HHS Public Access

Author manuscript

Am J Epidemiol. Author manuscript; available in PMC 2024 January 16.

Published in final edited form as:

Am J Epidemiol. 2022 October 20; 191(11): 1936–1943. doi:10.1093/aje/kwac117.

Using Machine Learning Techniques and National Tuberculosis Surveillance Data to Predict Excess Growth in Genotyped Tuberculosis Clusters

Sandy P. Althomsons[*],
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Kathryn Winglee,
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Charles M. Heilig,
Center for Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Sarah Talarico,
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Benjamin Silk,
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Jonathan Wortham,
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Andrew N. Hill,
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

[*]Correspondence to Sandy Althomsons, Division of TB Elimination, CDC, 1600 Clifton Road, NE MS US 12-4 Atlanta, GA 30333 (soa4@cdc.gov).
S.P.A. and K.W. contributed equally as first authors.

The data for this work contain information abstracted from the national tuberculosis (TB) case report form called the Report of Verified Case of Tuberculosis (OMB No. 0920-0026). These data have been reported voluntarily to the Centers for Disease Control and Prevention (CDC) by state and local health departments, and are protected under the Assurance of Confidentiality (Sections 306 and 308(d) of the Public Health Service Act, 42 U.S.C. 242k and 242m(d)), which prevents disclosure of any information that could be used to directly or indirectly identify patients. For more information, see the CDC/ATSDR Policy on Releasing and Sharing Data (at http://www.cdc.gov/maso/Policy/ReleasingData.pdf). Researchers seeking access to data may apply to analyze National TB Surveillance System data at CDC headquarters by contacting TBInfo@cdc.gov. All R code is available at https://github.com/CDCgov/Predicting_TB_cluster_growth.

We plan to present this work as a poster at the 2022 Joint Statistical Meeting, Washington, DC, August 6–11, 2022.

The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflict of interest: none declared.

Thomas R. Navin
Division of TB Elimination, National Center for HIV, Viral Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States

Abstract

The early identification of clusters of persons with tuberculosis (TB) that will grow to become outbreaks creates an opportunity for intervention in preventing future TB cases. We used surveillance data (2009–2018) from the United States, statistically derived definitions of unexpected growth, and machine-learning techniques to predict which clusters of genotype-matched TB cases are most likely to continue accumulating cases above expected growth within a 1-year follow-up period. We developed a model to predict which clusters are likely to grow on a training and testing data set that was generalizable to a validation data set. Our model showed that characteristics of clusters were more important than the social, demographic, and clinical characteristics of the patients in those clusters. For instance, the time between cases before unexpected growth was identified as the most important of our predictors. A faster accumulation of cases increased the probability of excess growth being predicted during the follow-up period. We have demonstrated that combining the characteristics of clusters and cases with machine learning can add to existing tools to help prioritize which clusters may benefit most from public health interventions. For example, consideration of an entire cluster, not only an individual patient, may assist in interrupting ongoing transmission.

Keywords

cluster growth; machine learning; surveillance data; tuberculosis

An outbreak of tuberculosis (TB) can be resource-intensive and time-consuming in terms of public health response (1-4). Preventing an outbreak reduces morbidity and mortality and might be more cost-effective than controlling one, particularly with limited public health resources (5). Prioritizing interventions on clusters of TB cases that are most likely to become outbreaks will have the greatest likelihood of preventing future cases of TB. Here, we describe continuing efforts to detect and predict TB outbreaks from routine national TB surveillance data by identifying which TB clusters are most likely to accumulate additional cases.

The feasibility of predicting TB outbreaks arising from incident TB genotype clusters—or clusters of TB cases where the beginning of a possible outbreak can be identified—was previously demonstrated (6). That study used decision trees to determine which characteristics of the first 3 cases best predicted which incident clusters were most likely to become an outbreak. However, that study was limited to incident clusters. Subsequently, we developed a statistical method to detect outbreaks arising from prevalent TB genotype clusters—clusters of TB cases where the first case in the cluster cannot be determined because cases with that genotype have been reported in a county for at least 2 years (7). This method is a negative binomial hurdle (NBH) model that identifies unexpected growth, a statistically significant number of cases above what is expected based on the baseline rate

of previously reported cases. However, we did not analyze characteristics associated with cluster growth or apply this second algorithm to incident clusters.

This investigation combines the previous 2 methods: first, by identifying unexpected growth among all (incident and prevalent) clusters of genotype-matched TB cases in the same county, and second, by using machine-learning (ML) techniques on case and cluster characteristics to predict which clusters are most likely to have continued growth above the baseline rate (6, 7). Whereas most epidemiologic analyses focus on describing and understanding the relationships between exposures and outcomes, ML techniques are designed to look for patterns to make predictions about new data, while allowing interactions and nonlinear relationships (8, 9). ML prioritizes improving model performance above determining the relationship between outcomes and predictors (often called features in ML publications). ML has been used to predict health outcomes, including predicting filariasis from social risk factors (10), identifying type 2 diabetes from electronic health records (11), and detecting cardiovascular risk from routine clinical data (12). Regarding TB, several studies used algorithms to diagnose TB cases (13-15). Others tried to predict transmission among reported TB cases based on *Mycobacterium tuberculosis* strain typing (16) or among contacts of TB patients based on demographic and clinical characteristics (17). Others used geospatial patterns to predict future transmission (18). Another study predicted whether a newly diagnosed case is associated with recent transmission (19). However, these studies were at the patient level, predicting an individual patient's disease state based on demographic, environmental, or clinical factors. To our knowledge, no other investigations have predicted future TB cases based on characteristics of both the TB cluster and the patients within that cluster.

TB surveillance in the United States captures over 200 variables for all reported US TB cases during the past 25 years (20). The current TB cluster detection system in the United States utilizes the count and geography of TB cases that have been genotyped since 2009 (21). Genotyping data was available for nearly all culture-confirmed cases, ranging from 87% in 2009 to 96% in 2018 (20). An alert system was developed based on a log-likelihood ratio derivation of expected cases in a county based on the prevalence of that *M. tuberculosis* genetic strain reported nationally (22). These alerts are reported weekly to state and local health departments and the Centers for Disease Control and Prevention (CDC) and are manually reviewed to prioritize clusters for public health action (21). However, these reviews are time-consuming, and the alert system has not been validated, nor does it consider cluster growth rate. We sought to determine whether patient surveillance data and cluster characteristics could be systematically used to predict additional cluster growth, thus prioritizing clusters that are likely to grow for public health intervention. Here we present an analytical framework using the combination of TB surveillance data, previously developed statistical methods, and ML techniques to predict which TB clusters are at risk of excess growth.

METHODS

Our analysis utilized the US National Tuberculosis Surveillance System (NTSS) and the National Tuberculosis Genotyping Service data sets, described elsewhere (20), and included

all genotyped culture-positive TB cases reported from the 50 states and the District of Columbia during 2009–2018. We defined a cluster as having ≥3 TB cases with a matching genotype (spoligotype and 24-locus mycobacterial interspersed repetitive unit variable-number tandem repeat analysis) reported in the same county or county-equivalent jurisdiction (7). Data management was performed in SAS, version 9.4 (SAS Institute, Inc., Cary, North Carolina) (23). Case counts for each cluster were aggregated by calendar-year quarters and exported to R (R Foundation for Statistical Computing, Vienna, Austria) (24) to identify unexpected growth by fitting an NBH model to each cluster. An 8-quarter baseline represents expected cases before the detection of unexpected growth. For each cluster, the NBH model sets the threshold for unexpected growth during the next quarter as the 95th percentile of the baseline. When the number of cases exceeds that threshold, unexpected growth is detected. Figure 1 shows a hypothetical example; details are described elsewhere (7).

Our data set included all clusters detected with unexpected growth during 2011–2017. Using an 8-quarter baseline and 4-quarter follow-up, this data set was based on all cases during 2009–2018. We constructed a training and testing data set (test data set) as the cohort of clusters in which the first instance of unexpected growth occurred between January 2011 and December 2015. The validation data set was a separate cohort containing clusters in which the first instance of unexpected growth occurred between January 2016 and December 2017. Clusters in which the third case was reported by 2015 (test data set) or 2017 (validation data set) were included in the analytical data sets.

We defined the outcome using the accrual of excess cases within a 4-quarter follow-up period (Figure 1). Excess cases were defined as the number of cases above what was expected, which was calculated by subtracting the baseline from the number of cases in each quarter after the detection of unexpected growth. Each quarter's excess cases are summed over the follow-up period. If the number of accumulated excess cases in the follow-up period exceeded zero, the result was more cases than expected. These clusters were labeled excess growth. If the accumulated excess cases in the follow-up period were zero or less, the number of cases were within the cluster's baseline. These were labeled expected growth.

In selecting predictors, we first considered which cases in the baseline were most likely to contribute to the outcome. The inclusion criteria for cases in the predictors was defined by the time prior to unexpected growth. We tested 4 time frames: 1 quarter (the quarter of unexpected growth, or QUG), 2 quarters (QUG plus 1 preceding quarter), 4 quarters (QUG plus 3 preceding quarters), and 9 quarters (QUG plus 8 preceding quarters).

Predictors included characteristics of patients reported to NTSS (20), including social risk factors: homelessness, incarceration, injection drug use, non-injection drug use, excess alcohol use, and residence of a long-term care facility. We also included patient clinical factors: previous case of TB, sputum smear positivity, cavitary disease, human immunodeficiency virus, and multiple drug resistance. Patient demographic and epidemiologic characteristics were also included: born in the United States, pediatric case (<15 years old), found via targeted testing or contact investigation, reported an epidemiologic link to another case, or attributable to recent transmission (25). Finally, we

included cluster characteristics during the designated time frame: average number of cases per quarter in the baseline, number of cases in a cluster, and difference in time (months) between the first and last case.

We next considered how to incorporate patient characteristics as predictors for our clusters. We tested 4 variations on how to quantify these characteristics, referred to as quantification. The “any” quantification was defined as at least 1 case in the cluster with the characteristic; predictors were given binary values (>1 or none). The “half” quantification was defined when at least half of the cases had the characteristic, assigned binary values. The “percent” quantification defined the percentage of cases in the cluster with that characteristic; predictors had continuous values between 0 and 100. Finally, we included a “mix” quantification, defined as a combination of the above options. For each quantification and time frame, we removed categorical predictors that were positive in <5 clusters in the test data set.

We tested 3 ML methods (tree-based ensembles, support vector machine (SVM), and regularized regression), with a total of 9 variations, using excess growth or expected growth as our outcome for binary classification (Table 1). See Web Appendixes 1-4 (available at https://doi.org/10.1093/aje/kwac117) for details on the ML methods, including hyperparameters used. By varying the options of the algorithm on the time frame of the predictors 4 ways (i.e., 1 quarter, 2 quarters, 4 quarters, 9 quarters), the quantification 4 ways (i.e., any, half, percent, mix), and 9 different ML variations, we compared the prediction results of 144 different models.

To assess the generalizability of each prediction algorithm, we trained the models on the test data set using 5-fold cross-validation, and averaged results over the 5 analytical runs. We limited results to those with a mean sensitivity >0.25 and then selected the model with the highest Youden index, defined as true positive fraction minus false positive fraction. A final model was built using the full test data set, using the parameters identified in the top model. We validated this final model by generating predictions on our validation data set, which was used to confirm that the model was not overfitted (i.e., whether the test data set and validation data set have similar performance metrics) and thus would perform similarly when used on future data. All ML analyses were performed using R (R Foundation for Statistical Computing), version 4.0.2 (24), and details about the code are in the data availability statement in the Acknowledgments. Further details can be found in Web Appendixes 3 and 4. CDC determined this activity to be research that does not involve identifiable human subjects, and institutional review board approval was not required.

RESULTS

During 2009–2017, 64,942 cases were reported to NTSS from the 50 states or District of Columbia that were culture positive for *M. tuberculosis* and had a genotype result, of which 46,624 were unique within the county, meaning they did not have another case with a matching genotype in the same county during the time period. The remaining 18,318 cases comprised 5,220 county-level clusters in total. By 2015, 332 clusters were flagged with unexpected growth by the NBH model (Figure 2). Of the 332 clusters used as the test

data set, 192 (58%) had expected growth and the remaining 140 (42%) had excess growth. During 2016–2017, 13,903 cases were reported to NTSS, and 45 clusters were flagged with unexpected growth. These clusters were used as our validation data set and had similar proportion of clusters (44%) with excess growth (Table 2).

Among the 144 ML models trained on the test data set, 62 had a sensitivity (i.e., recall) greater than 0.25. The 5 ML models with the highest Youden index are listed in Table 3 (all 62 are in Web Table 1). The prediction model with the highest Youden Index (0.165) also had the highest accuracy value (0.608). Since the mean sensitivity was 0.398, many clusters with excess growth were misclassified. This model used the 2-quarters time frame and the "half" quantification as the predictors for a random forest ML model. A summary of the predictors used in the final model can be found in Web Table 1. The final model performed similarly on the validation data set (Table 3), demonstrating the generalizability of the algorithm, suggesting that the model was not overfitted and would have similar performance metrics when applied to future data.

Figure 3 shows the importance of each predictor in our best model. At the top of the figure, the cluster characteristic predictors rank highest, with time between the first and last cases in the cluster prior to unexpected growth the most important predictor, as measured by the Gini index (see detailed description in Web Appendix 4). The patient characteristic predictors had much smaller importance when compared with cluster characteristic predictors. The most important of the patient characteristic predictors were the predictors indicating that at least half of the cases were US-born, were attributable to recent transmission, and reported positive sputum smear results.

We then used accumulated local-effects plots to further explore how each predictor contributes to the prediction (Web Figure 1) (26). For example, as the time between unexpected growth cases increases, the accumulated average probability that excess growth will be predicted decreases. The inverse is true for the next 3 most important predictors: baseline value, cases before baseline among prevalent clusters, and cases in quarter of unexpected growth, all of which are cluster characteristics. (Predictors are described in detail within Web Table 1.) For these predictors, the probability of predicting excess growth tends to increase as the value of these predictors increase, except for local minima or maxima. However, the data are concentrated at the lower values for these variables, so trends for higher values should be interpreted with caution. These results demonstrate that utilization of all data available incorporated into an algorithm can predict which clusters are likely to report excess growth.

DISCUSSION

Our algorithm is designed to predict which clusters will have excess growth within a 4-quarter follow-up period after unexpected growth is identified. We combined detection of unexpected growth from an NBH model, ML classification models, and surveillance data to predict likely TB cluster growth. The more important predictors in our model were cluster, not patient, characteristics. For example, as time decreased between cases within the baseline, excess growth was more likely to be predicted, suggesting that an

early accumulation of cases is predictive of excess growth. This finding is in line with epidemiologic experience suggesting that extensive transmission of TB leads to many cases occurring in a short period of time and can be indicative of an outbreak (6, 27). Our algorithm is modifiable and provides a systematic prioritization of clusters of concern.

TB cases are rare events in the United States. Although TB outbreaks are even more rare, when they do happen, they are resource-intensive and can challenge already resource-limited public health systems. If not brought under control, some outbreaks can last years, with subsequent cases presenting long after transmission occurred and with the possibility of instigating another outbreak. Preventing an outbreak can therefore reduce continued spread of strains that may otherwise become endemic within vulnerable populations. Predicting which clusters are most likely to grow will enable more targeted interventions to interrupt transmission. Still, even our best models make classification errors, meaning we can expect inaccurate predictions and suggesting that cluster growth may only be partially predictable using only patient- and cluster-level data.

Our methods have several limitations. Clusters were defined by geopolitical boundaries of a county (or equivalent), so predictions of future TB cases are also limited to this definition, even though transmission is not limited to borders. Furthermore, we excluded cases that were not culture-confirmed, which often included TB in children; pediatric TB is a known sentinel event for recent transmission. However, without genotype data, it can be unclear to which transmission chain a pediatric case belongs. Additionally, our results do not include whole-genome sequencing, which we expect to lead to overall smaller, more specific clusters. We propose applying these methods to update the prediction algorithm based on clusters defined using whole-genome sequencing data when they become available.

TB control is based on finding and treating cases of TB disease as well as finding and treating latent TB infection. Surveillance data only capture TB disease cases, are limited to patient- and cluster-level information, and do not include efforts to treat latent TB infection. For example, our results do not include a measure of contact investigations conducted around clusters of TB cases. CDC guidelines state that any patient with infectious TB should be interviewed to identify contacts who may have been exposed, that those contacts should be tested for TB infection, and that prophylactic treatment should be offered to those testing positive to prevent development into a TB disease case (1). Often, astute public health officials who convene extensive contact investigations find contacts to cases who have TB disease (and therefore start them on treatment to prevent further transmission) or latent TB infection (and therefore start them on treatment to prevent disease). The strength of contact investigations to stop transmission may play a greater role than individual patient or cluster characteristics to predict excess growth. Often the ability to stop transmission relies on the identification of these contacts, but the work done at the state and local level to prevent transmission is not captured by our models. The results presented here could be improved as additional data are incorporated.

Finally, the delayed diagnosis of a patient with infectious TB can result in extensive transmission. Outbreaks often occur because infectious individuals are in congregate settings, exposing many people in a short period of time, or because an infectious individual

with a large social network transmits to many individuals. If infectious individuals do not present to health-care providers early in their disease, they are unlikely to be the first cases documented in an outbreak, and therefore would not be included among predictor cases during unexpected growth. By the very nature of late identification, our methods would not facilitate preventing transmission by infectious individuals with delayed diagnoses. This may explain why cluster characteristics were more important than patient characteristics in our model. By considering an entire cluster, not only an individual patient, efforts to interrupt transmission may be more effective.

While it may be difficult to predict outbreaks based on characteristics of cases from surveillance data, an early response to unexpected growth of clusters can prevent the continuing transmission of TB. Data systems can aid in prioritizing clusters, but there is no substitute for on-the-ground responses to identify TB cases and contacts of cases to prevent further transmission. We have shown how ML with an extensive surveillance system can be used to predict ongoing transmission and expand our understanding of how clusters grow.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by Centers for Disease Control and Prevention, which provided support through employee salaries for this publication.

The authors acknowledge the state and local health department personnel who collect and report the data used for these analyses. CDC's Division of Tuberculosis Elimination provided funding support through employee salaries for this publication.

Abbreviations:

CDC	Centers for Disease Control and Prevention
ML	machine learning
NBH	negative binomial hurdle
NTSS	National Tuberculosis Surveillance System
TB	tuberculosis

REFERENCES

1. National Tuberculosis Controllers Association, Centers for Disease Control and Prevention. Guidelines for the investigation of contacts of persons with infectious tuberculosis. Recommendations from the National Tuberculosis Controllers Association and CDC. MMWR Recomm Rep. 2005;54(RR-15):1–47.
2. Mitruka K, Oeltmann JE, Ijaz K, et al. Tuberculosis outbreak investigations in the United States, 2002–2008. Emerg Infect Dis. 2011;17(3):425–431. [PubMed: 21392433]

3. Centers for Disease Control and Prevention. Tuberculosis outbreak associated with a homeless shelter—Kane County, Illinois, 2007–2011. MMWR Morb Mortal Wkly Rep. 2012;61(11):186–189. [PubMed: 22437912]
4. Powell KM, VanderEnde DS, Holland DP, et al. Outbreak of drug-resistant *Mycobacterium tuberculosis* among homeless people in Atlanta, Georgia, 2008–2015. Public Health Rep. 2017;132(2):231–240. [PubMed: 28257261]
5. Mindra G, Wortham JM, Haddad MB, et al. Tuberculosis outbreaks in the United States, 2009–2015. Public Health Rep. 2017;132(2):157–163. [PubMed: 28147211]
6. Althomsons SP, Kammerer JS, Shang N, et al. Using routinely reported tuberculosis genotyping and surveillance data to predict tuberculosis outbreaks. PLoS One. 2012;7(11):e48754. [PubMed: 23144956]
7. Althomsons SP, Hill AN, Harrist AV, et al. Statistical method to detect tuberculosis outbreaks among endemic clusters in a low-incidence setting. Emerg Infect Dis Mar. 2018;24(3):573–575.
8. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149–153. [PubMed: 29020316]
9. Bi Q, Goodman KE, Kaminsky J, et al. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188(12):2222–2239. [PubMed: 31509183]
10. Kondeti PK, Ravi K, Mutheneni SR, et al. Applications of machine learning techniques to predict filariasis using socio-economic factors. Epidemiol Infect. 2019;147:e260. [PubMed: 31475670]
11. Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform Jan. 2017;97:120–127.
12. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. [PubMed: 28376093]
13. El-Solh AA, Hsiao CB, Goodnough S, et al. Predicting active pulmonary tuberculosis using an artificial neural network. Chest Oct. 1999;116(4):968–973.
14. Cain KP, McCarthy KD, Heilig CM, et al. An algorithm for tuberculosis screening and diagnosis in people with HIV. N Engl J Med. 2010;362(8):707–716. [PubMed: 20181972]
15. Khan MT, Kaushik AC, Ji L, et al. Artificial neural networks for prediction of tuberculosis disease. Front Microbiol. 2019;10:395. [PubMed: 30886608]
16. Murase Y, Izumi K, Ohkado A, et al. Prediction of local transmission of *Mycobacterium tuberculosis* isolates of a predominantly Beijing Lineage by use of a variable-number tandem-repeat typing method incorporating a consensus set of hypervariable loci. J Clin Microbiol. 2018;56(1):e01016–e01017. [PubMed: 29046413]
17. Wang S. Development of a predictive model of tuberculosis transmission among household contacts. Can J Infect Dis Med Microbiol. 2019;2019:5214124–5214127. [PubMed: 31467622]
18. Asyary A, Prasetyo A, Eryando T, et al. Predicting transmission of pulmonary tuberculosis in Daerah Istimewa Yogyakarta Province, Indonesia. Geospat Health. 2019;14(673):171–177.
19. Mamiya H, Schwartzman K, Verma A, et al. Towards probabilistic decision support in public health practice: predicting recent transmission of tuberculosis from patient attributes. J Biomed Inform. 2015;53:237–242. [PubMed: 25460204]
20. Centers for Disease Control and Prevention. Reported tuberculosis in the United States, 2018. 2020. https://www.cdc.gov/tb/statistics/reports/2018/default.htm. Accessed March 3, 2020.
21. Ghosh S, Moonan PK, Cowan L, et al. Tuberculosis genotyping information management system: enhancing tuberculosis surveillance in the United States. Infect Genet Evol Jun. 2012;12(4):782–788.
22. Kammerer JS, Shang N, Althomsons SP, et al. Using statistical methods and genotyping to detect tuberculosis outbreaks. Int J Health Geogr. 2013;12:15. [PubMed: 23497235]
23. *SAS Statistical Software*, version 9.4. Cary, North Carolina: SAS Institute, Inc; 2020.
24. R Core Team. *R: A Language and Environment for Statistical Computing*, version 4.0.2. Vienna, Austria: R Foundation for Statistical Computing; 2019.
25. France AM, Grant J, Kammerer JS, et al. A field-validated approach using surveillance and genotyping data to estimate tuberculosis attributable to recent transmission in the United States. Am J Epidemiol Nov 1. 2015;182(9):799–807. [PubMed: 26464470]

26. Apley DWJingyu Z. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Series B Stat Methodology. 2020;82(4):1059–1086.

27. Wortham JM, Li R, Althomsons SP, et al. Tuberculosis genotype clusters and transmission in the U.S., 2009–2018. Am J Prev Med. 2021;61(2):201–208. [PubMed: 33992497]

28. Liaw A, Wiener M. Classification and regression by randomForest. R News. 2002;2(3):18–22.

29. Greenwell B, Boehmke B, Cunningham J, et al. gbm: Generalized boosted regression models. R package version 2.1.5. 2019. https://CRAN.R-project.org/package=gbm. Accessed August 14, 2020.

30. Meyer D, Dimitriadou E, Hornik K, et al. e1071: Misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7–2. 2019. https://CRAN.R-project.org/package=e1071. Accessed August 14, 2020.

31. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22. [PubMed: 20808728]

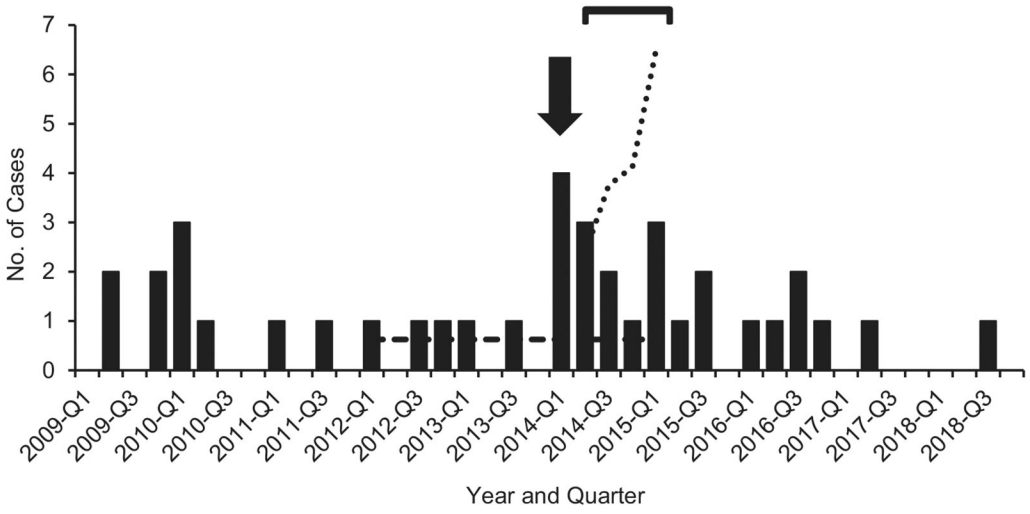

Figure 1.

Hypothetical epidemiologic curve of a cluster. Bars indicate the number of cases counted during the indicated quarter. Also indicated are an unexpected growth flag as an arrow, the corresponding baseline as a dashed line, follow-up period within a bracket, and accumulation of excess cases as a dotted line.

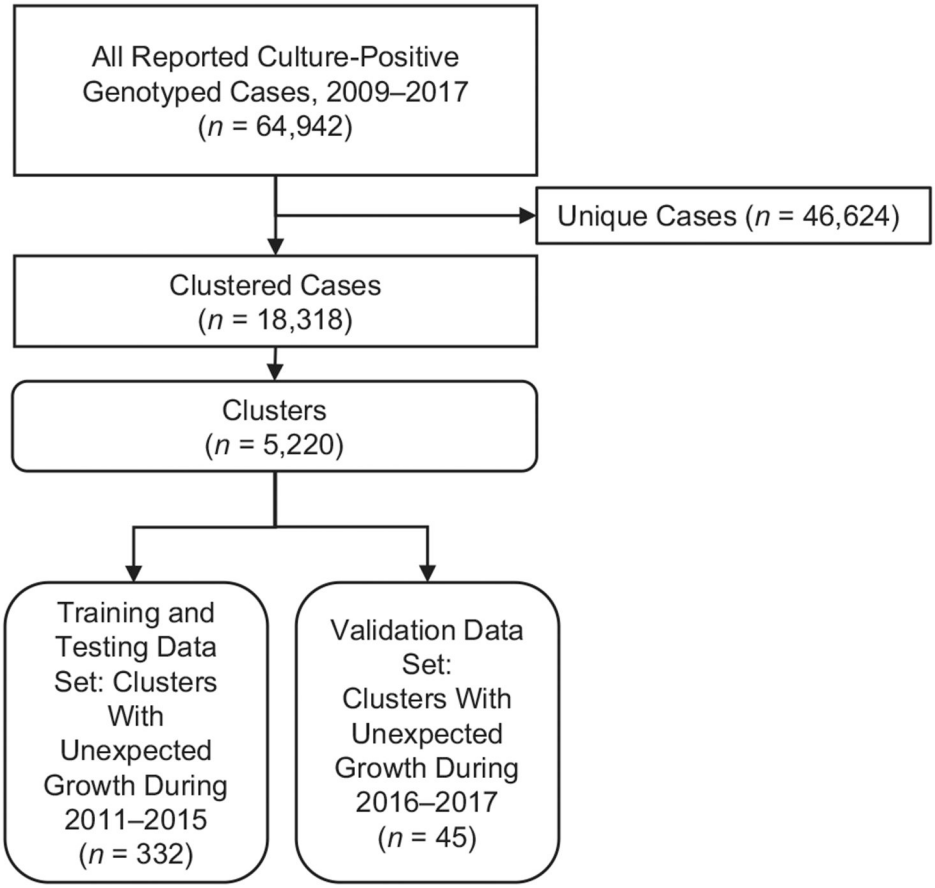

Figure 2.
Data set cohorts and tuberculosis surveillance data used to predict unexpected growth in cases, United States, 2011–2017.

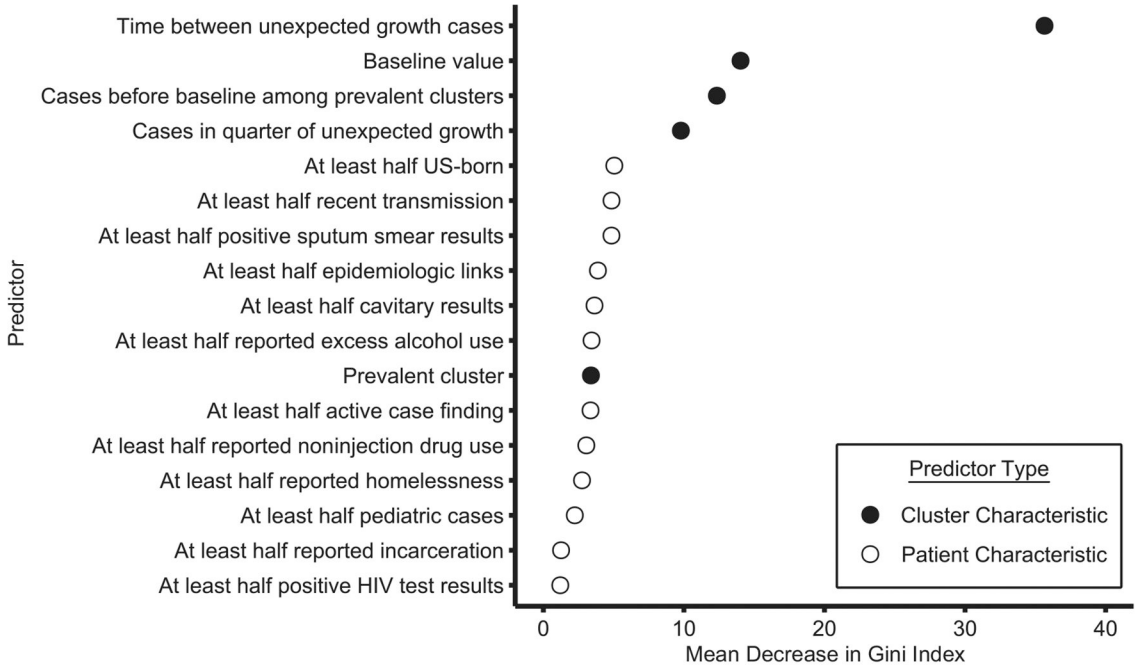

Figure 3.

Importance plot for all predictors in final model for predicting unexpected growth in tuberculosis cases, United States, 2011–2016. A final model was built using random forest on the 2-quarters time frame (the flagged unexpected growth quarter and 1 preceding quarter) with the “half” quantification (at least half the cases in the cluster were positive for the characteristic). We then calculated the variable importance for each of the predictors in this model. More details can be found in the Web Appendixes 3 and 4. Predictors are shown from high to low importance, where higher importance means that the predictor affects the model more. Filled points indicate predictors based on cluster characteristics, while outlined points indicate predictors based on patient characteristics. A description of each predictor can be found in Web Table 2. HIV, human immunodeficiency virus.

Table 1.

List of Machine Learning Methods and Variations Tested for Prediction Algorithm

Method	Variations Tested	R[a] Package	R Function	First Author, Year (Reference No.)
Tree-based ensembles	Random forest Gradient-boosting machine	randomForest gbm	randomForest gbm	Liaw, 2002 (28) Greenwell, 2019 (29)
Support vector machine	Radial kernel Linear kernel Polynomial kernel Sigmoid kernel	e1071	svm	Meyer, 2019 (30)
Regularized regression	Lasso regression Ridge regression Elastic net regression	glmnet	glmnet	Friedman, 2010 (31)

Abbreviation: Lasso, least absolute shrinkage and selection operator.

[a]R (R Foundation for Statistical Computing, Vienna, Austria).

Table 2.

Distribution of Outcomes in the Training and Testing Data Set (2011–2016) and Validation Data Set (2017–2018) for an Investigation of Predictors of Unexpected Growth in Tuberculosis Cases, United States

Data Set	No. of Clusters	Clusters With Excess Growth[a]		Median (Range) of Excess Growth Cases in Clusters With Excess Growth[a]	Clusters With Expected Growth		Median (Range) of Expected Growth Cases in Clusters With Expected Growth
		No.	%		No.	%	
Training and testing	332	140	42%	1.0 (0.5, 19.5)	192	58%	−0.5 (−3, 0)
Validation	45	20	44%	1.25 (0.5, 4)	25	56%	−0.5 (−2, 0)

[a]Excess growth is defined as observations with more than zero excess cases in the follow-up period.

Table 3.

Means and Standard Deviations for the Performance Metrics of the 5 Models With the Highest Youden Index From the Training and Testing Data Set (2011–2016) and the Final Model on the Validation Data Set (2017–2018) for an Investigation of Predictors of Unexpected Growth in Tuberculosis Cases, United States

Model[a]	Time Frame[b]	Quantification[c]	Method	Youden Index	Accuracy	Sensitivity	PPV	Specificity	NPV
1	2 quarters	Half	Random forest	0.165 (0.047)	0.608 (0.033)	0.398 (0.106)	0.563 (0.061)	0.768 (0.085)	0.637 (0.056)
2	9 quarters	Percent	SVM with linear kernel	0.149 (0.072)	0.593 (0.037)	0.386 (0.126)	0.555 (0.120)	0.763 (0.110)	0.630 (0.089)
3	9 quarters	Percent	Elastic net	0.137 (0.136)	0.594 (0.068)	0.331 (0.156)	0.567 (0.178)	0.806 (0.115)	0.624 (0.099)
4	9 quarters	Percent	SVM with sigmoid kernel	0.056 (0.092)	0.560 (0.044)	0.289 (0.050)	0.489 (0.144)	0.767 (0.089)	0.596 (0.066)
5	4 quarters	Percent	GBM	0.025 (0.069)	0.479 (0.024)	0.765 (0.118)	0.429 (0.056)	0.260 (0.087)	0.609 (0.096)
Final model	2 quarters	Half	Random forest	0.100	0.556	0.500	0.500	0.600	0.600

Abbreviations: CV, cross-validation; GBM, gradient-boosting machine; PPV, positive predictive value; NPV, negative predictive value; SVM, support vector machine.

[a]For the training and testing data set (first 5 models), the metrics are calculated from a 5-fold CV of 332 training/testing observations, and averaged across the 5 CV runs, with the mean (standard deviation) indicated here. Results were subset to models with a mean sensitivity >0.25, and then sorted from high to low Youden index. The top 5 models are shown here, while all models meeting these criteria are in Web Table 1. For the validation data set (final model), predictions were generated for all 45 clusters in the validation set (not used in the CV testing and training) using the final model (random forest built on the 2-quarters time frame with the half quantification). Excess growth was designated as the positive class while expected growth was designated as the negative class.

[b]Time frame indicates which cases were included in the predictors: 2 quarters included the flagged unexpected growth quarter and one preceding quarter; 4 quarters included the flagged unexpected growth quarter and 3 preceding quarters; and 9 quarters included the flagged unexpected growth quarter and 8 preceding quarters.

[c]Quantification indicates how the predictors were classified: Half means at least half the cases in the cluster were positive for the characteristic, while percent indicates the percentage of cases in the cluster positive for the characteristic.