



Article

# TB Hackathon: Development and Comparison of Five Models to Predict Subnational Tuberculosis Prevalence in Pakistan

Sandra Alba <sup>1,\*</sup>, Ente Rood <sup>1</sup>, Fulvia Mecatti <sup>2</sup>, Jennifer M. Ross <sup>3</sup>, Peter J. Dodd <sup>4</sup>, Stewart Chang <sup>5</sup>,  
Matthys Potgieter <sup>6</sup>, Gaia Bertarelli <sup>7</sup>, Nathaniel J. Henry <sup>8</sup>, Kate E. LeGrand <sup>9</sup>, William Trouleau <sup>10</sup>,  
Debebe Shaweno <sup>4</sup>, Peter MacPherson <sup>11,12,13</sup>, Zhi Zhen Qin <sup>14</sup>, Christina Mergenthaler <sup>1</sup>,  
Federica Giardina <sup>15</sup>, Ellen-Wien Augustijn <sup>16</sup>, Aurangzaib Quadir Baloch <sup>17</sup> and Abdullah Latif <sup>17</sup>

- <sup>1</sup> KIT Royal Tropical Institute, 1092 AD Amsterdam, The Netherlands; e.rood@kit.nl (E.R.); c.mergenthaler@kit.nl (C.M.)
  - <sup>2</sup> Department of Sociology and Social Sciences, University of Milano Bicocca, 20126 Milan, Italy; fulvia.mecatti@unimib.it
  - <sup>3</sup> Departments of Global Health and Medicine, University of Washington, Seattle, WA 98195, USA; jross3@uw.edu
  - <sup>4</sup> School of Health and Related Research, University of Sheffield, Sheffield S1 4DA, UK; p.j.dodd@sheffield.ac.uk (P.J.D.); d.shaweno@sheffield.ac.uk (D.S.)
  - <sup>5</sup> Institute for Disease Modeling, Seattle, WA 98109, USA; stchang@idmod.org
  - <sup>6</sup> Epcon, 2000 Antwerp, Belgium; thys@epcon.ai
  - <sup>7</sup> Sant'Anna School of Advanced Studies, 56127 Pisa, Italy; gaia.bertarelli@santannapisa.it
  - <sup>8</sup> Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK; nathaniel.henry@ndm.ox.ac.uk
  - <sup>9</sup> Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA 98109, USA; kel15@uw.edu
  - <sup>10</sup> École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; William.trouleau@epfl.ch
  - <sup>11</sup> Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Queen Elizabeth Central Hospital, Blantyre 312225, Malawi; Peter.MacPherson@lstmed.ac.uk
  - <sup>12</sup> Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK
  - <sup>13</sup> Clinical Research Department, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
  - <sup>14</sup> Stop TB Partnership, 1218 Geneva, Switzerland; zhizhenq@stoptb.org
  - <sup>15</sup> Department of Biostatistics, Radboud University Medical Centre, 6525 GA Nijmegen, The Netherlands; federica.giardina@radboudumc.nl
  - <sup>16</sup> Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7514 AE Enschede, The Netherlands; p.w.m.augustijn@utwente.nl
  - <sup>17</sup> Pakistan National Tuberculosis Control Programme, Islamabad 44000, Pakistan; aurangzaib.quadir@gmail.com (A.Q.B.); abdullah.latif@gmail.com (A.L.)
- \* Correspondence: s.alba@kit.nl



**Citation:** Alba, S.; Rood, E.; Mecatti, F.; Ross, J.M.; Dodd, P.J.; Chang, S.; Potgieter, M.; Bertarelli, G.; Henry, N.J.; LeGrand, K.E.; et al. TB Hackathon: Development and Comparison of Five Models to Predict Subnational Tuberculosis Prevalence in Pakistan. *Trop. Med. Infect. Dis.* **2022**, *7*, 13. <https://doi.org/10.3390/tropicalmed7010013>

Academic Editors: Jacob Creswell, Sode Matiku, Masja Straetemans and Luan Vo

Received: 27 September 2021

Accepted: 11 January 2022

Published: 17 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Pakistan's national tuberculosis control programme (NTP) is among the many programmes worldwide that value the importance of subnational tuberculosis (TB) burden estimates to support disease control efforts, but do not have reliable estimates. A hackathon was thus organised to solicit the development and comparison of several models for small area estimation of TB. The TB hackathon was launched in April 2019. Participating teams were requested to produce district-level estimates of bacteriologically positive TB prevalence among adults (over 15 years of age) for 2018. The NTP provided case-based data from their 2010–2011 TB prevalence survey, along with data relating to TB screening, testing and treatment for the period between 2010–2011 and 2018. Five teams submitted district-level TB prevalence estimates, methodological details and programming code. Although the geographical distribution of TB prevalence varied considerably across models, we identified several districts with consistently low notification-to-prevalence ratios. The hackathon highlighted the challenges of generating granular spatiotemporal TB prevalence forecasts based on a cross-sectional prevalence survey data and other data sources. Nevertheless, it provided a range of approaches to subnational disease modelling. The NTP's use and plans for these outputs shows that, limitations notwithstanding, they can be valuable for programme planning.

**Keywords:** small area estimation; tuberculosis burden; predictive modelling; subnational prevalence; spatial epidemiology; forecasting

---

## 1. Introduction

There is increasing demand for tuberculosis (TB) estimates at subnational level to inform TB programme planning in low and middle-income countries [1]. Indeed, there is substantial geographical heterogeneity in TB prevalence in high TB burden countries. Subnational estimates are therefore considered valuable by national TB control programmes (NTP) to optimise resource allocation.

Case notifications are the main source of subnational TB data, but even in countries with good health coverage, these data may not reflect patterns in disease burden [1]. Reasons include: Ref. [1] cases from one reporting area may be diagnosed in neighbouring areas due to better access and quality of care (or people's perception thereof); Ref. [2] even in countries where TB is a notifiable disease some service providers may not report through the national reporting system (e.g., private sector providers) [1]. Issues with subnational TB burden based on notifications are further exacerbated in contexts where access to care and reporting vary geographically.

National population-based prevalence surveys provide a direct measurement of the burden of disease. They are considered the gold standard in estimating the prevalence of TB, but they are typically not powered to provide subnational estimates of TB. Most TB prevalence surveys only allow the generation of reasonably precise estimates of TB prevalence at the national level in a small number of strata (e.g., two to three geographical regions) [1].

A number of approaches to produce sub-national estimates of TB burden have been proposed, but validating the accuracy of model predictions remains challenging. Indeed, these estimation methods are typically implemented in settings where the information need is greatest, which is almost always where there is no accurate empirical data against which to evaluate the validity of model predictions [2,3]. Notable examples of subnational TB burden estimation include the SUBsET model which aims to estimate subnational incidence of TB in Indonesia [4], Bayesian models to estimate subnational TB mortality [5] and incidence [6] in Brazil or TB prevalence in Cambodia [7]. Alternative approaches to subnational TB estimation include indirect methods to estimate sub-national notification gaps in Bangladesh [8], Nepal and Pakistan [9]. While the cited approaches implemented various statistical methods of validation, the extent to which predictions accurately reflect real burden often remains uncertain. To the best of our knowledge, there have been no attempts to apply more than one modelling approach in the same setting, meaning that cross-validation of approaches has not been attempted.

Pakistan's NTP is among the many programmes worldwide who value the importance of subnational estimates of TB burden but do not have reliable estimates to support the TB response. Pakistan is a very diverse country with a population of 217 million in four provinces and three regions. TB is a major public health concern, with an estimated 570,000 new cases in 2019 and 43,900 deaths attributable to the disease [10]. Pakistan was among the eight countries that accounted for two thirds of the total global number of incident cases worldwide in 2019 [10]. Moreover, under-diagnosis and under-reporting of cases are considered key barriers to ending TB. Pakistan is among the five countries which accounted for more than half of the people with TB who are missed globally, as measured by the gap between the number of incident cases and the number of people notified on TB treatment in 2019 [10]. This is partly attributed to under-reporting by public and private providers who operate outside the NTP [11].

Against this backdrop, the Pakistan NTP partnered with epidemiologists at KIT Royal Tropical Institute to launch a virtual hackathon for the estimation of the subnational TB burden in Pakistan. Hackathons are problem solving events where participants collaborate

intensively to develop a proposed solution for a specific issue within a short period of time. Hackathons originated in the field of software development and are increasingly proposed as a problem solving model in health. To date, most hackathons in global health have been conducted to develop new medical technologies [12,13] and especially solutions relying on mobile and wireless devices [14,15]. Examples also include the development of new analytical [16,17] methods and fostering multi-cultural dialogue [18]. One of hackathons' most appealing features is the potential for greater innovation due to the heterogeneity of participants' backgrounds and the encouragement of collaborations across institutions [16,19].

The Pakistan TB hackathon aimed to bring together various research groups interested in TB modelling to collaborate on a joint modelling exercise of the subnational TB burden in Pakistan in 2018. The hackathon's objectives were two-fold: (1) to compare and cross-validate different models for small area estimation of TB in Pakistan; (2) to provide Pakistan NTP with data to tailor their efforts to ending TB to different sub-national contexts.

## 2. Methods

The TB hackathon was a virtual event which did not require people to be physically present in the same space at the same time. It was launched in April 2019 by means of announcements on the KIT website, on social media and in mailings lists. Participants were offered three months and the same set of data (Table 1) to develop their own models to estimate sub-national TB burden.

**Table 1.** Datasets made available to TB hackathon modelers by Pakistan NTP.

Dataset	Disaggregation	Time Period
1. Prevalence survey data <sup>1</sup>	Individual	2010–2011
2. TB notifications	District	quarterly 2009–2018
3. Laboratory External Quality Assessment data	District	quarterly 2013–2017
4. Drug-sensitive TB treatment outcomes data	District	quarterly 2009–2015
5. Drug-Resistant TB notifications	District	quarterly 2009–2018
6. Master list of TB facilities	Health facility	2019
7. Sputum smear testing data	District	quarterly 2009–2017
8. Private sector notifications	District	Yearly 2017–2018
9. HIV registrations	Province	2001–2018
10. HIV testing rates among TB cases	District	quarterly 2009–2018
11. Census Population estimates	District	2017
12. Shape files	District	2019

<sup>1</sup> Including village names corresponding to the survey-clusters

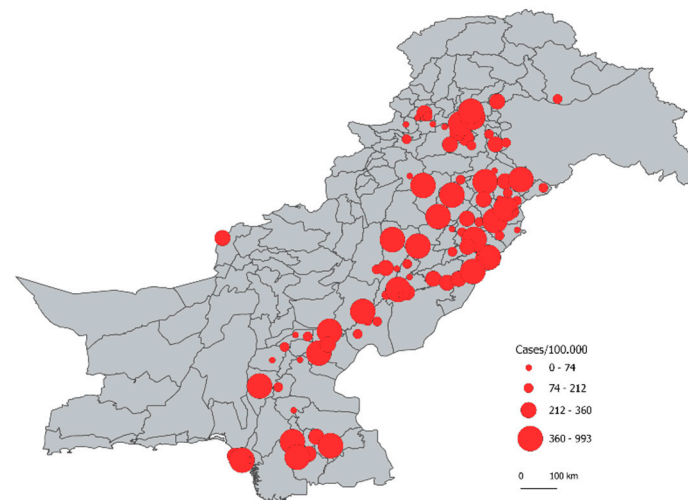
Participating teams were requested to produce district-level estimates of bacteriologically positive TB prevalence among adults (over 15 years of age) for 2018. The choice of 2018 as prediction year was the result of a compromise between the information needs of the NTP (who needed the most recent possible estimate for programme planning) and what was considered reasonable within the scope of available data for modelling (with auxiliary data for modelling expected to be available up until 2018 at the latest).

Nine teams applied for and signed data sharing agreements with the Pakistan NTP. Participating teams developed their models between mid-May and mid-September 2019. Five teams submitted district-level estimates for the whole country by the September 2019 deadline. Submitted models were appraised and compared by an evaluation panel comprising global TB and statistical experts including representatives from Pakistan's NTP and the World Health Organisation. Feedback from the evaluation panel was provided to the modellers.

### 2.1. Data Sources

Table 1 provides a complete overview of the data provided to hackathon modellers. Participants were invited to use the data provided, and any other publicly available data, for their model.

The Pakistan NTP provided case-based data for their 2010–2011 TB prevalence survey data. The national TB prevalence survey was a nationwide cross-sectional survey with multistage cluster sampling conducted in 95 clusters from 68 districts [20]. The clusters corresponded to *tehsils* (sub-districts) and were selected using sampling proportional to the estimated *tehsil* population size in 2010 projected from 1998 census data [20,21]. The Federally Administered Tribal Areas, district Dera Bugti in Balochistan, and 17 *tehsils*, of Khyber Pakhtunkhwa, were excluded from the survey due to serious security threats. Combined, these excluded areas account for 6.4% of Pakistan's population. A total of 105,913 adults ( $\geq 15$  years of age) participated in the survey, of whom 10,471 (9.9%) were eligible for sputum examination [20]. Of these, 8521 (81.4%) submitted at least one specimen for sputum examination. The proportion of TB bacteriologically positive people (out of all tested) by cluster according to the 2010–2011 prevalence survey is shown in Figure 1 [20]. (These do not take into account the Pakistan TB prevalence survey report adjustment for missing TB results among participants eligible for smear examination or non-participation.).



**Figure 1.** Proportion of TB bacteriologically positive people (out of all tested) by cluster in Pakistan in 2010–2011 prevalence survey [21].

The Pakistan NTP also shared data relating to TB testing and treatment, in both the private and public sectors, for the period between 2010 and 2018 (Table 1) to fill the gap between the year of the TB prevalence survey (2010–2011) and the year for which predictions were requested (2018). This was complemented by further data on HIV registrations as well as screening and testing for HIV in TB patients (since people living with HIV are more likely than others to develop TB disease). Participants were also directed to the 2017 census provisional province-wise population counts by sex and rural or urban location [22] and overall population counts by *tehsil* [23] (which were openly available online at the time). To ensure geographical consistency across all model predictions, participants were provided

with the latest district shapefiles obtained from GADM ([gadm.org](http://gadm.org)). In April 2019, the GADM versioning included 143 out of the 146 districts which existed up to the 2017 census (in addition to the frequent redrawing of district boundaries in Pakistan, there was a major overhaul as part of the 2017 census, leading to the current 156 districts).

## 2.2. Comparison of Models and Predictions

In the absence of an empirical ground truth to evaluate the predictions (e.g., a 2018 subnational prevalence survey), we first compared the models and then appraised the quality of predictions using maps and a series of bespoke data quality indicators.

To compare models, we described the modelling building approaches (modelling and inferential frameworks, covariance structures, selection of final model, post-modelling processing of predictions) and strategies for variable selection (choice of outcome and predictor variables, variable processing, lowest level of spatial aggregation).

To compare predictions, we first mapped each model's 2018 district estimate on common scales. We then calculated summary statistics and data quality indicators of completeness, pseudo-accuracy, precision, cross-validity and credibility. We assessed completeness as the proportion of Pakistan's districts for which predictions were available (out of 143). For pseudo-accuracy, hackathon modellers were requested to perform leave-one-out-cross-validation and provide the  $R^2$  comparing actual and predicted 2010–2011 cluster-level prevalence point estimates (since there was no prevalence survey or other empirical data in 2018 this calculation could not be made for the 2018 predictions). For precision, we calculated an approximated coefficient of variation as the difference between the upper limit and lower limit of 95% CI divided by the point estimate. Cross-validity was assessed by comparing model predictions using scatter plots and Pearson's correlation coefficient. We also produced a number prediction plots (including histograms, precision plots and pairwise correlations) to support this comparison of predictions.

The indicator of credibility differs from all others as it is based on expert opinion. We presented anonymised maps of each model's 2018 district estimates to four Pakistan TB experts (from the Pakistan NTP) and asked them to grade the estimates on a scale from 1 to 10 based on how credible they deemed the estimates, based on their knowledge of the TB epidemic in their country. We presented individual grades and summary statistics (mean) and calculated the average intra-class correlation coefficient to measure agreement between experts.

## 2.3. Identification of Districts with Most Under-Reporting

To provide data to support the NTP's programme planning, we identified the areas with most likely under-reporting according to the hackathon models. For this purpose, for each model we created maps displaying the ratio between new and relapse bacteriologically positive 2018 TB notification rate from NTP sources (numerator) and model predictions (denominator), by district. Low values of this 'notifications to prevalence ratio' for a particular district are assumed to correspond to the under-reporting of TB cases to and by the NTP. For each model, districts were allocated to a quantile based on this ratio and we identified those districts which consistently scored in the lowest quantile across all models.

# 3. Results

## 3.1. Comparison of Models

Details of the model building approaches can be found in Table 2 and Supplementary File S1. The hackathon models included a Bayesian binomial logistic regression with Markov Chain Monte Carlo (MCMC) inference (Model 1), an approximate Bayesian binomial logistic regression model with integrated nested Laplace approximations (INLA) inference (Model 2), an approximate Bayesian binomial-logistic model fit using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Model 3), a Small Area Estimation and Latent Markov model with MCMC inference (Model 4) and artificial neural network followed by a Bayesian network (Model 5). In other words, all models were fitted within Bayesian inference frameworks;

four were statistical models (Models 1–4) and one was a machine learning model (Model 5). Approaches to model selection included use of the log-scoring rule (Model 1), leave-one-out-cross-validation based on mean squared error and/or  $R^2$  (Models 2 and 3) and Chib's estimator (Model 4) (see Supplementary File S1 for more details).

**Table 2.** Model specifications.

	Model 1	Model 2	Model 3	Model 4	Model 5
<b>Modelling framework</b>	Binomial-logistic regression	Binomial-logistic regression	Binomial-logistic regression	Small Area Estimation (SAE) and Latent Markov (LM) modelling as linking model for SAE	Self-Organising Maps (SOM) on binomial
<b>Inference</b>	Bayesian inference with Markov Chain Monte Carlo with No-U-Turn-Sampler (NUTS)	Approximate Bayesian inference with integrated nested Laplace Approximations (INLA)	Approximate Bayesian inference with Broyden–Fletcher–Goldfarb–Shanno algorithm	Bayesian inference with Data Augmentation Markov Chain Monte Carlo and Gibbs sampler	Bayesian Artificial Neural Network
<b>Covariance structure</b>	Spatially explicit hierarchical model with fixed and random effects.	Spatially explicit hierarchical model with fixed and random effects	Spatially explicit hierarchical model with fixed and random effects	Hierarchical Discrete latent state model depending on a Gaussian linking model	N/A
<b>Outcome variable</b>	Bacteriologically-confirmed TB cases from TB prevalence survey at cluster-level by age and sex	Bacteriologically-confirmed TB cases from TB prevalence survey at cluster-level	Bacteriologically-confirmed TB cases from TB prevalence survey at cluster-level	Bacteriologically-confirmed TB cases from TB prevalence survey at district level	Bacteriologically-confirmed TB cases from TB prevalence survey at district level
<b>Final set of predictors<sup>1</sup></b>	SES, HH size, Indoor smoke, BMI, WAZ, Vaccination coverage, Prevalence of cough, Distance to health facility	Age 15–24 Female Age 15–24 * female Ag 65+ Age_65+ * Sindh Underweight Underweight * KPH	Population density Access to cities [10] Density of TB facilities Poverty Urban extents Locations of protests Locations of violent acts Aridity	Urban households Rural households Urban male pop Rural male pop Urban female pop Rural female pop Pop growth overall Bac+ notifications Bac-notifications EP notifications	All-forms TB notifications Bac+ TB notifications SS+ rate among tested Population density Average household size Percentage rural population Growth rate (urban, rural) Sex ratio (urban, rural) Log gross national income Life expectancy Expected years of schooling Mean years of schooling Human development index

<sup>1</sup> x1: x2 represents factor multiplication, while x1 \* x2 represents factor crossing and is equivalent to x1 + x2 + x1: x2.

Details of variable selection strategies can be found in Table 2 and Supplementary File S1. By and large, the modelling teams made similar choices in terms of the outcome variable and candidate predictors, with varying choices in terms of data sources, geographical linkage and processing. All modelling teams used the raw TB prevalence provided by the NTP without performing multiple imputation of missing data or adjustment for non-responses. Model 1 stood out as the only model fitting cluster-level data disaggregated by age and sex. Model 2 and Model 3 both fitted TB cases at cluster-level whereas Model 4 and Model 5 fitted models at district-level. Four out of five models used TB notification data as candidate predictors, Model 3 being the only exception. Most models used non-TB health data as well as socio-demographic data. In addition, Model 3 used climate (precipitation) and civil unrest (protests and violent acts) data. Model 5 was the only model to use macroeconomic and development data (such as gross national income and human development index). Model 2 and Model 3 were the most granular models, including fine-scale predictor data (1 km or 5 km) for several health and socio-demographic indicators. Model 2 performed spatial kriging of predictors to obtain granular cluster-level predictors, with inter-survey estimates derived by linear interpolation between survey years; whereas Model 3 extracted point estimates of predictors from gridded data sources, when available, using the georeferenced survey cluster locations.

Final prediction models included between 7 and 10 variables (Table 2). Three out of five models (Models 1–3) did not include routine TB data in their final selected predictive models. While Model 3 did not include these in the list of potential covariates to start with, both Model 1 and Model 2 did not observe a significantly strong enough correlation between programmatic data (such as TB testing and notifications) and cluster-level TB prevalence rate and thus did not carry them forward into the model selection step (as has been documented elsewhere [2]). Overall, the models were fairly consistent in which local socio-demographic risk factors were associated with TB prevalence rates: poverty, underweight, urban extents, aridity and gender. Model 2 stands out as the only model with interactions.

### 3.2. Comparison of Predictions

All model predictions and accompanying credible intervals are provided in Supplementary File S2. Data quality appraisal statistics for all models are presented in Table 3, while prediction plots (histograms, precision plots and pairwise correlations) are presented in Supplementary Figures S1–S3.

Model 3 stands out as the model with the lowest district-wise mean and median estimates due to the calibration with Global Burden of Disease study 2017 national estimates [24] (Model 3: mean = 192/100,000, median = 162/100,000) while Model 1 stands out as the model with the highest mean and median estimates (Model 1: mean = 754/100,000, median = 378/100,000) (Supplementary Figure S1). Model 4 was the only model which predicted zero prevalence for a number of districts, with the 10th percentile equal to zero. The most complete set of predictions were provided by Model 2 (143 districts) and Model 3 (142 districts) whereas Model 1 provided estimates for 131 districts (due to linkage issues with certain covariate data sources) and Model 4 for 94 districts (those that had complete predictors for 2010–2011). According to the LOOCV  $R^2$  statistics, Model 3 scored the best in terms of pseudo-accuracy, defined as a model's ability to predict cluster level TB in 2010–2011, when comparing with the actual measured values ( $R^2 = 0.733$ ), followed by Model 1 ( $R^2 = 0.404$ ), Model 2 ( $R^2 = 0.320$ ) and Model 5 ( $R^2 = 0.115$ ). Overall Model 2 and Model 4 provided the narrowest confidence intervals as can be seen by the lower precision ratios, whereas Model 3 has the widest confidence intervals. For Model 1 and Model 2, higher prevalence estimates tended to be less precise, while for Model 4 and Model 5, lower prevalence estimates tended to be less precise (Supplementary Figure S2). Within Model 3, no linear relationship was observed between the mean prevalence estimate for each district and the width of the uncertainty intervals surrounding that estimate.

**Table 3.** Predictions: data quality appraisal.

	Model 1	Model 2	Model 3	Model 4	Model 5
<b>Summary statistics</b> <sup>1</sup>	Min = 104 Max = 7425 Mean = 754 Median = 378	Min = 276 Max = 2050 Mean = 508 Median = 430	Min = 51 Max = 456 Mean = 192 Median = 162	Min = 0 Max = 1000 Mean = 362 Median = 382	Min = 44 Max = 906 Mean = 366 Median = 289
<b>Completeness</b> <sup>2</sup>	131	143	142	94	139
<b>Pseudo-accuracy by LOOCV for 2010</b> <sup>3</sup>	R <sup>2</sup> = 0.404	R <sup>2</sup> = 0.320	R <sup>2</sup> = 0.733 <sup>4</sup>		R <sup>2</sup> = 0.115
<b>Cross-validation</b> <sup>5</sup>	Model 2: r = -0.0882 Model 3: r = 0.2305 Model 4: r = -0.0041 Model 5: r = 0.0001	Model 1: r = -0.0882 Model 3: r = 0.4029 Model 4: r = 0.2492 Model 5: r = 0.1495	Model 1: r = 0.2305 Model 2: r = 0.4029 Model 4: r = 0.2402 Model 5: r = 0.1583	Model 1: r = -0.0041 Model 2: r = 0.2492 Model 3: r = 0.2402 Model 5: r = 0.0778	Model 1: r = 0.0001 Model 2: r = 0.1495 Model 3: r = 0.1583 Model 4: r = 0.0778
<b>Precision</b> <sup>6</sup>	Ratio = 2.69	Ratio = 0.78	Ratio = 5.30	Ratio = 0.63	Ratio = 2.06
<b>Credibility score</b> <sup>7</sup>	Rater 1: 3 Rater 2: 4 Rater 3: 3 Rater 4: 5 Mean score = 3.75	Rater 1: 5 Rater 2: 3 Rater 3: 3 Rater 4: 5 Mean score = 4	Rater 1: 7 Rater 2: 7 Rater 3: 7 Rater 4: 6 Mean score = 6.75	Rater 1: 4 Rater 2: 4 Rater 3: 3 Rater 4: 5 Mean score = 4	Rater 1: 7 Rater 2: 8 Rater 3: 6 Rater 4: 6 Mean score = 6.75

<sup>1</sup> Prevalence per 100,000 inhabitants. <sup>2</sup> Out of 143 districts. The difference between 136 and 143 is accounted for by districts in contested areas of Pakistan: 1 district in India-administered Kashmir, 1 district in Pakistan-administered Kashmir, 3 districts in the Federally Administered Tribal Area (FATA) and 2 districts in Balochistan. <sup>3</sup> LOOCV comparing final model estimates for 2010–2011 with actual prevalence survey cluster-level estimates for 2010–2011. This could not be calculated for Model 4 as LOOCV metrics are not practical for SAE-LM models (computationally too intensive) and were not produced by Model 5. <sup>4</sup> When performing cross validation, Model 3 excluded each cluster from the original survey; this meant that for cluster observations that were originally geo-matched to admin3 units and then resampled to multiple admin4 centroids, all down-sampled points corresponding to a single survey observation were excluded from a single out-of-sample run. <sup>5</sup> Pairwise correlations of district level central estimates of TB prevalence predictions (Pearson's correlation coefficient). <sup>6</sup> Ratio = [(upper limit of 95% credible interval) – (lower limit of 95% credible interval)]/(prevalence estimate). <sup>7</sup> Four TB experts from the Pakistan National TB control Programme were asked to grade models from 1–10 based on how credible they deemed the model estimates.

Overall there was very high heterogeneity in model predictions as can be seen from the five maps (Figure 2). Model 2 and Model 3 both present a more smoothed surface, which most likely reflects the high granularity of their approach as well as the spatially auto-correlated error structure used in the models. They are also the two most highly correlated estimates, by Pearson's correlation coefficient, albeit weakly ( $r = 0.4029$ ). Overall, Model 3 appears to be the most 'average' model, as it shares most similarities with other models (average pairwise correlations in district level central estimates of TB prevalence predictions) as can be seen in Table 3 and Figure 2 and Supplementary Figure S3. Model 3 and Model 5 obtained the highest average score (6.75) in the TB expert grading (Table 3). The average intra-class correlation coefficient between models was 0.92, indicating high clustering of grades within models (92% of the total variation in grades is between models) and thus high agreement between raters.

### 3.3. Identification of Districts with Most Under-Reporting

The maps displaying the ratio of 2018 new and relapse TB Notification rate over the predicted prevalence (Figure 3) provide information on areas with most under-reporting according to the model predictions. Areas consistently rated in the lowest notification rate to prevalence ratio quantile across all models include districts in Gilgit Baltistan in the north of the country (Ghizer and Ghanche districts); Khyber-Pakhtunkhwa in the north-west (Mohmand, Northern Waziristan, Southern Waziristan and Kurram) and Balochistan in the south-west (Kachhi, Musa Khel, Chagai, Kalat, Dera Bugti, Gwadar, Haranti and Lehri) (Supplementary Table S3).



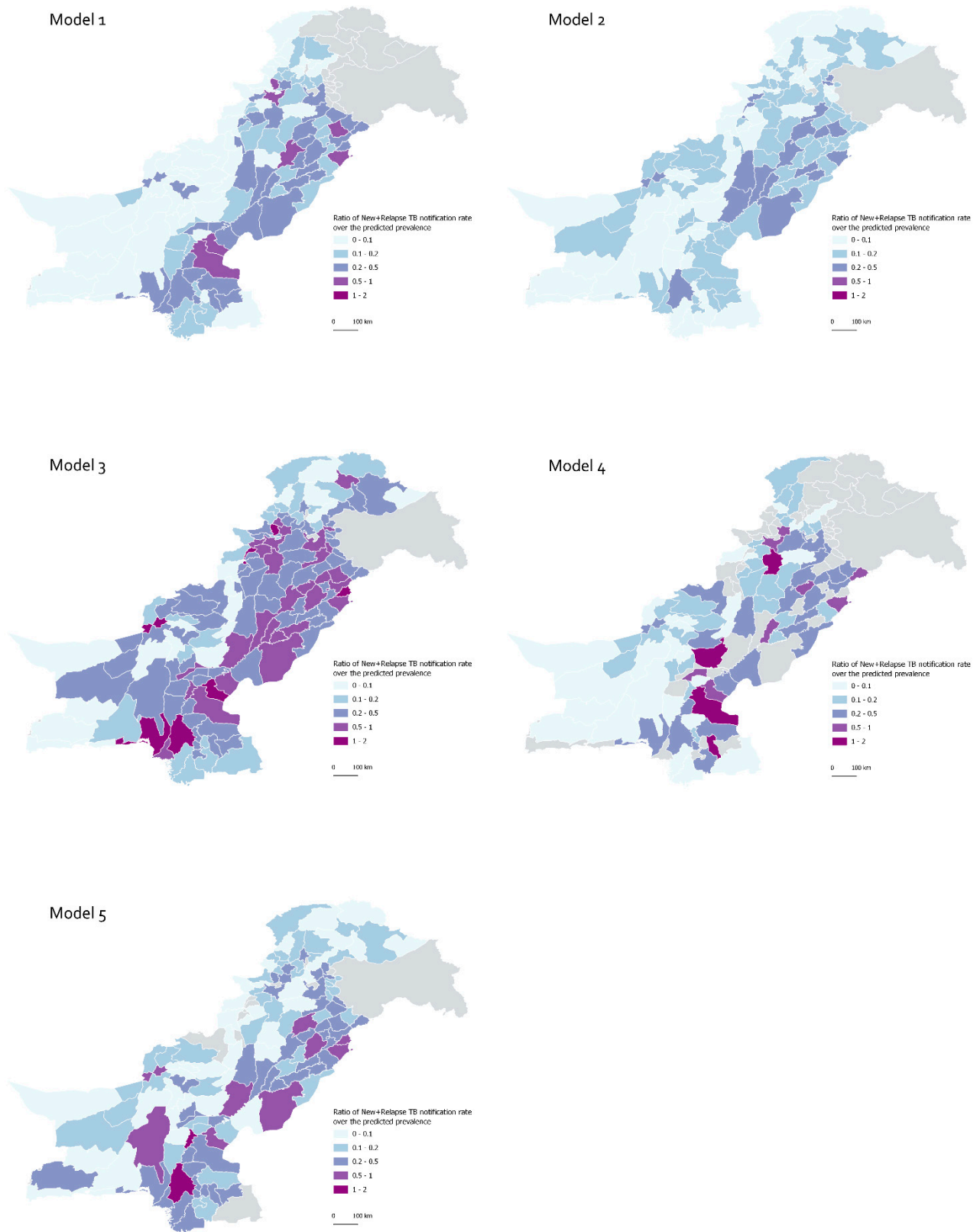
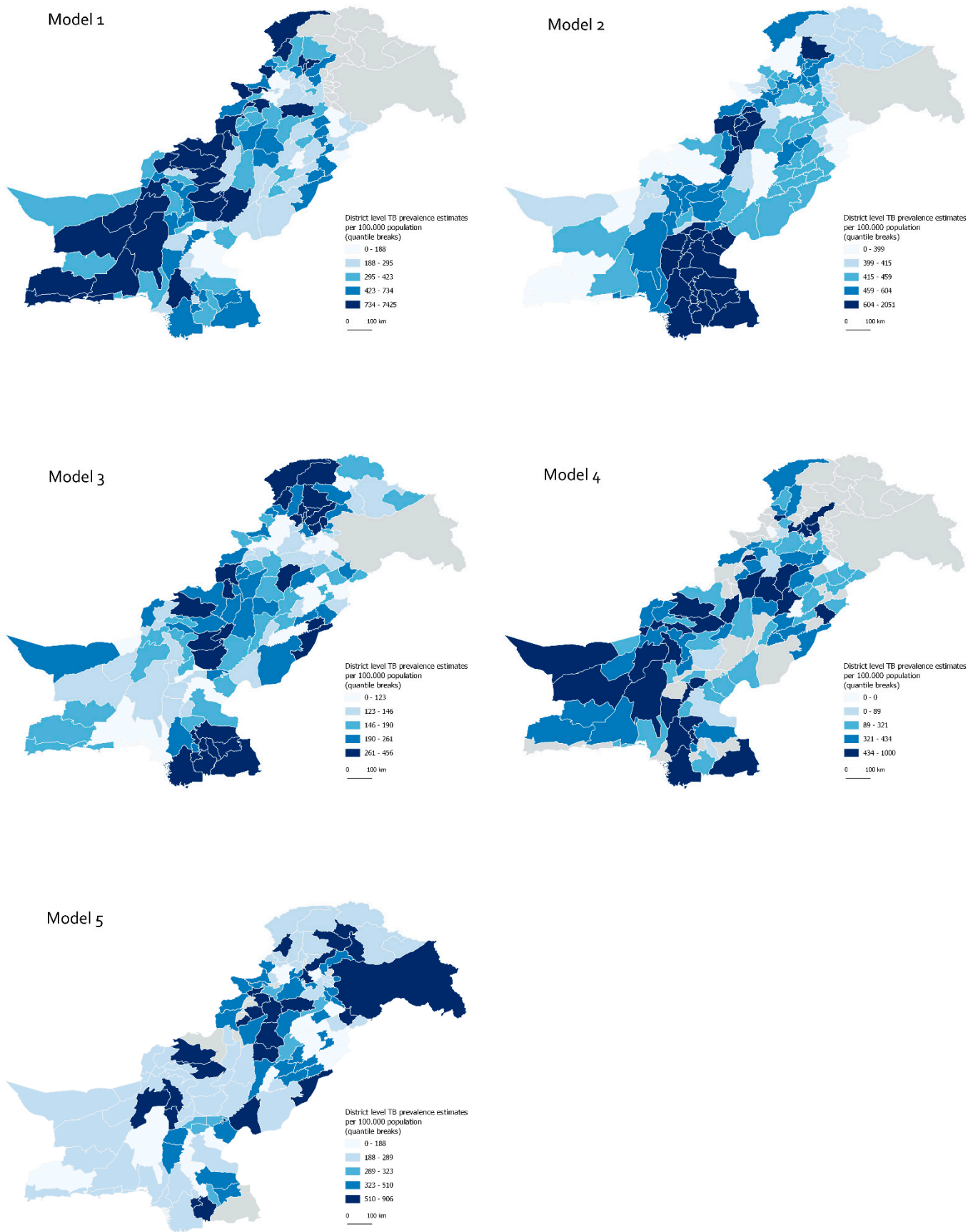


Figure 2. Predicted district-level TB prevalence in 2018, by model.



**Figure 3.** Ratio of 2018 new and relapse bacteriologically positive TB Notification rate over the predicted prevalence in each district, by model.

#### 4. Discussion

The TB hackathon provided five sub-national TB burden models that could be compared both in terms of their methodology and outputs. In doing so, the hackathon provided an opportunity to explore the utility of state-of-the-art modelling approaches to produce consistent TB prevalence predictions. It also proved useful to identify data sources which can be used to estimate TB prevalence at small spatial scales. The heterogeneity in model predictions shows that models based on a cross-sectional cluster-based prevalence surveys are limited in their ability to generate granular predictions into the future—even if they are complemented with other longitudinal and spatially disaggregated data sources. Indeed, model output comparisons highlighted the limited consensus between the different model outputs, and in the absence of an empirical ground truth against which to compare model predictions, it remains unclear which—and if any—predictive models produced reliable estimates. Nevertheless, we were able to identify a number of districts with consistently low notification to prevalence ratios across most models which could be prioritised for case finding activities.

Models 1–3 represent the more traditional statistical and epidemiological approach to TB modelling using a binomial logistic regression model. Although the predictions varied substantially across these three models, they shared a number of features. For instance, since these models were fitted using a logit link function, they could not generate estimates exactly equal to zero, and generally predicted values higher than the data in cases where prevalence estimates from the data are zero or relatively low. Furthermore, in Models 1 and 2, overall larger point estimates were associated with wider uncertainty. Models 4–5 changed the modelling perspective towards modern computational power and abundance of data. To the best of our knowledge, there are very few applications of Small Area Estimation and Latent Markov modelling (SAE-LM, Model 4) [25] and none on health data so far, and similarly there are limited examples of Bayesian Artificial Neural Network (ANN, Model 5) in health [26–28]. The main strength of these models lies in their flexibility, as there is no imposed function to link the outcome and predictors. This broadens options to the entire class of parametric probability distributions for the outcome. Both analytical and practical advantages follow. For instance, Model 4 was the only model able to estimate low district-level prevalence estimates and even zero. In contrast to Models 1 and 2, in Models 4 and 5 lower point estimates were associated with wider uncertainty.

Although we were not able to provide one set of validated estimates for sub-national TB planning, the NTP still found a number of practical applications for the hackathon models. The Pakistan NTP used Models 2 and 3 as a basis for sample size calculations for the upcoming TB prevalence survey (planned in 2022), given that they had both scored the highest as per the metrics presented in Table 3. Model 5, on the other hand, was used as a starting point to prioritise TB chest camps, a project the modellers became involved in shortly after participating in the hackathon. Chest camp data (including numbers of people screened, symptomatic, tested and positive) are now captured digitally in the field and are analysed in real-time. Incoming data are used to continuously update Model 5 predictions to provide ever more validated and accurate TB prevalence data at local level for the following decision making rounds on the location of chest-camps. Similarly, the NTP plans to use the hackathon models to operationalise other case-finding activities in their 2021–2023 National Strategic plan. In addition to chest camps, these include establishing sputum transportation mechanisms at primary health care levels, and engaging private providers in the diagnosis and treatment of TB. The hackathon data may be used, in combination with notification data, to understand the effectiveness of these interventions and further validate the accuracy of hackathon outputs.

One of the hackathon challenges was the combination of a spatial decomposition problem (from a national to district level estimates of TB) combined with a temporal forecast (projecting from 2010–2011 to 2018). While this was necessary to provide relevant information to NTP decision makers, it also added two layers of complexity for the modellers, each with their own theoretical and practical challenges. Spatial decomposition of

survey data has known challenges, further compounded by the fact that TB is a rare and unevenly scattered attribute [1]. As a result, all models' predictive power suffered from the following limitations: (1) data sparsity as a result of a limited number of clusters to base district-prevalence estimates (the 2010–2011 prevalence survey collected data from 95 [21] out of over 530 tehsils in the whole country [23]); (2) a small number of detected prevalent cases with over-dispersion in the distribution of clusters by case count (the mean number of bacteriologically positive cases was 3.3 per cluster with a standard deviation of 2.7 and 13/95 clusters with zero positive cases); and (3) extrapolation to areas where covariate data values fall outside the range observed in surveyed districts (the 95 survey clusters are in 68 unique districts whereas predictions were made for up to 136 districts). Moreover, both spatial and temporal forecasting rely on the strong assumption that the relationships observed in measured clusters and years (in our case 2010–2011) between TB estimates and covariates remain unchanged in non-measured clusters and future years (2018). However, there may be many violations to this assumption (migration patterns, consistency of quality of laboratory over time, lower TB bacteriology rate due to longer transport times, delays in transport, delays in testing, etc.).

The hackathon modellers also faced a number of more practical challenges. First of all, it is important to acknowledge that modellers were given a short turnaround and developed the models with limited human resources. Second, they faced a number of difficulties of working with the prevalence survey raw data, with limited information on data management steps (including imputation) needed beyond what was described in the survey report and scientific publication [20,21]. However, TB survey data are recognized as difficult to analyse, as there always are missing data and patterns of missingness are often associated with the outcome of interest (prevalence). WHO provides ample guidance on strategies to deal with these issues [29], but the specific choices of the Pakistan TB prevalence survey analysts were not available to the hackathon modellers. This underscores the importance of implementing TB prevalence surveys with transparent and reproducible procedures for data management and data analyses as recommended by most good epidemiological practice guidelines, including recently developed guidelines specifically for global health [30]. All underlying programming code and data for the hackathon models are available on an open access Zenodo repository (see Data Availability Statement). We invite interested modellers to access the code and data to improve on our estimates—either with methodological advancements or by including new high-quality predictors of TB prevalence.

## 5. Conclusions

The TB hackathon provided a unique opportunity to compare different TB subnational prediction models, including novel modelling techniques which had never been applied in this domain before. The technical difficulty of the hackathon assignment highlighted the known challenges of satisfying stakeholder information needs (most recent district-wise estimates) while attempting to fit complex statistical methodologies (subnational decomposition and temporal forecast). The large heterogeneity between the various outputs serves as an important note of caution for the future production and use of granular predictive models of TB based on cross-sectional cluster-based prevalence survey data. Nevertheless, by soliciting and contrasting different methodologies for this challenging problem, the hackathon was successful in providing examples of a range of applications for modellers interested in further developing or refining their approaches to subnational disease modelling. Moreover, the NTP's use and plans for these outputs shows that, limitations notwithstanding, they are valued by decision makers and planners.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/tropicalmed7010013/s1>, Table S1: Additional model specifications, by model; Table S2: Model predictions, by district and model; Table S3: Notification to prevalence, by district and model; Figure S1: Histograms of model predictions, by model; Figure S2: Precision vs. point estimate, by model; Figure S3: Pairwise correlations between model predictions.

**Author Contributions:** S.A., E.R., C.M., A.Q.B. and A.L. conceptualised the hackathon. C.M. was in charge of data curation and project administration. E.R. made all visualisations. A.Q.B. and A.L. provided resources and supervision. F.M., J.M.R., P.J.D., M.P., D.S., S.C., G.B., N.J.H., K.E.L., W.T., P.M. and Z.Z.Q. participated in the hackathon and therefore developed and fitted all models presented, including predictions. E.R., F.G., E.-W.A. and A.L. reviewed the models. S.A. wrote the first draft of the manuscript and all authors commented and edited it. All authors have read and agreed to the published version of the manuscript.

**Funding:** No funding was provided for hackathon modelling activities. KIT Royal Tropical Institute (S.A., E.R., C.M.) organised the hackathon within the frame of technical assistance to the Stop TB Partnership's Strategic Initiative, funded by The Global Fund. J.M.R., receives support from the National Institute of Allergy and Infectious Diseases (K01 AI138620). P.J.D. was supported by a fellowship from the UK Medical Research Council (MR/P022081/1); this UK funded award is part of the EDCTP2 programme supported by the European Union. P.M. was funded in whole, or in part, by the Wellcome Trust [200901/Z/16/Z] and for the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. F.G. acknowledges funding from the European Marie Skłodowska-Curie fellowship (MSCA-IF-EF-ST 846873).

**Institutional Review Board Statement:** KIT Royal Tropical Institute did not require ethical review for this study as analyses were performed on aggregated NTP data (by district) or anonymised publicly available data.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All underlying programming code and data for the hackathon models are available in an open access Zenodo repository (zenodo.org, DOI: <http://doi.org/10.5281/zenodo.5112022>).

**Acknowledgments:** The hackathon would not have been possible without Mirjam Bakker's support and valuable contributions to design and implementation. We are also thankful to and Philippe Glaziou for supporting this initiative from its inception and all throughout implementation. We are grateful to Javeriah Shamsi and Rana Muhammad Safdar for their expert review of hackathon model outputs. Model 3 was developed with contributions from Robert C. Reiner, Mingyou Yang, Emma Spurlock, Brigitte Blacker, and Audrey Batzel and Model 4 with contributions from M.W. Arisido.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Glaziou, P.; Floyd, K. *Latest Developments in WHO Estimates of TB Disease Burden*; World Health Organisation: Geneva, Switzerland, 2018. Available online: [https://www.who.int/tb/advisory\\_bodies/impact\\_measurement\\_taskforce/meetings/tf7\\_background\\_4a\\_burden\\_estimates.pdf](https://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/meetings/tf7_background_4a_burden_estimates.pdf) (accessed on 20 November 2020).
2. Alba, S.; Rood, E.; Bakker, M.I.; Straetemans, M.; Glaziou, P.; Sismanidis, C. Development and validation of a predictive ecological model for TB prevalence. *Int. J. Epidemiol.* **2018**, *47*, 1645–1657. [[CrossRef](#)]
3. Shaweno, D.; Karmakar, M.; Alene, K.A.; Ragonnet, R.; Clements, A.C.; Trauer, J.; MDenholm, J.; McBryde, E. Methods used in the spatial analysis of tuberculosis epidemiology: A systematic review. *BMC Med.* **2018**, *16*, 193. [[CrossRef](#)]
4. Mulder, C.; Nkiligi, E.; Kondo, Z.; Scholten, J.N. What to look for when using SUBsET for subnational TB incidence estimates. *Int. J. Tuberc. Lung. Dis. Off. J. Int. Union. Tuberc. Lung. Dis.* **2020**, *24*, 983–984. [[CrossRef](#)] [[PubMed](#)]
5. Ross, J.M.; Henry, N.J.; Dwyer-Lindgren, L.A.; Lobo, A.D.P.; De Souza, F.M.; Biehl, M.H.; Ray, S.E.; Reiner, R.C.; Stubbs, R.W.; Wiens, K.E.; et al. Progress toward eliminating TB and HIV deaths in Brazil, 2001–2015: A spatial assessment. *BMC Med.* **2018**, *16*, 144. [[CrossRef](#)]
6. Chitwood, M.H.; Pelissari, D.M.; Drummond Marques da Silva, G.; Bartholomay, P.; Rocha, M.S.; Sanchez, M.; Arakaki-Sanchez, D.; Glaziou, P.; Cohen, T.; Castro, M.C.; et al. Bayesian evidence synthesis to estimate subnational TB incidence: An application in Brazil. *Epidemics* **2021**, *35*, 100443. [[CrossRef](#)]
7. Prem, K.; Pheng, S.H.; Teo, A.K.J.; Evdokimov, K.; Nang, E.E.K.; Hsu, L.Y.; Saphonn, V.; Tieng, S.; Mao, T.E.; Cook, A. Spatial and temporal projections of the prevalence of active tuberculosis in Cambodia. *BMJ Glob. Health* **2019**, *4*, e010183. [[CrossRef](#)]
8. Rood, E.; Khan, A.H.; Modak, P.K.; Mergenthaler, C.; Van Gurp, M.; Blok, L.; Bakker, M.A. Spatial Analysis Framework to Monitor and Accelerate Progress towards SDG 3 to End TB in Bangladesh. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 14. [[CrossRef](#)]
9. van Gurp, M.; Rood, E.; Fatima, R.; Joshi, P.; Verma, S.C.; Khan, A.H.; Blok, L.; Mergenthaler, C.; Bakker, M.I. Finding gaps in TB notifications: Spatial analysis of geographical patterns of TB notifications, associations with TB program efforts and social determinants of TB risk in Bangladesh, Nepal and Pakistan. *BMC Infect. Dis.* **2020**, *20*, 490. [[CrossRef](#)] [[PubMed](#)]

10. World Health Organisation. *Global Tuberculosis Report 2020*; WHO: Geneva, Switzerland, 2020. Available online: <https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf?ua=1> (accessed on 20 November 2020).
11. Fatima, R.; Harris, R.J.; Enarson, D.A.; Hinderaker, S.G.; Qadeer, E.; Ali, K.; Bassili, A.; Bassilli, A. Estimating tuberculosis burden and case detection in Pakistan. *Int. J. Tuberc. Lung. Dis. Off. J. Int. Union. Tuberc. Lung. Dis.* **2014**, *18*, 55–60. [[CrossRef](#)]
12. DePasse, J.W.; Carroll, R.; Ippolito, A.; Yost, A.; Santorino, D.; Chu, Z.; Olson, K.R. Less noise, more hacking: How to deploy principles from MIT’s hacking medicine to accelerate health care. *Int. J. Technol. Assess Health Care* **2014**, *30*, 260–264. [[CrossRef](#)]
13. Olson, K.R.; Walsh, M.; Garg, P.; Steel, A.; Mehta, S.; Data, S.; Petersen, R.; Guarino, A.J.; Bailey, E.; Bangsberg, D.R. Health hackathons: Theatre or substance? A survey assessment of outcomes from healthcare-focused hackathons in three countries. *BMJ Innov.* **2017**, *3*, 37–44. [[CrossRef](#)]
14. Li, C.; Xiong, Y.; Sit, H.F.; Tang, W.; Hall, B.J.; Muessig, K.E.; Wei, C.; Bao, H.; Wei, S.; Zhang, D.; et al. A Men Who Have Sex With Men-Friendly Doctor Finder Hackathon in Guangzhou, China: Development of a Mobile Health Intervention to Enhance Health Care Utilization. *JMIR MHealth UHealth* **2020**, *8*, e16030. [[CrossRef](#)]
15. Angelidis, P.; Berman, L.; Casas-Perez, M.D.L.L.; Celi, L.A.; Dafoulas, G.E.; Dagan, A.; Escobar, B.; Lopez, D.; Noguez, J.; Osorio-Valencia, J.S.; et al. The hackathon model to spur innovation around global mHealth. *J. Med. Eng. Technol.* **2016**, *40*, 392–399. [[CrossRef](#)]
16. Ghouila, A.; Siwo, G.H.; Entfellner, J.-B.D.; Panji, S.; Button-Simons, K.; Davis, S.Z.; Fadlelmola, F.M.; Ferdig, M.T.; Mulder, N.; Participants, T.D.O.M.H. Hackathons as a means of accelerating scientific discoveries and knowledge transfer. *Genome Res.* **2018**, *28*, 759–765. [[CrossRef](#)]
17. Ferreira, G.C.; Oberstaller, J.; Fonseca, R.; Keller, T.E.; Adapa, S.R.; Gibbons, J.; Wang, C.; Liu, X.; Li, C.; Pham, M.; et al. Iron Hack—A symposium/hackathon focused on porphyrias, Friedreich’s ataxia, and other rare iron-related diseases. *F1000Research* **2019**, *8*, 1135. [[CrossRef](#)]
18. Ramadi, K.; Srinivasan, S.; Atun, R. Health diplomacy through health entrepreneurship: Using hackathons to address Palestinian-Israeli health concerns. *BMJ Glob. Health* **2019**, *4*, e001548. [[CrossRef](#)] [[PubMed](#)]
19. Wu, D.; Ong, J.J.; Tang, W.; Ritchwood, T.D.; Walker, J.S.; Iwelunmor, J.; Tucker, J.D. Crowdsourcing Methods to Enhance HIV and Sexual Health Services: A Scoping Review and Qualitative Synthesis. *J. Acquir. Immune Defic. Syndr.* **2019**, *82* (Suppl. S3), 271–278. [[CrossRef](#)]
20. Qadeer, E.; Fatima, R.; Tahseen, S.; Samad, Z.; Kalisvaart, N.; Tiemersma, E. *Prevalence of Pulmonary Tuberculosis among the Adult Population of Pakistan 2010–2011*; Islamabad TB Care I: Islamabad, Pakistan, 2013.
21. Qadeer, E.; Fatima, R.; Yaqoob, A.; Tahseen, S.; Haq, M.U.; Ghafoor, A.; Asif, M.; Straetmans, M.; Tiemersma, E.W. Population Based National Tuberculosis Prevalence Survey among Adults (>15 Years) in Pakistan, 2010–2011. *PLoS ONE* **2016**, *11*, e0148293. [[CrossRef](#)] [[PubMed](#)]
22. Pakistan Bureau of Statistics. Provisional Province Wise Population by Sex and Rural/Urban—Census 2017 Pakistan. Islamabad. Available online: [http://www.pbs.gov.pk/sites/default/files//DISTRICT\\_WISE\\_CENSUS\\_RESULTS\\_CENSUS\\_2017.pdf](http://www.pbs.gov.pk/sites/default/files//DISTRICT_WISE_CENSUS_RESULTS_CENSUS_2017.pdf) (accessed on 20 November 2020).
23. Pakistan Bureau of Statistics. Provisional Province Wise Population—Census 2017 Pakistan. Islamabad. Available online: [http://www.pbs.gov.pk/sites/default/files//DISTRICT\\_WISE\\_CENSUS\\_2017.pdf](http://www.pbs.gov.pk/sites/default/files//DISTRICT_WISE_CENSUS_2017.pdf) (accessed on 20 November 2020).
24. James, S.L.; Abate, D.; Abate, K.H.; Abay, S.M.; Abbafati, C.; Abbasi, N.; Briggs, A.M. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **2019**, *393*, e44. [[CrossRef](#)]
25. Bertarelli, G.; Ranalli, G.; Bartolucci, F.; d’Alò, M.; Solari, F. Small area estimation for unemployment using latent Markov models. *Surv. Methodol.* **2018**, *44*, 167–192.
26. Philemon, M.D.; Ismail, Z.; Dare, J. A Review of Epidemic Forecasting Using Artificial Neural Networks. *Int. J. Epidemiol. Res.* **2019**, *6*, 132–143.
27. López-Martínez, F.; Núñez-Valdez, E.R.; Crespo, R.G.; García-Díaz, V. An artificial neural network approach for predicting hypertension using NHANES data. *Sci. Rep.* **2020**, *10*, 10620. [[CrossRef](#)]
28. Mollalo, A.; Mao, L.; Rashidi, P.; Glass, G.E. A GIS-Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States. *Int. J. Environ. Res. Public Health* **2019**, *16*, 157. [[CrossRef](#)] [[PubMed](#)]
29. *Tuberculosis Prevalence Surveys: A Handbook*; World Health Organization: Geneva, Switzerland, 2011; Available online: [http://www.who.int/tb/advisory\\_bodies/impact\\_measurement\\_taskforce/resources\\_documents/thelimebook/en/](http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/resources_documents/thelimebook/en/) (accessed on 30 March 2021).
30. Alba, S.; Verdonck, K.; Lenglet, A.; Rumisha, S.F.; Wienia, M.; Teunissen, I.; Straetmans, M.; Mendoza, W.; Jeannotet, D.; Weibel, D.; et al. Bridging research integrity and global health epidemiology (BRIDGE) statement: Guidelines for good epidemiological practice. *BMJ Glob. Health* **2020**, *5*, e003236. [[CrossRef](#)] [[PubMed](#)]