifiable.)

Specifically designed empirical studies or auxiliary data analysis could help calibrate these methods. Fogliato et al. [36] note that sensitivity of their regression analysis to the sample suggests model misspecification and ultimately caution "that the utility of data from NIBRS for the estimation of sampling bias in RAIs seems fairly limited."

Of approaches based on classical inferential approaches to measurement error, proxy-based approaches posit structure on how unobserved or observed noisy measurements reflect the underlying true data. (Our use of "proxy" draws on the measurement error literature. In CJ, "proxy" variables are additionally posited to be primary contributors to another variable, rather than aan auxiliary mediator.) It is expected that there are inherent limitations to the informativity of covariate-based proxy-based approaches: typically approaches based on proxies for race improve in estimation as the proxy becomes a more accurate predictor for race. Typically these proxies such as population by race in neighborhood cannot be perfectly predictive. Assessing the implications of using proxy variables is similar to assessing the ability to ensure fairness constraints based only X information alone, which Lipton et al. [117] argues can simply reallocate errors to those who misfit the "stereotypes" associated with the predictive relationship of covariates and protected attribute information. Another category of proxy variable simply models the observed attribute as a noisy misclassification of its true value. Then descriptive statistics such as those reported in Lum et al. [38], based on a unique structure of multiple records associated with an individual, are crucial to calibrate these mismeasurement probabilities.

For the problem of *distribution shift*, standard methodological approaches typically assume "missingness at random," which is often not the case due to judicial discretion and other human-driven discretion in CJ decisions. These issues would be particularly relevant in trying to assess the impacts of alternative RAIs on outcomes. The typical "missing at random" assumption posits that decisions resulting in missing data (e.g. detention) are conditionally independent of the downstream outcomes, conditional on observed data. Therefore, observed data are sufficient to statistically adjust for the resulting selection bias. This assumption is violated because administrative data records a limited subset of information relative to the judge's discretion. Unobserved confounders are the norm, rather than the exception. That is, historical selection decisions by judges likely take into account additional information that is unavailable to downstream analysts: case details, subjective information such as courtroom aspects that can provide evidence of community ties, etc. Judge "leniency scores", as used elsewhere in instrumental variable analysis [57, 118], provides quantitative empirical evidence of the extent of variability in sentencing decisions. Jung et al. [119] applies Bayesian sensitivity analysis to

assess sensitivity of algorithmic performance assessment to potential unobserved confounders under parametric restrictions.

One domain level consideration in CJ is that of historical normative considerations that govern the encodings for covariates. For example, not only the number of prior charges, but the timeframe of when these prior charges occurred is salient. RAIs are generally desired to be simple for transparency or interpretability. Therefore they discretize of these and other continuous variables. The number of prior FTAs is discretized into recent 2-year vs. older than 2-year, or just recent 2-year FTAs; others count "two or more" types of different charges [120].

A.2 Issues with COMPAS Dataset

We include elaboration on issues with the COMPAS dataset from Table 1. Classification Thresholds: Follow-up work that uses COMPAS (including the ProPublica analysis) translates continuous, probabilistic risk scores to classification discrepancies to aid disparity analysis. However, even the highest-risk individuals have low rates of rearrest. For example only 26% of people with the highest risk scores on the Public Safety Assessment were re-arrested when validated in Kentucky [121]. Classification discrepancies may not be the best way to assess disparity and papers that use COMPAS often assess arbitrary thresholds.

Follow-up period: The ProPublica dataset reportedly used a two year follow up period [122] which resulted in some people being followed even after their case was closed. However, once a person's case closes, a new arrest that happens after that doesn't count as an additional rearrest predicted by the initial RAI, it's a new arrest for which a new (updated) RAI should be filled out and the prediction period starts over.

Differing timeframes: In the two subsamples created with the ProPublica data, to examine general rearrest and violent rearrest, a two year cutoff was used for people who were not rearrested but this cutoff was not used for people who were rearrested [123]. Instead, a longer timeframe was used for people who were rearrested resulting in a greater number of people with rearrests being retained in the dataset. This resulted in an artificial inflation of the rearrest rate.

Range restriction: It is difficult to tell exactly how the COMPAS dataset dealt with people who were incarcerated during the pretrial period (or at any point in the timeframe during which people could be included in the data set). ProPublica authors mention removing people who were incarcerated [122], and since it is possible that people who were incarcerated were also people who received higher risk scores, this may restrict the range of risk scores included in the data which can impact assessments of the predictive ability of assessments. Lowder et al. show that with high detain rates for moderate- and high-risk individuals, range restriction may significantly impact predictive validity estimates [124].

Further evidence of criminal-justice-system induced endogeneity. Lum and Isaac [125] study these concerns in a predictive policing context, considering biased reporting of drug crimes and benchmarking against survey data. Ensign et al. [126] study these "runaway feedback loops" in a partial monitoring model (bandit learning with censored feedback). Akpinar and Chouldechova [127] also studies differential reporting and shows concerns about biases in downstream algorithms can arise in general from geospatially differentiated policing practices; even if drug crime information is not used. While these papers focus on the predictive policing context and implications for predictive policing algorithms, this fundamental endogeneity could be of concern for RAI assessments because many "static risk factors" deemed fair game for RAI tools, such as number of prior arrests and charges, are themselves outputs from processes in the CJ system that may disproportionately track members of different groups.