Predicting Financial Toxicity Risk in Revenue Cycle Management

A Random Forest Approach Using Synthetic EHR Data

Riya Deepak Shet

January 2026

Abstract

Medical debt creates substantial burden for patients and healthcare providers alike. This study develops a propensity-to-pay framework using synthetic Kentucky electronic health record data to identify patients at high risk of bad debt. Five linked tables were integrated to construct a patient-level financial view, with features grounded in financial toxicity literature including comorbidity count and income. A Random Forest classifier achieved 0.206 mean cross-validated average precision (baseline: 0.022) but demonstrated fundamental limitations: 5-fold cross-validated F1 score of 0.146 ($\pm$0.316) reflecting severe class imbalance (2.1% prevalence). Hyperparameter tuning improved best CV F1 to 0.235. Critically, permutation importance analysis revealed comorbidity count as the only feature with meaningful predictive signal (importance: 0.062), while income showed minimal contribution (0.023) and age showed negative importance (-0.050). These results suggest that in synthetic data, clinical complexity—not socioeconomic factors—drives bad-debt classification, though external validity remains limited. The analysis demonstrates feasibility of EHR-based financial risk stratification while highlighting the substantial methodological challenges posed by rare outcome prediction.

1 Introduction

Medical debt has become a defining challenge for the US healthcare system. The Kaiser Family Foundation reports that uncompensated care costs for the uninsured averaged $42.4 billion annually during 2015–2017 (Coughlin et al., 2021). For individual patients, substantial proportions of adults report problems paying medical bills, with many delaying or foregoing necessary care (Hamel et al., 2022). These patient-level struggles translate into rising bad debt for hospitals, especially in settings where financial margins are already constrained.

Traditional Revenue Cycle Management (RCM) workflows are reactive: hospitals extend care, generate claims, then devote resources to chasing overdue balances with limited success. Recent work suggests a more sustainable approach: stratifying patients proactively by *propensity to pay*, using clinical and sociodemographic data to distinguish patients unable to pay from those who can, offering early financial assistance to the former (Davis et al., 2021).

This study grounds propensity-to-pay modelling in the concept of *financial toxicity*, originally coined in oncology to capture objective financial burden of treatment (Zafar and Abernethy, 2013). The financial toxicity literature demonstrates that hardship arises not only from low income or lack of insurance, but also from high treatment intensity and multimorbidity (Altice et al., 2017; Fastiggi et al., 2022). Patients with multiple chronic conditions face greater out-of-pocket costs and administrative complexity, placing them at higher default risk even when insured.

1.1 Research Question

Using Synthea-generated synthetic Kentucky electronic health record (EHR) data (Walonoski et al., 2018), this analysis addresses: **Can routinely collected EHR data—specifically comorbidity burden, income, age, and insurance status—predict high-risk bad-debt patients, and which features provide meaningful predictive signal?**

Importantly, this analysis uses a modest analytic sample: of 1,701 synthetic patients, only 186 (10.9%) had at least one billed claim, and 137 held non-zero outstanding balances. This small effective sample constrains statistical power and limits generalisability, but provides a realistic test bed for examining whether routinely available data fields

carry predictive signal.

The objective is demonstrating methodological feasibility of EHR-based financial risk stratification while critically evaluating model performance under severe class imbalance.

2 Methodology

2.1 Data Sources and Extract, Transform, Load (ETL)

This analysis used synthetic EHR data generated by Synthea, a validated open-source model simulating realistic patient trajectories (Walonoski et al., 2018). The Kentucky cohort comprised 1,701 patients across five integrated tables: `patients.csv` (demographics, income), `conditions.csv` (diagnoses), `claims.csv` (claim headers, outstanding balances), `claims_transactions.csv` (payments), and `payers.csv` (coverage). Among 1,701 patients, 186 (10.9%) had at least one billed claim, with 137 holding non-zero outstanding balances totalling $22,521.

Data preparation followed the Observational Health Data Sciences and Informatics (OHDSI) community conventions for transparent cost attribution (Blacketer and Voss, 2021). Condition-level data were aggregated at the patient level to derive a comorbidity count (total distinct diagnosis codes per patient). Transaction-level records were grouped by claim, with a transfer-type field distinguishing payments made by insurers from those made by patients. Financial integrity was verified by checking that paid amounts plus outstanding balances approximately equalled billed amounts—exact equality is not expected in claims data because of contractual adjustments between billed charges and negotiated rates, coordination-of-benefits transfers across payers, and minor rounding in transaction records.

2.2 Feature Engineering and Outcome Definition

Four interpretable features operationalised financial toxicity theory:

- **Age**: Calculated from birth date
- **Income**: Annual household income from patient demographics
- **Comorbidity count**: Number of distinct diagnosis codes per patient

- **Insurance status**: Binary indicator (insured vs. uninsured)

The binary high-risk outcome was defined using a compound rule: patients in the top 10% of outstanding debt (≥$205.30) **or** with payment ratio <1% on non-zero balance. This yielded 36 high-risk patients, creating severe class imbalance (47:1 ratio).

2.3 Predictive Modelling

Given severe class imbalance, a Random Forest classifier was trained with balanced subsample weighting, which automatically assigns higher penalty to misclassifying the rare high-risk class (Breiman, 2001). Data were split 70:30 with stratification preserving class distribution. The decision threshold was lowered from the default 0.5 to 0.2 to increase sensitivity for high-risk patient detection, accepting a trade-off of reduced precision. This trade-off is appropriate for financial screening: missed high-risk patients generate uncompensated care—estimated at $42.4 billion annually for the uninsured alone (Coughlin et al., 2021)—whereas false positives merely trigger a low-cost review by financial counselling staff. Proactive identification of at-risk patients enables earlier assistance interventions that can reduce eventual bad debt (Davis et al., 2021).

Model evaluation employed multiple approaches appropriate for imbalanced data:

1. **5-fold stratified cross-validation** for robust performance estimation
2. **Average precision**, summarised as the area under the precision-recall curve (AUC-PR), rather than the area under the receiver operating characteristic curve (AUC-ROC), as recommended for imbalanced classification (Saito and Rehmsmeier, 2015)
3. **Grid-search hyperparameter tuning** (systematic evaluation of parameter combinations using cross-validation) optimising F1 score on the minority class
4. **Permutation importance** alongside mean decrease in impurity (MDI)—a tree-based measure of how much each feature reduces prediction error—to assess feature contributions (Breiman, 2001)

2.4 Ethical Considerations

Propensity-to-pay scores were explicitly framed as mechanisms for identifying patients needing financial assistance—not for denying services. Synthetic data avoids direct pri-

vacy risk but limits external validity. Real-world deployment would require systematic fairness evaluation across demographic subgroups.

3 Results and Discussion

3.1 Cohort Characteristics

Distribution of patient-level outstanding balances was highly right-skewed (Figure 1). Most patients with non-zero balances owed modest amounts ($35–$200), with a small number of outliers toward $8,000. The 90th percentile threshold of $205.30 defined the high-risk group, which showed mean outstanding balance of $488.63 (median: $205.30; max: $7,965) compared to $48.81 for non-high-risk patients (Table 1).

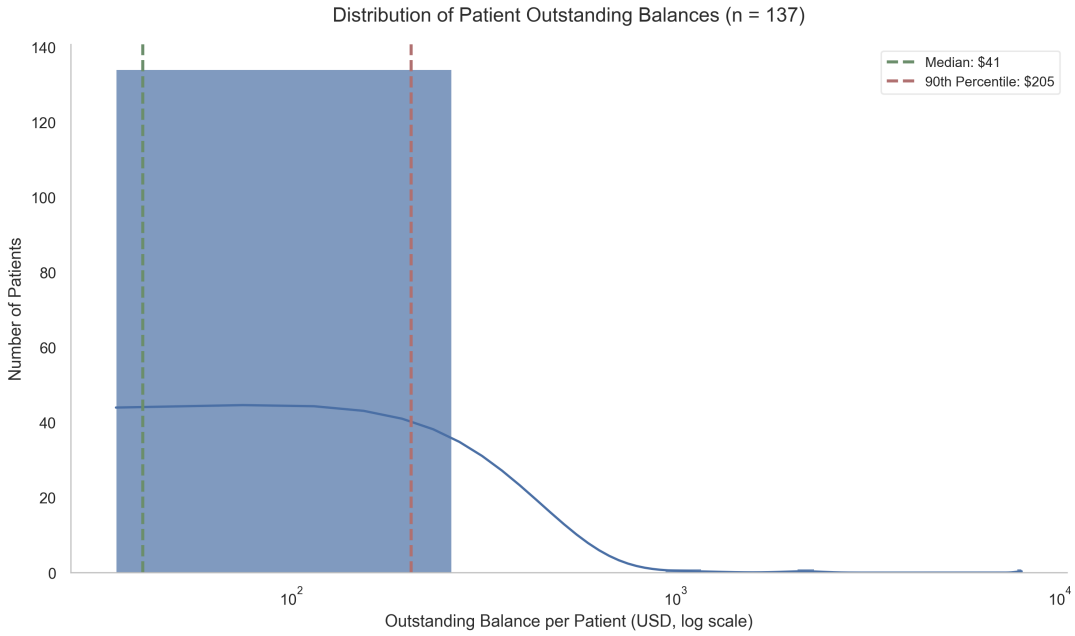

Figure 1: Distribution of non-zero patient outstanding balances (log scale, N=137). Green dashed line indicates median ($41.06); red dashed line indicates 90th percentile high-risk threshold ($205.30).

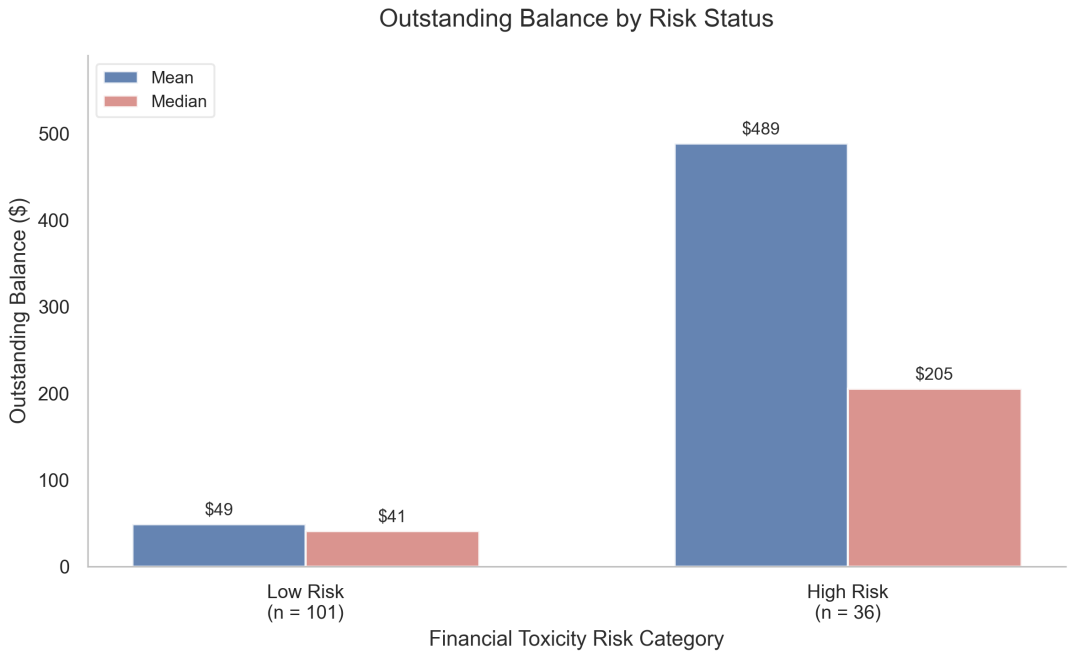

Figure 2: Comparison of outstanding balance statistics between risk groups, showing mean and median values for non-high-risk (n = 101) and high-risk (n = 36) patients.

Table 1: Summary statistics of outstanding balances by high-risk status

Group	N	Mean ($)	Median ($)	Max ($)
Non-high-risk	101	48.81	41.06	141.76
High-risk	36	488.63	205.30	7,965.00

The distributional shape differs markedly between the two groups (Figure 2, Table 1). For non-high-risk patients the mean ($48.81) lies close to the median ($41.06), indicating a roughly symmetric, concentrated distribution where most patients owe similar modest amounts. In the high-risk group, by contrast, the mean ($488.63) is more than double the median ($205.30), revealing a right-skewed distribution pulled upward by a small number of patients with very large outstanding debts (maximum $7,965). This skew suggests that the high-risk group is itself heterogeneous: some patients sit just above the 90th-percentile threshold while others carry substantially greater financial exposure, an observation with practical implications for triaging financial assistance resources.

3.2 Model Performance

Model discrimination was assessed using both threshold-dependent and threshold-independent metrics. Figure 3 compares Precision-Recall and ROC curves, highlighting the importance of metric selection under class imbalance. While the ROC curve suggests strong performance (AUC-ROC = 0.719), this is misleading given the dominance of true negatives. The Precision-Recall curve (AUC-PR = 0.156) provides more realistic assessment, demonstrating 7.3× improvement over the random baseline of 0.022.

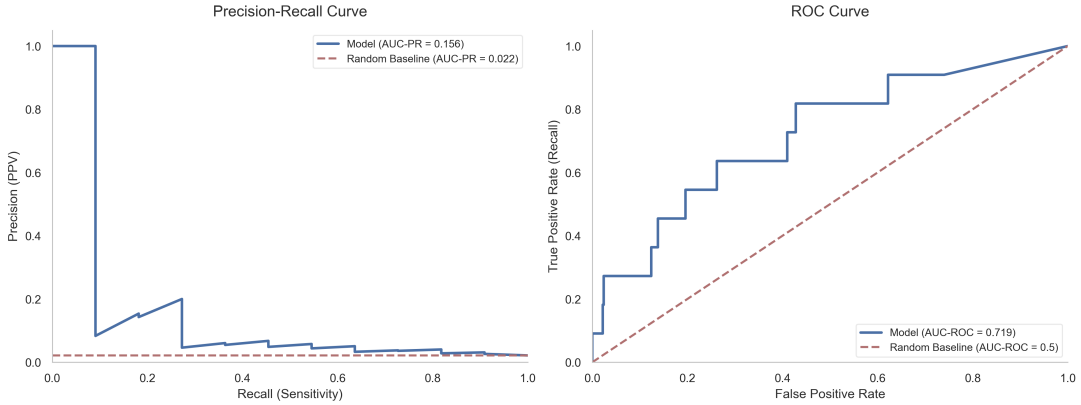

Figure 3: Precision-Recall curve (left, AUC-PR = 0.156) versus ROC curve (right, AUC-ROC = 0.719). The PR curve provides more informative evaluation under severe class imbalance, showing 7.3× improvement over random baseline.

To operationalise the model for patient screening, the decision threshold was lowered from 0.5 to 0.2, prioritising sensitivity over specificity. At this threshold, the classifier identified 4 of 11 high-risk patients in the test set (36.4% recall) while flagging 65 false positives (Figure 4). Overall accuracy of 85.9% illustrates the well-documented problem that accuracy is misleading for imbalanced classification—the model could achieve 97.8% accuracy by predicting all patients as low-risk.

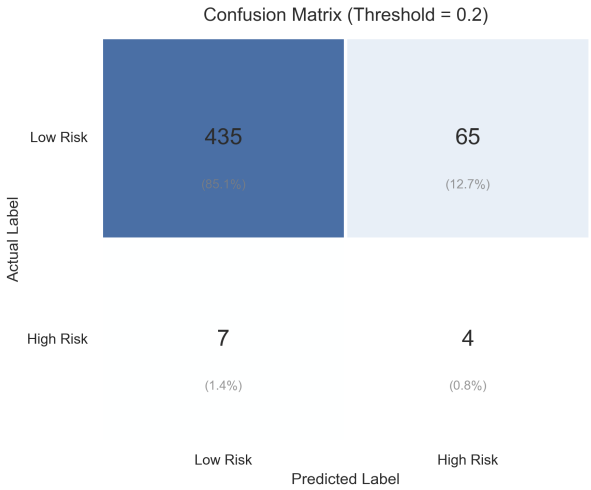

Figure 4: Confusion matrix for propensity-to-pay classifier at decision threshold 0.2, evaluated on held-out test set ($n = 511$). The model detected 4 of 11 true high-risk patients while generating 65 false positives among 500 low-risk patients.

Cross-validation provided more robust performance estimates. Five-fold stratified cross-validation (CV) yielded mean average precision of 0.206 ($\pm$0.182), substantially exceeding the random baseline (9.4$\times$ improvement). Minority-class metrics showed recall of 0.143 ($\pm$0.313), precision of 0.156 ($\pm$0.329), and F1 of 0.146 ($\pm$0.316), with high variance reflecting instability across folds given only 7 high-risk patients per fold on average.

Hyperparameter tuning via grid search (324 combinations, 1,620 fits) identified optimal parameters: 100 trees, unrestricted depth, a minimum of 10 samples to split a node, a minimum of 2 samples per leaf, and balanced class weighting. This improved best cross-validated F1 to 0.235—modest but meaningful given 2.1% prevalence.

3.3 Feature Importance Analysis

The critical finding emerged from comparing MDI and permutation importance (Table 2, Figure 5). MDI suggested income (0.362) and comorbidity count (0.354) were near-equally important, with age contributing substantially (0.267). However, permutation importance—which measures actual predictive contribution by randomly shuffling features—revealed a starkly different picture.

Table 2: Comparison of feature importance methods.

Feature	MDI Importance	Permutation Importance
Comorbidity count	0.354	**0.062 (±0.046)**
Income	0.362	0.023 (±0.059)
Age	0.267	−0.050 (±0.032)
Insurance status	0.017	−0.001 (±0.003)

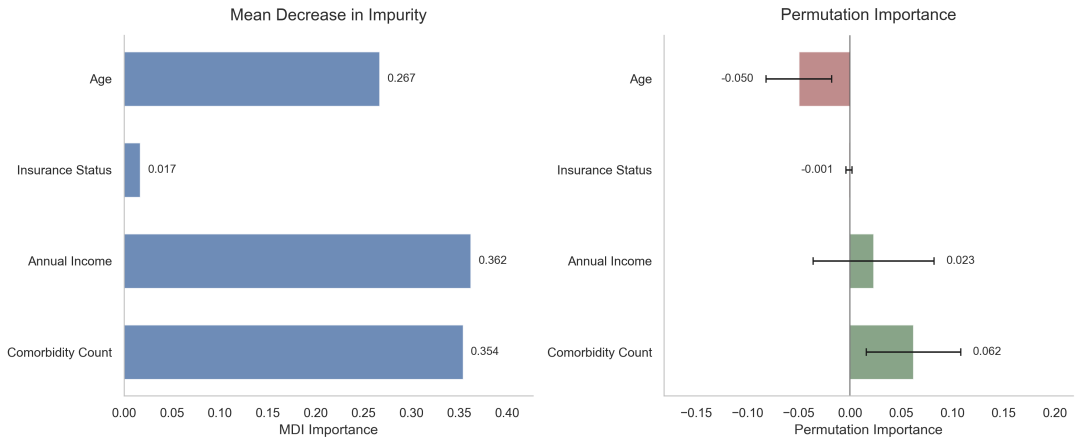

Figure 5: Feature importance comparison: MDI (left) versus permutation importance with standard deviation error bars (right). Green bars indicate positive contribution; red bars indicate features that degrade model performance when included.

Only **comorbidity count** showed positive permutation importance (0.062), indicating genuine predictive contribution. Income's permutation importance (0.023) was small relative to its variance, suggesting limited reliable signal. Strikingly, **age showed negative permutation importance** (−0.050), meaning removing age *improved* model performance—its apparent MDI importance was spurious, likely reflecting overfitting to noise.

This discrepancy is methodologically significant. MDI is biased toward high-cardinality continuous features and can indicate importance even when features add no predictive value (Strobl et al., 2007). Permutation importance provides the more trustworthy estimate of genuine predictive contribution.

3.4 Implications for Financial Toxicity Theory

The finding that comorbidity count alone provides meaningful signal aligns with financial toxicity literature emphasising clinical complexity as a driver of financial hardship, though the signal is weak. Comorbidity burden increases bad-debt risk through multiple pathways: higher out-of-pocket accumulation toward deductibles, greater care coordination failures, and increased likelihood of disability affecting employment-based income (Altice et al., 2017; Fastiggi et al., 2022). However, the minimal contribution of income contradicts expectations that socioeconomic factors would dominate. This likely reflects Synthea's synthetic data generation: income may not be realistically correlated with payment behaviour in simulated data.

The negligible insurance status contribution (permutation importance: -0.001) suggests that in this synthetic cohort, coverage status does not differentiate bad-debt risk once other factors are considered. Real-world data would likely show stronger insurance effects.

3.5 Limitations

Several limitations warrant acknowledgment:

First, synthetic data fundamentally limits external validity. Synthea generates structurally plausible but not behaviourally realistic financial data. The near-perfect payment ratios (78.2% of patients with 100% payment ratio) are unrealistic compared to empirical US medical debt estimates (Hamel et al., 2022).

Second, severe class imbalance (2.1% prevalence) constrained model performance despite balanced class weights. With only 36 high-risk patients (11 in test set), performance estimates remain unstable. Production deployment would require larger datasets or alternative approaches such as cost-sensitive learning.

Third, the high-risk label is a constructed proxy rather than observed default. Performance metrics demonstrate methodological feasibility, not deployment-ready prediction.

Fourth, permutation importance revealing that most features contribute minimally suggests the feature set may be insufficient for this prediction task. Richer features (prior payment history, social determinants, neighbourhood-level data) may be necessary.

4 Conclusion

This study developed a propensity-to-pay framework demonstrating both the feasibility and challenges of EHR-based bad-debt prediction. Key findings include:

1. Cross-validated average precision (0.206) exceeded random baseline (0.022) by 9.4×, indicating meaningful signal despite severe imbalance

2. Hyperparameter tuning improved best CV F1 from 0.146 to 0.235

3. Permutation importance analysis revealed comorbidity count as the only feature with genuine predictive contribution; income and age showed minimal or negative importance

4. MDI feature importance was misleading, highlighting the need for robust importance methods

The dominance of comorbidity count supports the financial toxicity hypothesis that clinical complexity drives financial hardship. However, the weak performance of socioeconomic features likely reflects synthetic data limitations rather than true relationships.

Real-world deployment requires: (1) validation on actual EHR/claims data, (2) explicit class imbalance handling, (3) systematic fairness evaluation, and (4) richer feature engineering. Nevertheless, this analysis demonstrates a practical pathway for health systems to develop targeted financial assistance interventions using routinely collected data.

5 AI Usage Statement

The author acknowledges the use of generative AI tools (Claude, Perplexity, and Google AI Studio) in accordance with University of Birmingham guidelines on acceptable AI use. These tools were employed in the following capacities:

Literature Search: AI tools assisted in identifying relevant publications and research directions during initial project ideation. All sources identified were independently verified and critically evaluated by the author before inclusion.

Code Development Support: AI tools were consulted to generate code templates and clarify programming concepts. All code generated was reviewed, tested, and modified by the author to ensure correctness and appropriateness for the analytical objectives. The author takes full responsibility for all code implementation and results.

Writing Assistance: AI tools provided feedback on draft structure and clarity, functioning as editorial assistance for improving academic writing conventions. All substantive content, arguments, analysis, interpretation of results, and scientific reasoning were developed independently by the author. AI tools were not used to generate, alter, or develop the core ideas, arguments, or explanations presented in this work.

The author confirms that all work submitted represents their own understanding, analysis, and interpretation of the research question and findings.

References

Altice, C.K., Banegas, M.P., Tucker-Seeley, R.D. and Yabroff, K.R. (2017), 'Financial hardships experienced by cancer survivors: a systematic review', *Journal of the National Cancer Institute*, 109(2), djw205.

Blacketer, C. and Voss, E. (2021), 'Extract, transform, load', in Hripcsak, G., Suchard, M. and Schuemie, M. (eds), *The Book of OHDSI*, Observational Health Data Sciences and Informatics, pp. 69–88.

Breiman, L. (2001), 'Random forests', *Machine Learning*, 45(1), 5–32.

Coughlin, T.A., Samuel-Jakubos, H. and Garfield, R. (2021), 'Sources of payment for uncompensated care for the uninsured', *Journal of Health Care for the Poor and Underserved*, 32(3), 1283–1301.

Davis, S., Nourazari, S., Granovsky, R. and Fard, N. (2021), 'Predicting a need for financial assistance in emergency department care', *Healthcare*, 9(5), 556.

Fastiggi, M., Sim, J. and Huang, I.C. (2022), 'Association of co-morbidities with financial hardship in survivors of adult cancer', *Supportive Care in Cancer*, 30(2), 1655–1665.

Hamel, L., Norton, M., Pollitz, K., Levitt, L., Claxton, G. and Brodie, M. (2022), *The Burden of Medical Debt: Results from the Kaiser Family Foundation/New York Times Medical Bills Survey*, Kaiser Family Foundation.

Saito, T. and Rehmsmeier, M. (2015), 'The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets', *PLOS ONE*, 10(3), e0118432.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007), 'Bias in random forest variable importance measures: illustrations, sources and a solution', *BMC Bioinformatics*, 8(1), 25.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T. and McLachlan, S. (2018), 'Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic

health care record', *Journal of the American Medical Informatics Association*, 25(3), 230–238.

Zafar, S.Y. and Abernethy, A.P. (2013), 'Financial toxicity, part I: a new name for a growing problem', *Oncology*, 27(2), 80–81.