

25 **ABSTRACT**

26 **BACKGROUND:**

27 Drug-resistant (DR) tuberculosis treatment is challenging and frequently leads to poor outcomes. An
28 international collaboration, the National Institute of Allergy and Infectious Diseases (NIAID) TB Portals
29 develops, maintains, and supports a multi-national database of tuberculosis cases, with an emphasis on
30 drug-resistant tuberculosis. Patient records include clinical, radiological, genomic, and socioeconomic
31 features. Establishing factors associated with unsuccessful treatment may help optimize treatment for the
32 most challenging infections.

33 **METHODS:**

34 Association analysis and machine learning algorithms were applied to identify important factors
35 associated with treatment outcome and predict the outcome for three patient cohorts, selected by drug
36 resistance level representing 1575 patients in total. The predicted probabilities of poor treatment outcome
37 from models were calibrated as a risk score ranging from 0 to 100 corresponding to confidence level of
38 the model for treatment outcome.

39 **RESULTS:**

40 The features most associated with treatment success in all cohorts were body mass index (BMI), onset age,
41 employment, education, smear-negative microscopy, and percent of abnormal volume in X-ray images,
42 confirming previously reported findings, and identifying novel factors such as pathogen genomic markers.

43 **CONCLUSIONS:**

44 The identified features might help in establishing high-risk patients at the time of admission for
45 tuberculosis treatment. This study integrates clinical, radiological, and pathogen genomics into a patient
46 risk model, a way of determining risk through the application of machine learning on real-world data.

47

48

49 **Keywords:** Machine Learning, Tuberculosis, Outcome Prediction

50

51 **BACKGROUND**

52 Tuberculosis (TB) is a communicable disease that is amongst the top 10 causes of death and the
53 leading cause from a single infectious agent (1). With the Covid-19 pandemic's ongoing global impact on
54 healthcare systems, TB diagnostic and clinical management may experience challenges (2). The World
55 Health Organization (WHO) released results of modeling work that if global TB case detection decreases
56 by 25% over a period of 3 months (compared to level of detection pre-pandemic), an additional 190,000
57 TB deaths might occur (3). Thus, it is essential to identify the most vulnerable or hard-to-treat TB patients
58 at the time of diagnosis. Identifying the most at-risk groups within the population may allow for better
59 prophylactic measures and monitoring in advance, to prevent outbreaks of this highly contagious disease.
60 While many tuberculosis databases or registries contain critical clinical information such as drug
61 resistance status, time of diagnosis, sputum sampling, etc., it is rarer to find the integration of drug
62 resistance with other important factors such as pathogen genomic status and social determinants of health.

63 Drug-resistant TB continues to be especially challenging to treat and control. Multidrug resistant
64 TB (MDR-TB) is TB that is resistant to both rifampicin and isoniazid, two of the most widely used anti-
65 TB drugs. Extensively drug-resistant TB (XDR-TB) is defined as MDR-TB plus resistance to at least one
66 of the fluoroquinolones and one of the injectable agents used in MDR-TB treatment regimens (1). The
67 latest global data show a treatment success rate of 85% for drug-susceptible TB, 56% for MDR-TB and
68 39% for extensively drug-resistant TB [1]. Given the lower success rate of treatment in drug resistance
69 TB, the dynamics and nature of the disease show differences when comparing to drug sensitive TB and it
70 is necessary to identify what clinical, genomic, or social determinants of health risk factors might be
71 common or distinct between these disease types to highlight patient risk.

72 Prior studies have shown the importance of analyzing and interpreting the connections between
73 clinical and social determinants of health. For example, clinical factors like culture-positivity at two
74 months of treatment, history of treatment with second-line drugs were identified as risk factors of poor

75 drug resistant tuberculosis treatment outcome in Eastern Europe and Central Asia (4). In the same study,
76 the socio-economic factor of homelessness was also identified as correlating with poor outcome. Another
77 study demonstrated a similar result where a combination of clinical factors such as MDR-TB patients
78 with HIV, high smear grade, or a history of previous MDR-TB treatment were identified together with a
79 socio-economic factor, malnutrition, as risk factors of poor treatment outcome (5). Another identified
80 social determinants of health such as employment and education status along with clinical factors such as
81 drug resistance status statistically associated and predictive of treatment failure (6). By analyzing a much
82 larger dataset of clinical, radiological, and genomic features in a unique multi-national cohort of drug
83 sensitive and drug resistant cases, our study extends earlier works that have investigated potential
84 combinations of clinical and socio-economic factors impacting treatment outcome.

85 The TB Portals (<https://tbportals.niaid.nih.gov/>) is the largest, open-access, patient-centric
86 database connecting clinical, socio-economic, pathogen genomics and patient radiological data from 16
87 countries (6-8). As part of the National Institute of Allergy and Infectious Diseases (NIAID) strategic plan
88 for TB research and National Institute of Health (NIH) strategic plan for data science(9, 10), the resource
89 is focused on how to translate real-world data collected from TB cases, primarily drug resistant, into
90 actionable information for public health following the FAIR data principles. As part of this effort, the
91 program considers researchers and public health experts with various backgrounds, providing ways to
92 visualize, interact, and analyze the data. A request for data can be made by an investigator using the web-
93 accessible data use agreement and application process, <https://tbportals.niaid.nih.gov/download-data>. The
94 data can be accessed via API or directly downloaded to facilitate visualization, analysis, modeling, and
95 other data science approaches. For users who prefer point and click interaction and analysis within a
96 website, an ecosystem of tools appropriate for each type of data have been developed (8, 11, 12). The data
97 and underlying suite of tools provides an unprecedented opportunity for public health researchers looking
98 to understand the real-world impact of TB. For example, the program and its collaborators have
99 demonstrated the applicability of TB Portals collection of data, publishing research on epidemiological

100 aspects of tuberculosis, the evolutionary processes related to drug resistance, the factors involved with
101 tuberculosis relapse versus reinfection, detailed analysis of various strains in sputum versus surgical
102 samples of tuberculosis, analysis of radiological data to discover distinguishing features of drug resistant
103 tuberculosis (7, 13, 14).

104 Here, we have analyzed real-world patients' data (i.e. not intended as a representative
105 epidemiological survey) as a retrospective case-control study using TB Portals data. Machine learning
106 techniques were applied on 1575 TB cases with complete set of multi-domain data to predict the
107 treatment outcome (either "cured" or "died"). Models were developed to each subgroup by drug resistance
108 level in order to derive a case-level risk score to facilitate understanding of the relative contribution of
109 clinical or social determinants of health factors towards the success of treatment. The insights provided
110 by these models might assist in identifying the important risk factors that could inform public health
111 policies and programs as part of a holistic, data-informed, and evidence-based approach. Moreover, this
112 study is intended to demonstrate the potential of TB Portals for real-world studies and some of the
113 important considerations of public health researchers interested in leveraging the resource.

114 **METHODS**

115 ***Data acquisition and initial processing***

116 This analysis uses publicly shared, deidentified data from The Tuberculosis Data Exploration
117 Portal (TB DEPOT) as of Mar 2021, which can be obtained by anyone who signs and agrees to a [data use](#)
118 [agreement](#) (DUA). Only cases resulting in death (died, negative outcome) or recovery (cured, positive
119 outcome) at the end of the treatment with available BMI and age measures were included. There were
120 1575 subjects with features of interest, including 299 drug-sensitive (DS) TB patients, 883 MDR-TB
121 patients, and 393 XDR-TB patients. The clinically reported type of resistance was used throughout the
122 study. These patients came from multiple countries spanning the consortium (Table 1 and Table S1).
123 Categorical feature levels with limited samples were combined: for gender, female and other were

124 combined as non-male; for education, basic school and no education were combined as basic school or
125 lower, College (bachelor) and Higher (university) were combined as college or higher; for employment,
126 Homemaker and Self-employed were combined as Homemaker or self-employed. Missing categories
127 (“Not reported”, “NA”, blank) were combined into a single missing category for any categorical data. A
128 case was considered as resistant if the pathogen showed resistance in any of First-line Drug Line Probe
129 Assay, Second-line Drug Line Probe Assay, Solid medium Lowenstein, BACTEC MGIT 960, or
130 GeneXpert MTB/RIF (Xpert) tests. Certain variables from lung images are subdivided by sextant of the
131 lung corresponding to upper, middle, lower sextant of right or left lung. A case might have multiple lung
132 sextants involving a feature from an available image or multiple images per case. In the following
133 analysis, the levels are combined into "Upper sextant - Yes" or “No”, "Middle or Lower sextant - Yes" or
134 "No" to indicate the combination of features from available image data.

135 *Association Analysis*

136 Measure of effect size in ANOVA (analysis of variance) quantifies degree of association between
137 an effect and a continuous dependent variable. Eta squared (η^2) was used to measure size of observed
138 effect with age and BMI. Cohen defined small, medium, and large effects as Eta squared values of 0.01,
139 0.06, 0.14 respectively (15). Variables with statistically significant difference in the mean of age or BMI
140 by drug-sensitive and drug-resistant subgroups were reported.

141 Uncertainty coefficient (UC) was applied to measure the association between pairs of variables
142 (e.g., gender and other non-continuous variables). UC represents a percent reduction in error when
143 predicting dependent variable from independent variable. When UC is 0, the independent variable lacks
144 information to predict the dependent variable (16). The association of variables that were statistically
145 significant with outcome by Fisher exact test were reported. The UC association analysis excluded
146 missing values.

147 *Prediction of Treatment Outcome*

148 Categorical covariates with more than 2 levels were encoded into binary variables. Features
149 were standardized by min-max normalization. Features with highest UC with treatment outcome were
150 selected for each drug-sensitive and drug-resistant TB subgroup respectively. Seventy percent of data
151 were selected as a training dataset and the remainder were used as an independent testing dataset.
152 Machine learning algorithms, Logistic Regression, Random Forest, Support Vector Machine, and
153 XGBoost, were trained on the training set to predict treatment outcome. Repeated grid search 3-fold cross
154 validation was used for parameter tuning. Features with more than 75% missing data or less than 1%
155 variation were removed. Synthetic Minority Oversampling Technique (SMOTE) was used to oversample
156 the minority class, of “died”. Area Under the Receiver Operating Characteristics Curve (AUROC) and
157 Area Under the Precision and Recall Curve (PRAUC) were used to evaluate model performance on the
158 testing dataset. Feature importance was generated for the model with highest AUROC and PRAUC.
159 Model coefficients and Shapely Additive Explanations (ShAP) indicate feature importance for Logistic
160 Regression and tree-based models respectively.

161 ***Risk Score***

162 The predicted probabilities of poor treatment outcome from models for each drug-sensitive or
163 drug-resistant TB patient subgroup were calibrated as a risk score ranging from 0 to 100 corresponding to
164 confidence level of the model for treatment outcome. The risk scores of the training dataset were
165 calculated by out-of-bag cross-validation while the risk scores of the testing dataset were calculated by
166 the best model trained on the training dataset.

167 ***Statistical Software***

168 The data cleaning, inferential statistics, and association analysis were done using R statistical
169 software version 4.0.2 (RStudio version 1.2.5033) (17) with packages fastDummies, dplyr, ggplot2,
170 finalfit. The predictive modeling was done using Python 3 (18) with packages numpy, pandas, matplotlib,
171 sklearn, random, and xgboost.

172 **RESULTS**

173 ***Demographics of identified cohort stratified by drug-resistance***

174 Given our focus on predicting the outcome within the specific drug resistance subtypes and
175 ensuring that the identified factors are generalizable, we first assessed the resistance status of the selected
176 cohort of patients with regards to outcome. The treatment success rate for DS-TB cases was high (88.6%),
177 while it was lower for MDR-TB treatment (79.5%), and even lower for XDR-TB treatment (73.3%)
178 (Table S2). While the rates of treatment success outcome correlate with the ones from the WHO report in
179 2019 (1), the absolute rates are distinct because only the most definitive outcomes of died and cured
180 patients were considered in this analysis while the WHO report considered died, cured, treatment failed,
181 and lost-to-follow-up patients. Importantly, there was a significant difference of success rates among
182 these drug-sensitive and drug-resistant subgroups ($p < 0.0001$, Chi-Square Test) supporting the strategy of
183 stratifying by resistance subgroup in the machine learning analyses.

184 ***Association Analysis***

185 With the diverse demographic, clinical, radiological, microbiological, and genomic domains of
186 data available in TB Portals, we studied the most significant inter-domain relationships using association
187 analysis. The goal was to explore these inter-domain relationships prior to including these variables in the
188 modeling. This approach attempts to identify any potential biases as well as highlight significant
189 associations between types of data (e.g., pathogen genomics and imaging) taking advantage of the patient-
190 centric, multi-domain nature of the TB portals resource. As these domains of data have not been
191 combined previously, there was an opportunity for discovery as well as comparing known associations
192 from prior studies.

193 Among the top variables associated with outcome in DS, MDR, and XDR TB were imaging
194 related pathologies like nodule, cavity, and fibrosis as well as overall area of abnormality. Moreover,
195 non-imaging features including social determinants of health like employment and education status are
196 present across resistance types. The complete list of these factors is in Table 2. The heatmaps of the most
197 important variables associated with treatment outcome in MDR-TB and XDR-TB groups are visualized

198 (Figure 1). In the XDR-TB group, UC with outcome were between 0.090 and 0.192. The top associated
199 feature *Pleural effusion percent of hemithorax* was in the same cluster with outcome. In MDR-TB group,
200 UC with outcome were between 0.07 and 0.237. For full code and complete overview of the association
201 analysis pipeline, please refer to our GitHub repository ([https://github.com/niaid/tb-portals-association-](https://github.com/niaid/tb-portals-association-and-prediction)
202 [and-prediction](https://github.com/niaid/tb-portals-association-and-prediction)).

203 ***Modeling and Prediction***

204 We were able to obtain reasonable predictive capacity across a variety of algorithms; however,
205 we did observe certain models performing better depending upon the specific resistance subgroup. The
206 performance of different predictive models was compared (Table 3 and Figure 2). XGBoost outperformed
207 other models in the DS-TB patient subgroup (AUROC = 0.8188, PRAUC = 0.5836); Logistic Regression
208 with Lasso regularization outperformed other models in the MDR-TB patient subgroup (AUROC =
209 0.8353, PRAUC = 0.6037); in XDR-TB patient subgroup, Random Forest outperformed other models
210 (AUROC = 0.8665, PRAUC = 0.7260). The top features associated with treatment outcome with the
211 highest feature importance were selected in DS-TB, MDR-TB, XDR-TB patient subgroups (Figure 3).
212 This could explain why specific models performed better in particular resistance groups since the
213 variables may show non-linear or linear dependencies with the outcome of interest. The boxplot of top
214 features was generated for each drug-sensitive and drug-resistant subgroup (Figure S1). The frequency
215 tables and univariate odds ratios between top features with treatment outcome were listed in Table S5.

216 ***Risk Scores***

217 The risk scores were derived from the predictive probabilities of all-cause mortality from the best
218 predictive model for each subgroup respectively. There is a statistically significant difference in risk
219 scores between cured and died outcome for TB patients by Two-sample t-test (Figure S2). For most cases,
220 risks aligned with the expected outcome (Figure S3). In public health and clinical practice, it is important
221 to identify and highlight situations where predicted risks do not align with expected outcome due to
222 potential biases in data collection or other causes. We examined these on a case-by-case basis to account

223 for potential impacts of social determinants of health or other factors on predicted outcome. Our risk
224 analysis provides new understanding of the mechanisms of risk that can inform future clinical study and
225 healthcare policy specific to important subtypes of TB (e.g., drug-sensitive versus drug-resistant).

226 **DISCUSSION**

227 TB Portals presents the largest, publicly available real-world dataset of tuberculosis cases to
228 understand the important clinical and socio-economic features of the case that predict treatment outcome.
229 We analyzed the data from TB Portals and identified factors spanning the distinct domains of information
230 including clinical and socio-economic features to discover associations and interactions between these
231 domains that might impact the probability of a poor treatment outcome. We confirmed prior risk factors
232 identified in earlier studies while extending the analysis to include stratification of drug resistance status,
233 which allows for a more accurate understanding of the common or divergent factors across subgroups. Six
234 features were associated with treatment outcome in all three subgroups: BMI, onset age, employment,
235 education (college or higher), smear-negative microscopy, and overall percent of abnormal lung volume
236 as determined by a radiologist in X-ray images (less than 50 percent). We confirmed in a large, stratified
237 analysis several previously reported clinical as well as socio-economic factors like BMI, age, microscopy,
238 and imaging as important towards the probability of successful treatment outcome.

239 BMI was found to be the most predictive feature inversely associated with all-cause mortality,
240 consistent with population-based cohort studies by Hsien-Ho, et al. (19). Employment is usually
241 associated with better healthcare outcomes as it is indicative of higher income, health insurance coverage,
242 and socio-economic status. College or higher education as well as employment status is inversely
243 associated with all-cause mortality. Onset age is a predictor of all-cause mortality consistent with an
244 earlier study finding that older age associated with treatment failure (20). Smear microscopy is commonly
245 used as a part of the primary diagnostic protocol and monitoring of treatment efficacy of TB in many low-
246 or medium-income countries. In our analysis, we found that negative smear microscopy was associated
247 with successful treatment, which is consistent with clinical guidelines indicating treatment efficacy.
248 Aside from critical demographic and social determinants of health factors, our analysis identified several

249 radiological (overall abnormal lung volume, presences of nodules, upper lobe involvement, etc.) as well
250 as pathogen genomic features (octal spoligotype and genomic variants in select resistance genes). These
251 distinct domains of data have not frequently been analyzed together; and we believe it is necessary to
252 consider them together to understand the holistic nature of the disease and treatment.

253 There are some caveats that need to be carefully considered. The temporal dynamics in a TB case
254 are complex and require an expanded dataset with information captured at distinct timepoints of the
255 treatment, which is also available from TB Portals (<https://analytic.tbportals.niaid.nih.gov/index.html>).
256 The aforementioned API as well as data sharing website describing the data model
257 (<https://datasharing.tbportals.niaid.nih.gov/#/about-the-data>) present an opportunity to assess each case
258 temporally through relational organization of the timing of important events (imaging, treatment, culture,
259 and microscopy) in number of days from the earliest registration date in the record. The presented results
260 in this study focus on the case level summary and we plan to expand these analyses in the future to
261 account for temporal dynamics in the case. For example, we summarized information about radiologist
262 reported lung pathology that is captured across one or more images and involving segments within the
263 lung into categories (e.g., "Upper sextant - Yes" or "No", "Middle or Lower sextant - Yes" or "No" to
264 indicate the combination of features from available image data). We focused on understanding the factors
265 associated with cured and died outcomes to examine the predictors of the most definitive outcomes;
266 however, we plan to follow up with additional analyses that examine other outcomes since competing
267 risks might overlap between end points such as all-cause-mortality or treatment failure.

268 TB portals is a real-world data resource focusing on the most challenging TB cases and the
269 programmatic priorities of participating clinical centers and so it is enriched in highly-drug resistant cases
270 as a natural history study. The result from any machine learning and association study should not be
271 considered for making actionable clinical decisions until a clinical study or clinical trial demonstrates
272 efficacy of an intervention. Understanding the importance of bias or areas of failure in model performance
273 is also important for the application of machine learning in public health. We explored some of these
274 dynamics using Shapely Additive Explanations (SHAP) force plots (Figure S3) in a case-by-case basis for

275 outlier or unexpected predictions to highlight situations in which the predictions from our models conflict
276 with expected observations. Given that these are real-world data, our findings need to be interpreted
277 cautiously due to the potential of confounding from observed or unobserved variables; however, the
278 analysis of cross-domain information can provide valuable insight for future translational medicine efforts
279 and study. We plan to periodically update this analysis with new information and cases as data become
280 available from the rapid growth of the TB Portals resource.

281

282 **LIST OF ABBREVIATIONS**

Abbreviation	Full Word
DR	Drug resistant
BMI	Body mass index
NIAID	National Institute of Allergy and Infectious Diseases
NIH	National Institute of Health
TB	Tuberculosis
MDR-TB	Multidrug resistant TB
XDR-TB	Extensively drug-resistant TB
WHO	World Health Organization
TB DEPOT	Tuberculosis Data Exploration Portal
DUA	Data usage agreement
DS	Drug-sensitive
ANOVA	Analysis of variance
UC	Uncertainty coefficient
SMOTE	Synthetic Minority Oversampling Technique
AUROC	Area Under the Receiver Operating Characteristics Curve
PRAUC	Area Under the Precision and Recall Curve

SHAP	Shapely Additive Explanations
------	-------------------------------

283

284 **DECLARATIONS**

285 **ETHICS APPROVAL AND CONSENT TO PARTICIPATE**

286 The data is provided by the TB Portals program stripped of all identifiers as a de-identified dataset
287 according to their data use agreement ([https://tbportals.niaid.nih.gov/pdf/TB-Portals-Data-Use-](https://tbportals.niaid.nih.gov/pdf/TB-Portals-Data-Use-Agreement.pdf)
288 [Agreement.pdf](https://tbportals.niaid.nih.gov/pdf/TB-Portals-Data-Use-Agreement.pdf)). All methods were carried out in accordance with relevant guidelines and regulations.

289 **CONSENT FOR PUBLICATION**

290 Not applicable

291 **AVAILABILITY OF DATA AND MATERIALS**

292 The deidentified datasets are publicly available in the Tuberculosis Data Exploration Portal (TB DEPOT),
293 <https://depot.tbportals.niaid.nih.gov/#/home>.

294 **COMPETING INTERESTS**

295 The authors declare that they have no competing interests

296 **FUNDING**

297 This project has been funded in part with Federal funds from the National Institute of Allergy and
298 Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services
299 under BCBB Support Services Contract HHSN316201300006W/HHSN27200002 to MSC, Inc.

300 **AUTHORS' CONTRIBUTIONS**

301 QW applied machine learning algorithms to predict the outcome of tuberculosis patients and wrote the
302 manuscript. JG conducted association analysis to identify important factors associated with treatment
303 outcome. AG initialized the project, helped interpreting the results clinically and writing the conclusion.
304 GR edited the manuscript and assisted with submission. MQ, DH, AR provided guidance on the project
305 and revised the manuscript. All authors read and approved the final manuscript.

306 **ACKNOWLEDGEMENTS**

307 We thank Kurt Wollenberg, PhD from NIAID for his suggestions and support on this project; Alina
308 Grinev, MD, MSBM from NIAID for managing this project; Alyssa Long, BS from NIAID for accessing
309 the TB Portals database.

310 REFERENCES

- 311 1. WHO. Global tuberculosis report 2019 2019 [Available from:
312 <https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-report-2019>.
- 313 2. Fuady A, Houweling TAJ, Richardus JH. COVID-19 and Tuberculosis-Related
314 Catastrophic Costs. *Am J Trop Med Hyg.* 2020.
- 315 3. Glaziou P. Predicted impact of the COVID-19 pandemic on global tuberculosis deaths in
316 2020. medRxiv. 2020.
- 317 4. Auer C, Mazitov R, Makhmudov A, Pirmahmadzoda B, Skrahina A, Dobre A, et al.
318 Factors contributing to drug-resistant tuberculosis treatment outcome in five countries in the
319 Eastern Europe and Central Asia region. *Monaldi Arch Chest Dis.* 2020;90(1).
- 320 5. Van LH, Phu PT, Vinh DN, Son VT, Hanh NT, Nhat LTH, et al. Risk factors for poor
321 treatment outcomes of 2266 multidrug-resistant tuberculosis cases in Ho Chi Minh City: a
322 retrospective study. *BMC Infect Dis.* 2020;20(1):164.
- 323 6. Sauer CM, Sasson D, Paik KE, McCague N, Celi LA, Sanchez Fernandez I, et al. Feature
324 selection and prediction of treatment failure in tuberculosis. *PLoS One.* 2018;13(11):e0207491.
- 325 7. Rosenfeld G, Gabrielian A, Wang Q, Gu J, Hurt DE, Long A, et al. Radiologist
326 observations of computed tomography (CT) images predict treatment outcome in TB Portals, a
327 real-world database of tuberculosis (TB) cases. *PLoS One.* 2021;16(3):e0247906.
- 328 8. Gabrielian A, Engle E, Harris M, Wollenberg K, Juarez-Espinosa O, Glogowski A, et al.
329 TB DEPOT (Data Exploration Portal): A multi-domain tuberculosis data analysis resource. *PLoS*
330 *One.* 2019;14(5):e0217410.
- 331 9. Group TNTRSPW. NIAID Strategic Plan for Tuberculosis Research. In: NIAID, editor.
332 2018. p. 17-8.
- 333 10. Strategy OoDS. NIH Strategic Plan for Data Science. In: ODSS, editor. 2018. p. 14-5.
- 334 11. Long A, Glogowski A, Meppiel M, De Vito L, Engle E, Harris M, et al. The technology
335 behind TB DEPOT: a novel public analytics platform integrating tuberculosis clinical, genomic,
336 and radiological data for visual and statistical exploration. *J Am Med Inform Assoc.*
337 2021;28(1):71-9.
- 338 12. Rosenthal A, Gabrielian A, Engle E, Hurt DE, Alexandru S, Crudu V, et al. The TB
339 Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data
340 Sharing and Analysis. *J Clin Microbiol.* 2017;55(11):3267-82.
- 341 13. Engle E, Gabrielian A, Long A, Hurt DE, Rosenthal A. Performance of Qure.ai automatic
342 classifiers against a large annotated database of patients with diverse forms of tuberculosis. *PLoS*
343 *One.* 2020;15(1):e0224445.
- 344 14. Gabrielian A, Engle E, Harris M, Wollenberg K, Glogowski A, Long A, et al.
345 Comparative analysis of genomic variability for drug-resistant strains of *Mycobacterium*
346 tuberculosis: The special case of Belarus. *Infect Genet Evol.* 2020;78:104137.
- 347 15. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L.
348 Erlbaum Associates; 1988. xxi, 567 p. p.

349 16. De Muth JE. Basic statistics and pharmaceutical statistical applications. Third edition. ed.
350 Boca Raton: CRC Press, Taylor & Francis Group; 2014. xxvi, 821 pages p.
351 17. RStudio. RStudio: Integrated Development for R. RStudio. Boston, MA2020.
352 18. Van Rossum GD, F.L. Python 3 Reference Manual. Scotts Valley, CA2009.
353 19. Lin HH, Wu CY, Wang CH, Fu H, Lonroth K, Chang YC, et al. Association of Obesity,
354 Diabetes, and Risk of Tuberculosis: Two Population-Based Cohorts. Clin Infect Dis.
355 2018;66(5):699-705.
356 20. Mulu W, Mekonnen D, Yimer M, Admassu A, Abera B. Risk factors for multidrug
357 resistant tuberculosis patients in Amhara National Regional State. Afr Health Sci.
358 2015;15(2):368-77.
359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377 **FIGURES AND TABLES**

378 **Table 1 – Demographics of Different Drug-Sensitive and Drug-Resistant Subgroups**

	MDR non XDR		Sensitive		XDR	
	Cured (N=702)	Died (N=181)	Cured (N=265)	Died (N=34)	Cured (N=288)	Died (N=105)
gender						
Male	470 (67.0%)	153 (84.5%)	179 (67.5%)	25 (73.5%)	194 (67.4%)	78 (74.3%)
Non-Male	232 (33.0%)	28 (15.5%)	86 (32.5%)	9 (26.5%)	94 (32.6%)	27 (25.7%)
bmi						
Mean (SD)	21.0 (3.47)	19.6 (3.35)	21.1 (3.68)	18.1 (3.69)	21.7 (3.70)	19.3 (3.40)
Median [Min, Max]	20.6 [12.8, 40.3]	19.6 [13.2, 34.8]	21.0 [10.8, 34.0]	17.5 [12.1, 27.7]	21.1 [13.4, 38.6]	18.9 [11.8, 29.4]
age_of_onset						
Mean (SD)	38.9 (13.4)	46.6 (12.6)	44.3 (16.6)	52.3 (16.7)	38.7 (12.9)	43.8 (12.3)
Median [Min, Max]	38.0 [11.0, 81.0]	45.0 [18.0, 83.0]	43.0 [13.0, 85.0]	54.0 [20.0, 79.0]	37.5 [15.0, 69.0]	42.0 [21.0, 75.0]
employment						
Disabled	39 (5.6%)	22 (12.2%)	6 (2.3%)	2 (5.9%)	42 (14.6%)	24 (22.9%)
Employed	271 (38.6%)	14 (7.7%)	102 (38.5%)	5 (14.7%)	96 (33.3%)	14 (13.3%)
Homemaker or Self-employed	5 (0.7%)	3 (1.7%)	16 (6.0%)	2 (5.9%)	1 (0.3%)	0 (0%)
Not Reported	40 (5.7%)	6 (3.3%)	23 (8.7%)	0 (0%)	10 (3.5%)	1 (1.0%)
Retired	36 (5.1%)	19 (10.5%)	30 (11.3%)	5 (14.7%)	23 (8.0%)	8 (7.6%)
Student	36 (5.1%)	0 (0%)	10 (3.8%)	0 (0%)	18 (6.2%)	0 (0%)
Unemployed	275 (39.2%)	115 (63.5%)	78 (29.4%)	20 (58.8%)	98 (34.0%)	55 (52.4%)
Unofficially employed	0 (0%)	2 (1.1%)	0 (0%)	0 (0%)	0 (0%)	3 (2.9%)
education						
Basic school or lower	77 (11.0%)	49 (27.1%)	73 (27.5%)	18 (52.9%)	38 (13.2%)	28 (26.7%)
College or higher	284 (40.5%)	49 (27.1%)	103 (38.9%)	2 (5.9%)	141 (49.0%)	32 (30.5%)
Complete school (a-level, gymnasium)	139 (19.8%)	63 (34.8%)	74 (27.9%)	13 (38.2%)	61 (21.2%)	32 (30.5%)
Not Reported	202 (28.8%)	20 (11.0%)	15 (5.7%)	1 (2.9%)	48 (16.7%)	13 (12.4%)

379
 380 The demographics of the patients selected for study in this analysis are shown stratified by drug-resistance subgroup.
 381 Categorical variables (e.g., gender, employment, education) are displayed with the number of patient cases with
 382 percentages of total for each drug-resistance subgroup and outcome grouping. Numerical variables (e.g.,
 383 age_of_onset and bmi) are shown with the means [standard deviations] as well as medians [minimums and
 384 maximums] for each subgroup and outcome grouping.

385
 386 **Table 2 – Top variables associated with Outcome by UC**

DS-TB	MDR-TB	XDR-TB
First lung cavity size (0.167)	Overall percent of abnormal volume (0.237)	Pleural effusion percent of hemithorax involved (0.192)
First cure nodule (0.125)	Gene name (0.199)	Octal spoligotype (0.188)
First cure fibrosis (0.063)	Social risk factors (0.158)	Social risk factor (0.168)
Education (0.063)	First affected segments (0.141)	Any multiple nodules exists (0.113)
Infiltrate low ground glass density (0.063)	Employment (0.117)	First microscopy (0.100)

387 The top five features associated with outcome by drug resistance subgroup is shown in Table 2. The values next to
388 the features are uncertainty coefficient (UC), which can vary from 0 to 1. When UC is 0, the feature is of no value in
389 predicting the outcome.

390

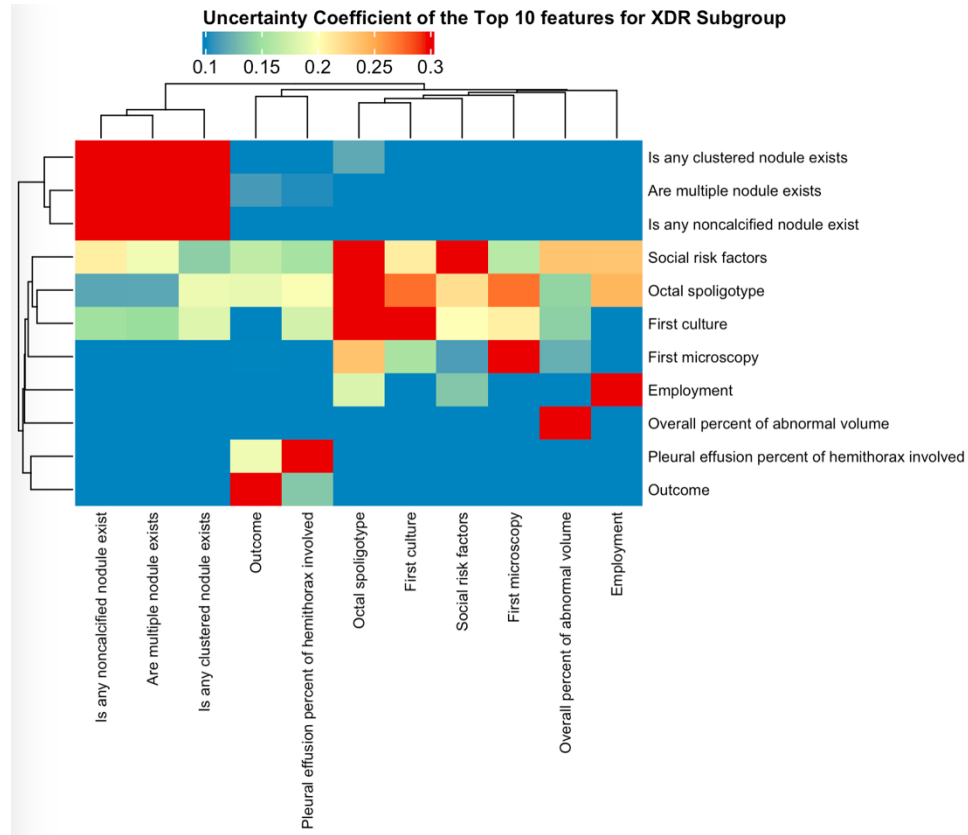
391

Table 3 – Comparison of model performance

Models	DS-TB (N=299)		MDR-TB (N=883)		XDR-TB (N=393)	
	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC
Logistic Regression	0.8688	0.3758	0.8353	0.6037	0.8438	0.7234
Random Forest	0.8050	0.4144	0.8138	0.5562	0.8665	0.7260
Support Vector Machine	0.8288	0.3258	0.8180	0.5572	0.8020	0.6523
XGBoost	0.8188	0.5836	0.8198	0.5382	0.8230	0.6846

392 The performance of different predictive models in Area Under the Receiver Operating Characteristics Curve
393 (AUROC) and Area Under the Precision and Recall Curve (PRAUC) are shown stratified by drug-resistance
394 subgroup. The models with the highest PRAUC are highlighted.

395

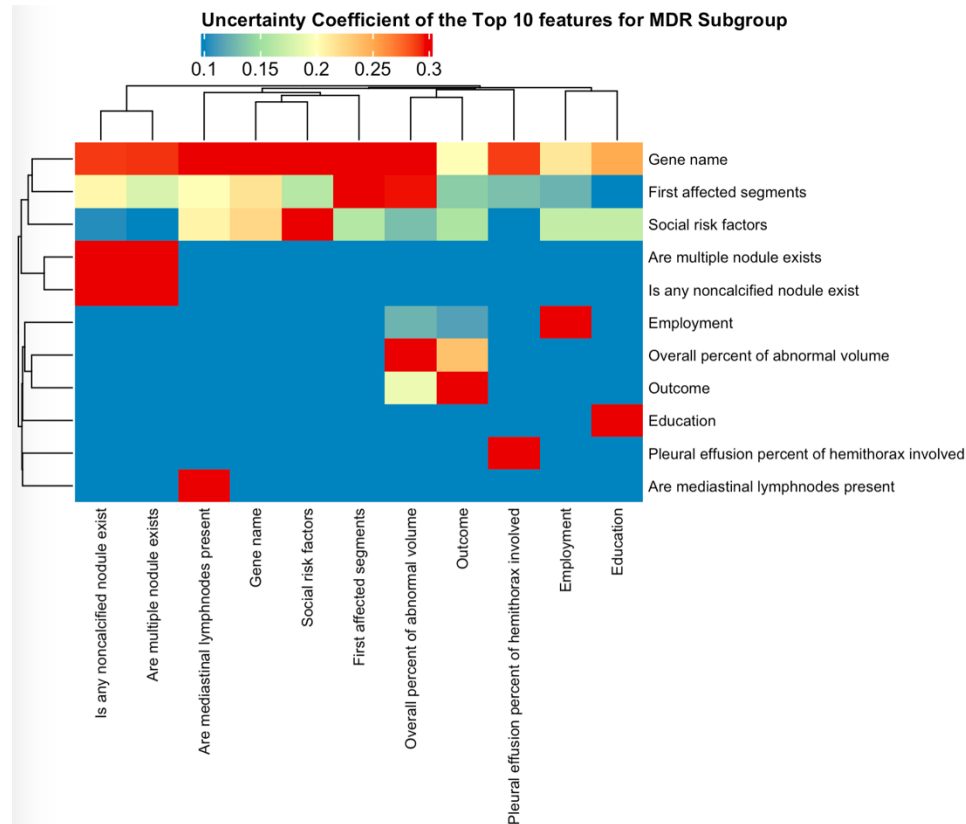


396

397

398

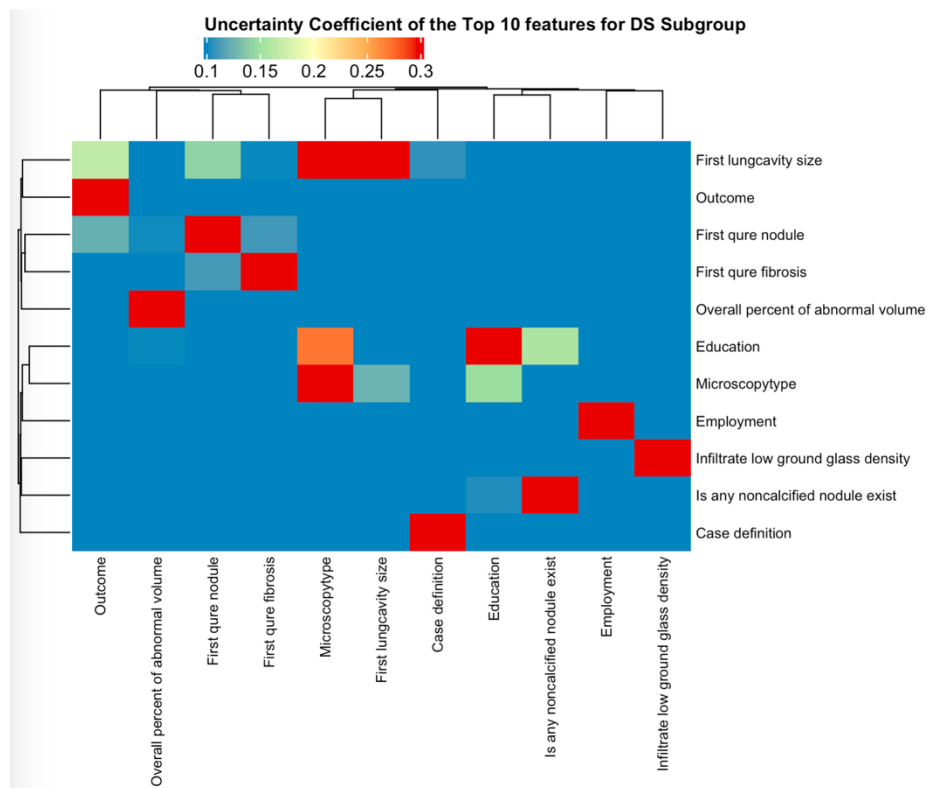
(a) XDR-TB



399

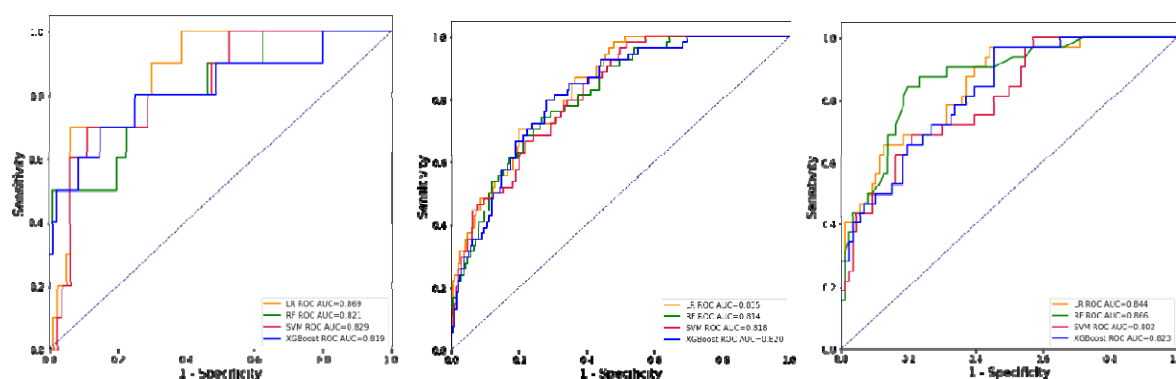
400

(b) MDR-TB



401

402 (c) DS-TB
 403 Figure 1 – Heatmap of Top Ten Features associated with Treatment Outcome. The top ten features associated with
 404 Treatment Outcome are shown in relation to each other with a higher uncertainty coefficient (UC) indicating a
 405 stronger association between sets of features. In A), the relationship by UC for the top 10 features for XDR-TB (N
 406 = 430) are shown while in B), the relationship is shown for MDR-TB (N = 1019) and C), the relationship for DS-TB
 407 (N = 811). In both, the sets of features are clustered by hierarchical clustering to show clusters of features showing
 408 associations.
 409



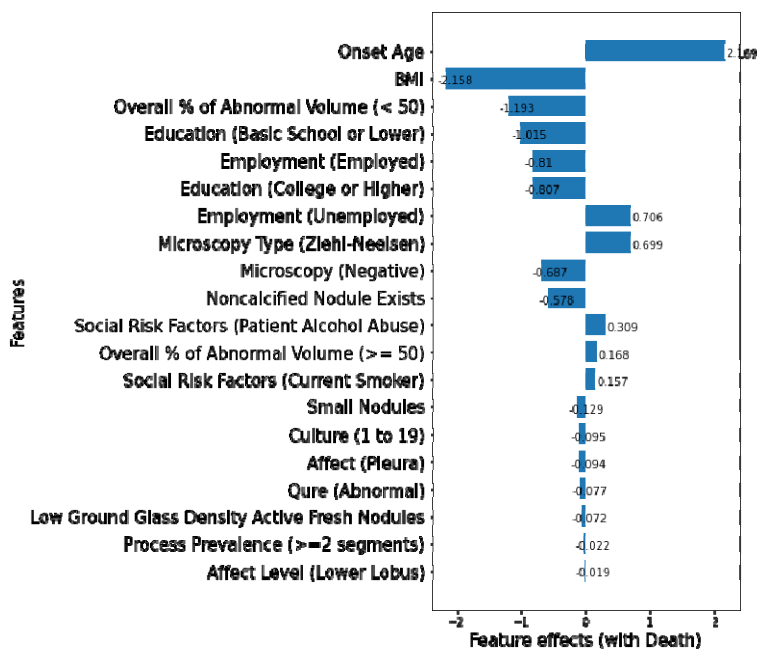
410 (a) DS-TB (N=299)

411 (b) MDR-TB (N=883)

412 (c) XDR-TB (N=393)

413 Figure 2 – ROC Curves of Drug-Sensitive and Drug-Resistant Subgroups

414 The receiver operating characteristic (ROC) Curves of different predictive models are shown stratified by drug-
 415 resistance subgroup. The Area Under the Receiver Operating Characteristics Curve (AUROC) for a given curve is
 416 the area beneath it. The AUROC values are listed in the legends.
 417

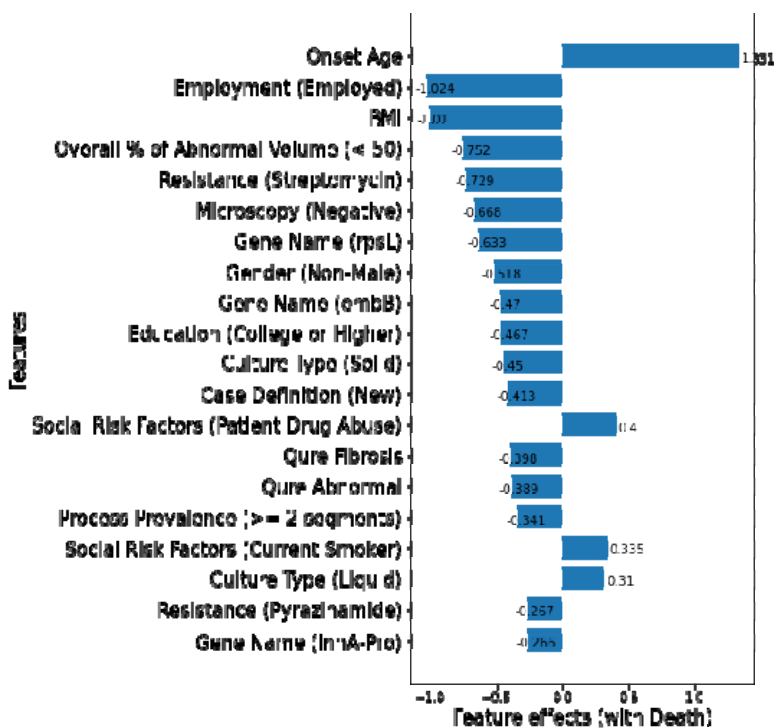


418

419

420

(a) XGBoost on DS-TB patients (N=299)

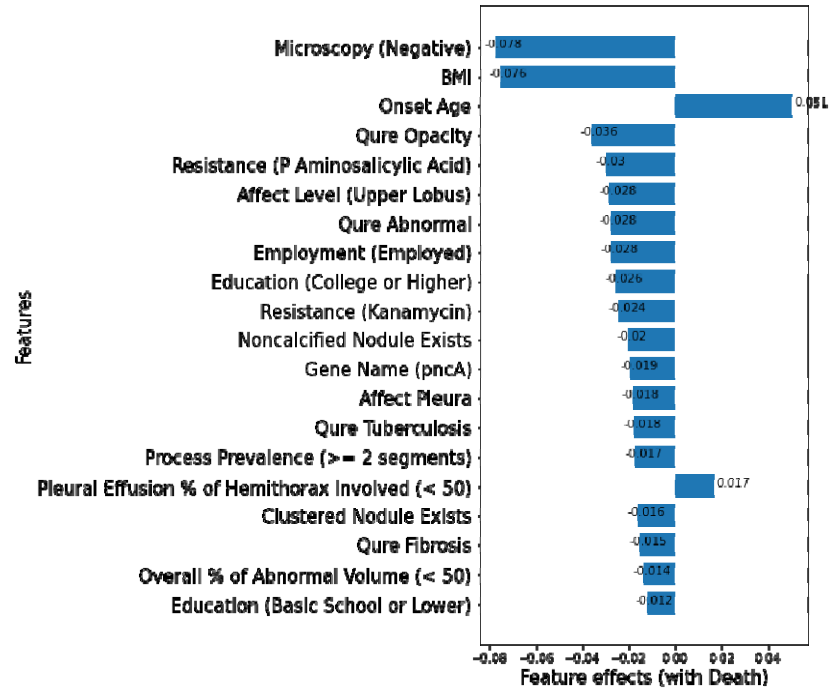


421

422

423

(b) Logistic Regression on MDR-TB patients (N=883)



424

425

(c) Random Forest on XDR-TB Patients (N=393)

426

Figure 3 – Feature Importance of top 20 features from the Best Models

427

For Logistic Regression, the plot shows that normalized features with positive feature effects increase the odds of outcome of death compared to cured whereas normalized features with negative feature effects show the opposite.

428

429

For tree-based models, the plot shows Shapley values and features with larger absolute Shapley values are more

430

important contributors to the prediction.

431