BMJ Public Health

Predicting communities with high tuberculosis case-finding efficiency to optimise resource allocation in Pakistan: comparing the performance of a negative binomial spatial lag model with a Bayesian machine-learning model

Christina Mergenthaler ,[1,2] Jake D Mathewson,[1] Stephanie Lako,[1] Andreas Werle van der Merwe,[3] Matthys Potgieter,[3] Vincent Meurrens,[3] Abdullah Latif,[4,5] Hasan Tahir,[4] Tanveer Ahmed,[5] Zia Samad,[6] Frank Cobelens,[2] Daniella Brals,[2] Mirjam I Bakker,[1] Ente Rood[7]

To cite: Mergenthaler C, Mathewson JD, Lako S, *et al.* Predicting communities with high tuberculosis case-finding efficiency to optimise resource allocation in Pakistan: comparing the performance of a negative binomial spatial lag model with a Bayesian machine-learning model. *BMJ Public Health* 2025;**3**:e001424. doi:10.1136/bmjph-2024-001424

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2024-001424).

Received 6 May 2024
Accepted 10 February 2025

For numbered affiliations see end of article.

Correspondence to
Dr Christina Mergenthaler; c.mergenthaler@kit.nl

ABSTRACT

Introduction Despite progress in tuberculosis (TB) treatment coverage in past years, an estimated 183 000 people with TB may not have been diagnosed in Pakistan in 2022. Therefore, there is a need to develop models which help to steer active case finding (ACF) towards populations with a high probability of having undetected TB. The aim of this study was to cross-validate TB positivity rate predictions in ACF settings of an existing Bayesian machine learning (BML) with a simpler frequentist model.

Methods We conducted a retrospective analysis of cross-sectional data to identify predictors for detection of bacteriologically confirmed TB cases during ACF events in Pakistan. A predictive negative binomial regression (NBR) model was created, and the presence of spatial autocorrelation was examined to account for spatial dependencies in the outcome variable. The NBR and BML models were compared on their respective predictive precisions for the identification of TB hotspots, based on Root Mean Square Error values, k-fold cross-validation and tehsil-level (sub-district) prediction rankings.

Results 407 (1.9%) bacteriologically confirmed cases among 21 227 visitors were detected in 414 ACF events between September 2020 and January 2022. In the final NBR, the spatial lag variable explained most variation in TB positivity rates across ACF events. NBR and BML predictions were similar at tehsil level. While the BML had a slightly lower root mean squared error (1.02 vs 1.03) the NBR had a slightly better fit based on the Akaike information criterion.

Conclusions Statistical models can be effective in predicting TB hotspots for ACF planning, and the relatively simpler NBR model was nearly as effective as a more complex BML model. The predictions of different modelling approaches were similar, suggesting that predictions are more driven by covariates rather than modelling framework. The agreement between model results increases confidence in the potential utility of models to spatially target ACF activities in high need, low access areas.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ It is practical and appropriate to base tuberculosis (TB) active case-finding (ACF) locations on notification data, information on yield from previous ACF and built relationships with local public and private sector partners. However, ACF is resource-intensive, so it is worthwhile to consider how predictive modelling may support the selection of high case-finding efficiency locations for TB ACF.

WHAT THIS STUDY ADDS

⇒ Our study demonstrates that frequentist and Bayesian modelling frameworks predicted TB positivity rates with similar accuracy (an error of 1 person with TB per 51 people screened) when fitted using the same variables and data, at the tehsil level in Pakistan.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The models presented in this manuscript provide a practical level of discrimination when the objective is prioritising higher need and lower access locations where TB positivity rates found through ACF can be optimised.

INTRODUCTION

Globally, tuberculosis (TB) killed an estimated 1.3 million people in 2022.[1] The COVID-19

Table 1 Study population characteristics at chest camp, tehsil and district level

Chest camp characteristics	Level	Median N	TB care cascade %* %	IQR Q1	IQR Q3
Number of chest camp visitors	Chest camp	51	–	11.9	17
Number of presumptive TB cases identified (% of chest camp visitors)	Chest camp	12	23.5	8	16
Number of X-rays performed (% of chest camp visitors)	Chest camp	48	94.1	39	54
Number of CAD scores above 70 (% of X-rays performed)	Chest camp	2	4.2	2	5
Number of lab tests done (% of presumptive TB cases identified)	Chest camp	11	91.7	8	15
Number of bacteriologically confirmed TB cases identified (% of sputum samples examined)	Chest camp	1	9.1	0	2
Number of all forms TB cases identified (% of sputum samples examined)	Chest camp	2	18.2	1	4
Population characteristics	**Level**	**Mean**	**SD**	**Min**	**Max**
TB epidemic indicators and risk factors					
Demographics and risk factors for tuberculosis					
Average household size	Tehsil	6.6	0.6	5.9	8.9
Indoor air pollution (%)	District	55.8	19.4	15.8	83.0
Migration rate (%)	District	6.3	2.1	1.1	10.6
Population density (number of people per square hectare)†	Tehsil	8.15	--	4.45	28.4
Proportion of population living in urban areas (%)†	Tehsil	15.6	--	11.8	31.7
HIV case notification rate per 100 000 population (2021)	Tehsil	21.9	23.4	3.2	147.5
COVID-19 case notification rate per 10 000 population (2021)	District	4.92	2.75	1.22	9.38
Socioeconomic factors					
Proportion of population considered as literate (%)	District	60.8	8.6	50.0	74.0
Proportion of population living in poverty (%)	Tehsil	55.6	9.7	33.3	73.7
Local health system response					
Tuberculosis programme performance					
New and relapse B+TB CNR per 100 000 population (2021)	Tehsil	53.1	33.4	9.9	148.3
All forms of tuberculosis CNR per 100 000 population (2021)	Tehsil	159.7	91.90	21.9	407.4
Access to healthcare					
Average distance (kilometres) to nearest healthcare facility	Chest camp	6.4	4.72	0.03	24.3
Vaccination coverage rate (%)	District	80.6	5.55	69.0	91.0

*TB care cascade % refers to the numerator and denominator described under chest camp characteristics column.
†Median presented instead of mean; Q1 and Q3 IQR presented instead of min and max for non-normally distributed covariates.
CAD, computer-aided detection; CNR, case notification rate; TB, tuberculosis.

pandemic had a disruptive effect on TB diagnosis and treatment, and for the first time in over a decade, the estimated number of TB deaths increased in 2020 and 2021, due most likely to the reduction in the number of people receiving care for TB, estimated at 18% globally in 2020 compared with 2019.[1] Pakistan is among the eight countries accounting for two-thirds of all incident cases globally and continues to be ranked as having the fifth highest burden, accounting for 5.7% of all incident cases in 2022.[2] However, with an estimated treatment coverage (notified cases divided by estimated incident cases) of 70% (51%–95% UI) among 608 000 all forms incident cases in 2022, Pakistan appears to have potentially made a large step forward from the previous year's estimated treatment coverage rate of 55%.[3]

Many well-documented factors explain this TB treatment coverage gap, including the long and unpredictable incubation period, gaps in health system responses, many of which are interlinked with barriers to health service accessibility.[4–6] People with TB may not end up being tested in a timely manner due to distance barriers, stigma and/or having less knowledge about TB symptoms or services, or symptoms not being recognised by healthcare providers.[5]

The majority of people who start on TB treatment are detected and reported through passive case finding

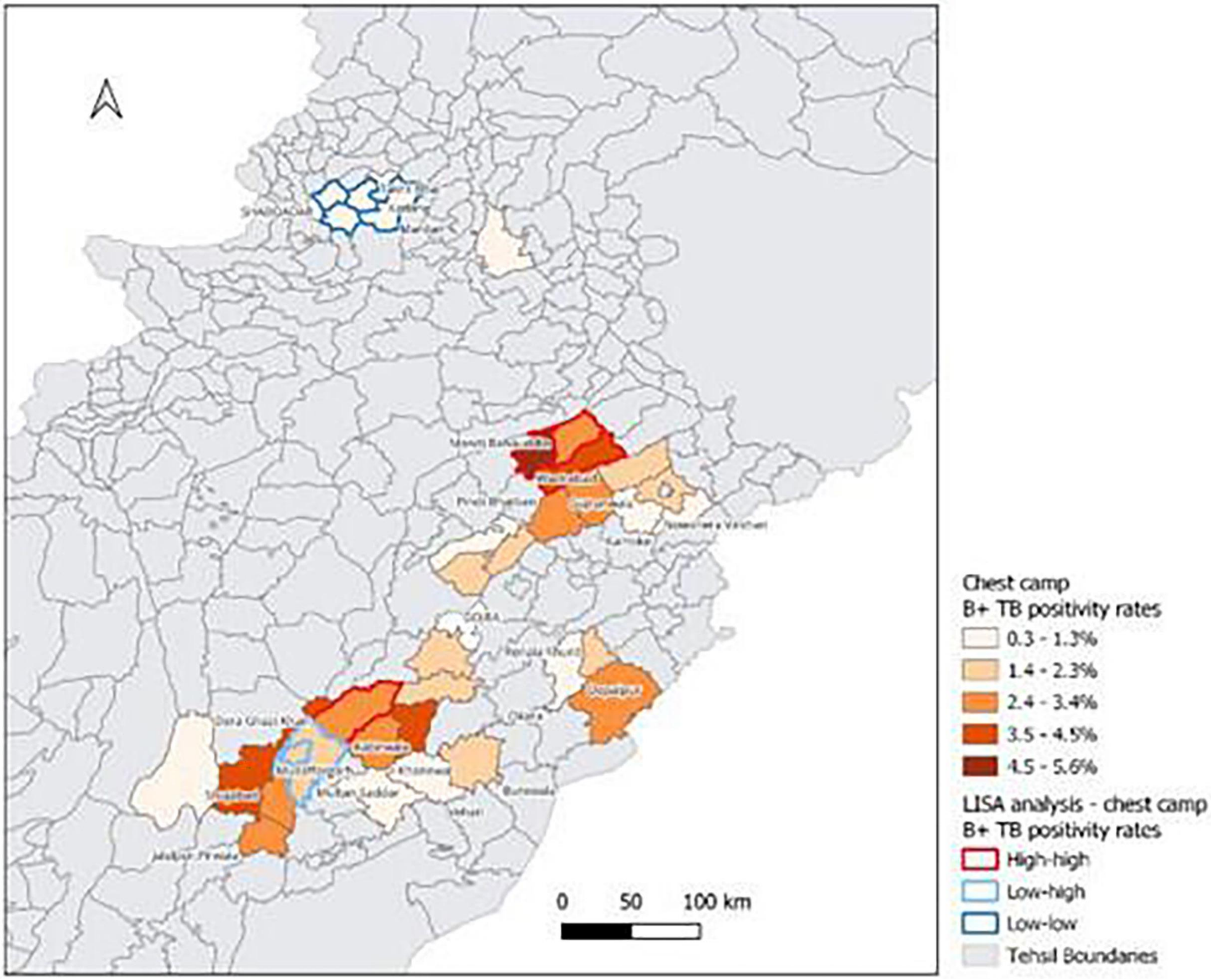

Figure 1 Observed chest camp bacteriologically positive TB positivity rates aggregated to tehsil level and local indicators of spatial autocorrelation analysis. TB, tuberculosis.

(PCF), in which people actively seek medical care in health facilities.[7 8] However, PCF is usually ineffective in closing the treatment coverage gap in many settings.[8–10] Contrary to PCF, active case finding (ACF) relies on screening through outreach activities instead of waiting for people to reach health facilities.[11] Different types of ACF events have been implemented worldwide, sometimes organised as mobile 'chest camps', in which people are screened for TB.[12] However, especially in low-income and middle-income countries, ACF is often constrained by limited human and financial resources. Therefore, there is a need to identify populations with a high probability of having undetected TB to prioritise communities for ACF scale-up.[5 13]

Predictive models may potentially help to fill this gap, supporting the identification of subnational TB disease patterns to reveal the locations of low treatment coverage areas but also those in which case-finding efficiency could be optimised.[14] Such models could be very useful in the context of a multiyear TB ACF programme, in which a key objective is to steer case-finding efforts towards communities in which the probability of having TB is higher than that of the general population, so that TB case-finding can be done efficiently. This efficiency can be measured by the number needed to screen (NNS) to find one case of TB, and when it is minimised then resources spent on ACF can be optimised.

While there is a huge unmet need for subnational TB burden estimates, more predictive modelling exercises are starting to be developed for TB.[14–18] With the increasing availability of large datasets, machine learning (ML) models are starting to be used in public health decision making[19 20] including subnational disease modelling.[21–23] ML models can also handle complex non-linear relationships between covariates and may also adopt a Bayesian approach of updating prior beliefs with new information to inform and fine-tune predictions. For an ACF intervention, a Bayesian ML (BML) model's predictive accuracy could be improved whenever new NNS results become

Table 2 Predictive negative binomial and Bayesian regression analysis of community TB positivity rates

	Univariate negative binomial regression	Non-spatial multivariate negative binomial regression	Spatial lag multivariate negative binomial regression	Bayesian machine learning model
	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	MIC*
Covariate				
Constant	–	0.23 (0.04, 1.45)	0.15 (0.22, 0.99)†	
Tuberculosis programme performance				
All forms of tuberculosis CNR‡	1.00 (0.99, 1.00)	–	–	0.768
Demographic and risk factors for tuberculosis				
Average household size	0.68 (0.55, 0.84)†	0.77 (0.63, 0.94)†	0.83 (0.67, 1.02)	1.505
Indoor air pollution (%)	1.00 (0.99, 1.00)	–	–	1.547
Migration rate (%)	0.93 (0.88, 0.98)†	0.87 (0.83, 0.92)†	0.92 (0.86, 0.98)†	1.779
Population density (per square hectare)	0.96 (0.94, 0.99)†	–	–	0.537
Proportion of population living in urban areas (%)	1.00 (0.99, 1.01)	–	–	1.556
HIV case notification rate‡	1.00 (1.00, 1.01)	–	–	0.820
COVID-19 case notification rate‡	0.99 (0.99, 1.00)†	0.96 (0.92, 1.00)	0.96 (0.92, 1.00)	1.534
Smoking rate (%)	1.08 (1.04, 1.11)†	–	–	0.994
Socioeconomic factors				
Proportion of population considered to be literate (%)	1.02 (1.01, 1.04)†	1.03 (1.02, 1.05)†	1.02 (1.00, 1.04)	
Proportion of population living in poverty (%)	0.99 (0.98, 1.00)	–	–	1.973
Access to healthcare				
Average distance (km) to nearest healthcare facility	1.00 (0.97, 1.02)	–	–	1.676
Vaccination coverage rate (%)	1.00 (0.98, 1.02)	–	–	1.715
Under five mortality rate (%)	0.96 (0.94, 0.97)†	0.97 (0.96, 0.99)†	0.98 (0.96, 1.00)	
Spatial lag of outcome variable			1.45 (1.08, 1.93)†	1.682
Akaike information criterion	–	1046.83	1042.72	1061.31
Average root mean squared Error	–		1.03	1.02
Global Moran's I of the residuals	–	0.016	−0.03	
P value of the Global Moran's I	–	0.260	0.179	

*MIC Bayes set of predictors reflects those included in the final multivariate NBR regression.
†Indicates statistical significance (p<0.05).
‡Case notification rate per 100 000 population (2021) for TB and HIV; per 10 000 population for COVID-19.
IRR, incidence rate ratio; MIC, mutual information criterion; NBR, negative binomial regression; TB, tuberculosis.

available to train the model, hence its attractiveness. However, given the substantial resources required for ACF, it is particularly important to cross-validate BML predictions with other modelling approaches,[14 24 25] including other well understood and broadly applied statistical approaches.[26] Therefore, the aim of this study was to develop a frequentist model and compare its predictions with an existing BML model[21] for the prediction of TB positivity rates detectable through ACF in Pakistan.

METHODS

Study design and population

We conducted a retrospective analysis of cross-sectional data collected during 510 TB chest camps conducted between 7 September 2020 and 29 January 2022. Chest camps were coordinated and conducted across 15 districts in Punjab and Khyber Pakhtunkhwa provinces (online supplemental figure S1). Mobile TB ACF campaigns were conducted by generating awareness for TB in rural communities, and then conducting free screening and

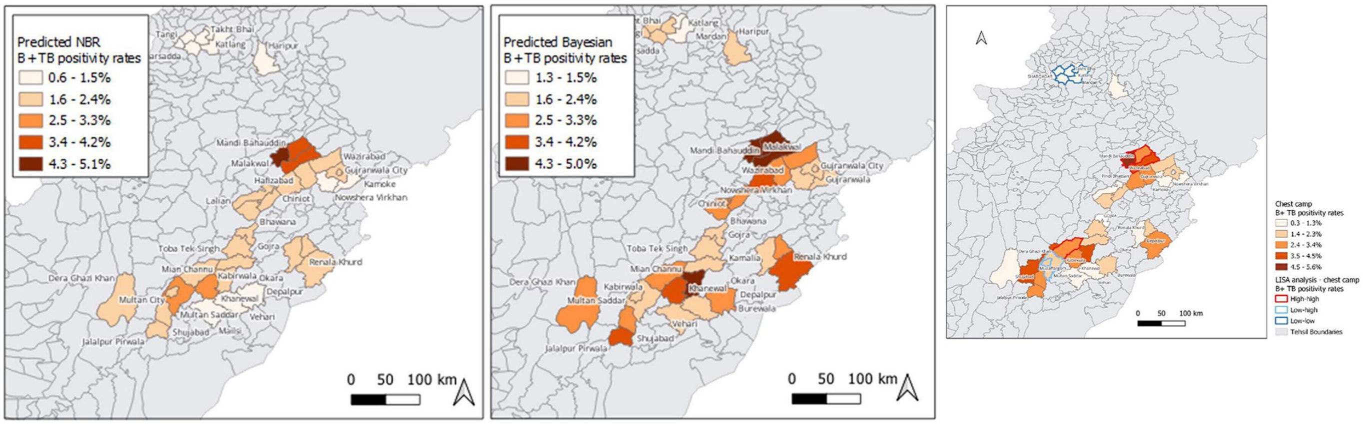

Figure 2 Predicted B+TB positivity rates for negative binomial and Bayesian machine learning models. NBR, negative binomial regression; TB, tuberculosis.

diagnosis for TB on the following day (chest camps). In agreement with community leaders, vans outfitted with X-ray and (usually) GeneXpert equipment was placed in a central spot to gain maximum visibility. The locations where vans conducted chest camps were selected using a scoring system prioritising the following criteria: four or more newly reported TB cases in the last quarter, the presence of any drug-resistant TB cases notified in the last quarter, the location being more than 3 km away from the nearest TB facility, and no prior chest camps having been conducted.

All chest camp attendees were first informed of the chest camp procedure and then asked to provide consent to be screened and tested for TB. After providing consent on paper, attendees were eligible for verbal symptom screening, assessing presence of cough (duration), expectoration, weight loss, chest pain, fever, loss of appetite, night sweats and chills, and history of TB contact. X-rays at most chest camps were offered to people with a positive symptom screen; however, in some chest camps in which there were relatively fewer participants, all participants were offered X-ray. X-rays were scored from 0 to 100 using computer-aided detection (CAD) software, with scores 70 and above classified as abnormal and requiring sputum collection. X-ray exclusion criteria included children under the age of 14 as well as pregnant women, although both groups were eligible for verbal screening and sputum testing. Sputum tests were conducted on all individuals identified as presumptive, defined as either presence of any symptom or abnormal chest X-ray. Sputum tests were done using Xpert MTB/RIF PCR or fluorescence smear microscopy when PCR was not available. All positive PCR or microscopy results are referred to as bacteriologically confirmed. Primary data were collected at individual level: age and sex, verbal symptoms, X-ray results, clinical evaluation notes, and lab results. Data from the case-based paper forms were then aggregated and entered at the chest camp into Microsoft Excel. Field teams collected the address information or nearest landmark such as school, mosque or cross streets of the chest camp. These details were entered into Google Maps and other open-source tools to approximate the camp's GPS coordinates.

Conceptual framework for model development

Covariates included in the building process for both models were identified based on their known association with TB according to the literature, and as conceptualised by the MATCH conceptual framework (Mapping and Analysis for Tailored Disease Control and Health System Strengthening), which categorises variables into one of three domains: indicators of the local TB epidemic; indicators of the local health system response; and indicators of the local high-risk populations and locations.[5]

Data collection, processing and descriptive analysis

Tehsil-level (the first administrative level below the district) quarterly new and relapse and all forms of TB notifications from 2019 to 2021 were provided by Pakistan's National Tuberculosis Programme (NTP) through their DHIS2 database. Quarterly HIV notifications were provided at the tehsil level by Pakistan's National AIDS Control Programme, and COVID-19 notifications were retrieved from the Johns Hopkins University COVID-19 database. We applied a distance-based nearest neighbour analysis to impute data for tehsils with missing HIV notifications. 2021 case notification rates (CNRs) for TB, HIV and COVID-19 standardised to the population based on the population census database.[27] We collected secondary data for TB risk factors including under-5 child mortality rate, malnutrition and smoking from Pakistan's Demographic and Health 2017–2018 Survey database.[28] Literacy, migration, vaccination coverage and indoor air pollution rates were retrieved from Pakistan's 2019–2020 Social and Living Standards Measurement Survey.[29] Population density and poverty rates were extracted from WorldPop, while population census and urban/rural stratifications and average household sizes were accessed from the Pakistan Bureau of Statistics' 2017 census.[27,30] We included a proxy variable measuring access to TB services by performing a nearest neighbour analysis from each chest camp to the nearest healthcare

Table 3 Comparison of spatial negative binomial regression and Bayesian regression predictions of tehsil-level rates of bacteriologically positive TB

Highest observed B+TB positivity rates		Highest predicted B+TB rates—spatial lag negative binomial regression		Highest predicted B+TB rates—ML Bayesian model	
Tehsil	**B+TB rate %**	**Tehsil**	**B+TB rate % (95% CI)**	**Tehsil**	**B+TB rate % (credible interval)**
1 Malakwal	5.57	1 Malakwal	5.14 (4.59, 5.69)	1 Mian Channu	4.19 (0.73, 8.84)
2 Phalia	4.07	2 Mandi Bahauddin	3.62 (3.10, 4.13)	2 Malakwal	3.96 (0.50, 8.99)
3 Mian Channu	3.77	3 Khanewal	3.30 (2.75, 3.84)	3 Mandi Bahauddin	3.85 (1.05, 8.25)
4 Khanewal	3.36	4 Phalia	3.08 (2.59, 3.57)	4 Phalia	3.75 (1.19, 7.73)
5 Mandi Bahauddin	3.33	5 Depalpur	2.57 (1.78, 3.35)	5 Kabirwala	2.76 (0.62, 7.14)
Lowest observed B+TB positivity rates		**Lowest predicted B+TB rates—spatial lag negative binomial regression**		**Lowest predicted B+TB rates—ML Bayesian model**	
1=lowest; 5=5th lowest		**1=lowest; 5=5th lowest**		**1=lowest; 5=5th lowest**	
Tehsil	**B+TB rate %**	**Tehsil**	**B+TB rate % (95% CI)**	**Tehsil**	**B+TB rate % (credible interval)**
1 Shabqadar	0.00	1 Katlang	0.69 (–0.23, 1.61)	1 Mardan	0.92 (0.04, 3.45)
2 Gojra	0.00	2 Shabqadar	0.72 (–0.21, 1.65)	2 Takht Bhai	0.94 (0.04, 4.39)
3 Katlang	0.30	3 Charsadda	0.76 (–0.11, 1.63)	3 Haripur	0.96 (0.04, 5.02)
4 Lalian	0.40	4 Tangi	0.77 (–0.12, 1.66)	4 Katlang	1.04 (0.05, 4.42)
5 Takht Bhai	0.80	5 Mardan	0.78 (–0.05, 1.62)	5 Tangi	1.04 (0.05, 4.67)

B+TB %, bacteriologically confirmed tuberculosis positivity rate among people screened; ML, machine learning.

facility according to Pakistan's NTP in QGIS.[31] Once each chest camp event was paired with a healthcare facility, we measured the distance in kilometres based on Euclidean distance. We retrieved province, district and tehsil shape files from the GADM database.[32] An overview of all considered covariates is presented in online supplemental table S1, grouped by MATCH pillar, although the presence of high-risk populations is missing as no geo-located high-risk group data were identified.

Outcome and exposure variables

Per chest camp (the unit of analysis), the number of people diagnosed with bacteriologically confirmed (B+) TB was defined as the outcome variable, and the number of people screened as the exposure variable.

All covariates were merged with the outcome variable at the chest camp level, and one dataset was created for the statistical analyses. Chest camps conducted in the same tehsil were assigned the same tehsil-level data.

We calculated the mean, SD and range across chest camps for all variables included in the model, and we also presented median and interquartile ranges for non-normally distributed data. For variables with missing data (max=3.9% per covariate), we imputed the observations by performing the multiple imputation by chained equations method.[33 34] All non-spatial data processing and univariate and multivariate regression analyses were performed using Stata V.15.0.[35] All covariates were included as candidates in both models.

Assessment of spatial autocorrelation in the outcome variable

The presence of spatial autocorrelation (ie, spatial clustering) of the outcome variable was checked to determine whether an adjustment for spatial dependency was needed in the final multivariate models. We conducted a local indicators of spatial autocorrelation (LISA) test and calculated the global univariate Moran's I based on the GPS coordinates of the chest camps.[36] A distance weight was created using a queen contiguity matrix order of one neighbour for this test using GeoDa V.1.20.0.8.[36]

Bayesian ML

The reference BML model examined the number of B+TB diagnoses during chest camps as the outcome and the number of people screened in the same camps as the exposure. It employed a directed Bayesian network to represent the interactions of covariates listed in online supplemental table S1, constructing the network by mining Bayesian structures from the data using mutual information and a minimum description length criterion. To query the model, belief propagation was used, enabling efficient inference of the probability distributions for the outcome variable based on observed evidence within the network. This method propagates probabilistic information through the network's structure, allowing the model to update beliefs about the number of B+TB diagnoses when new data from chest camp events or other covariates are introduced, predicting the most probable value as well as 95% credible interval for each prediction.[37]

Frequentist model development

The B+TB count was overdispersed with many zeros; therefore, a negative binomial regression (NBR) model was fitted to the data, using the number of people screened for TB as the exposure variable. Univariate NBR associations were calculated for all covariates with the outcome variable. When collinearity was observed (correlation coefficients <–0.7 or >0.7), the variable with the smaller correlation coefficient with the outcome variable was removed.[38] For the NBR model, we performed backward stepwise selection until all covariates were significant in the final non-spatial multivariate model (p<0.1).

We reported incidence rate ratios and 95% CIs for all NBR univariate and multivariate analyses, and the global Moran's I statistic of the model's residuals for the multivariate NBRs.

We evaluated the fit of the spatial negative binomial model using the Akaike information criterion (AIC), which provides an estimate of the likelihood of the model fitting the data while accounting for the number of parameters used in the model to prevent overfitting.

Comparison of models and internal validation

We compared outcomes of the spatial NBR and the BML models. We conducted k-fold cross-validation using five-fold, and we calculated the average root mean square error (RMSE) of both models to compare the predictive precision. To compare accuracy of the predicted B+TB positivity rates, we ranked tehsils from highest to lowest positivity rates for both observed and predicted, including predicted 95% CIs. NBR and BML models' predicted B+TB positivity rates were compared with the observed rates using a non-parametric Spearman's rank correlation test. Modelled rates were presented in choropleth maps using QGIS V.3.16.16.[31]

RESULTS

Descriptive analysis

Among the 510 chest camp events conducted in the study period, 96 (23%) could not be geolocated and were, therefore, dropped from the dataset, as the geolocation was needed for spatial analysis, leaving a total of 414 chest camp events in 40 tehsils in the final dataset. The 96 dropped chest camps appeared to be randomly geographically distributed across all tehsils.

During the study period, a total of 21 227 individuals attended 414 chest camps, where 407 (1.9%) people were diagnosed and confirmed with B+TB. On average, one person with B+TB was diagnosed per chest camp, out of a median of 51 people appearing at the camp and agreeing to be verbally screened. Predominantly rural populations were visited (77.2%) while there was relatively even gender participation, with females making up 54.3% of all chest camp participants. Descriptive statistics of chest camp, tehsil-level and district-level characteristics are presented in table 1.

A median of 2 abnormal chest X-rays was detected out of 48 X-rays taken (CAD score above 70), with one B+TB case among two all-forms TB cases diagnosed. With regard to routinely conducted passive case-finding, the mean new and relapse bacteriologically confirmed TB CNR per 100 000 was 53, and 160 for the mean all-forms TB CNR. The mean distance between villages where chest camps were conducted and the nearest health facility was 6.39 km, with an average vaccination coverage rate of 80%. Literacy rates were approximately 60% with approximately 55% of the screened population estimated to be living in poverty.

Spatial analysis

The three tehsils with the highest observed B+TB positivity rates were Malakwal (5.6%), Phalia (4.1%) and Muzaffargarh (4.0%), while the three tehsils with the lowest observed B+TB positivity rates were Gojra, Khanpur and Shabqadar at 0.0% each (figure 1). The B+TB positivity rates showed significant spatial clustering with two hotspots (two locations where tehsils with high B+TB positivity rates clustered together), one coldspot (one location where tehsils with low B+TB positivity rates clustered together) and one low spatial outlier where a low B+TB positivity rate tehsil is surrounded by higher B+TB positivity rate tehsils (Moran's I=0.441; p<0.001). Only Malakwal and Phalia formed part of the northern hotspot.

Regression analysis

Univariate analysis showed a number of factors significantly associated with chest camp B+TB positivity rates. Household size, migration rate, COVID-19 notification rate, population density and malnutrition rate were negatively associated with B+TB positivity rates, while higher smoking rate and literacy rate were associated with higher B+TB positivity rates (table 2). The Global Moran's I of the non-spatial NBR model's residuals was not statistically significant (I=0.016; p=0.260); therefore, the final NBR did not require a spatial lag variable, yet as the outcome variable did show modest autocorrelation (I=0.132; p<0.01) this was added to compare model performance. In the spatial lag NBR model, only migration rate and the spatial lag remained significant (and positively) associated with the observed TB positivity rates. The Global Moran's I of the model's residuals remained insignificant (I=−0.030; p=0.179), indicating that any spatial autocorrelation has been accounted for in the model. The BML model showed the smoking rate, followed by migration rate and population density to have the highest mutual information criterion values.

Model comparison and internal validation

Figure 2 and table 3 show that predicted B+TB positivity rates at tehsil level were very similar between the NBR and BML model, with higher predictions overall, and stronger clustering of predicted higher positivity rates in Mandi Bahauddin, Malakwal and Phalia for the BML model, and the same tehsils identified as a high-high cluster through the LISA analysis. The range of predicted values was very similar between models and only minor differences in the minimum and maximum values. Both models predicted similarly low positivity rates (0.6%–1.5%) in the low-low (coldspot) cluster in the northwest shown in the LISA analysis (Mardan, Katlang, Tangi, Takht Bhai and Charsadda tehsils). Both models predicted the lowest B+TB positivity rate tehsils less accurately compared with the highest rate tehsils, with only Katlang and Shabqadar captured in the NBR list, and Katlang and Takht Bhai captured in the BML list. The average RMSE following fivefold cross-validation of the spatial NBR was 1.03, and 1.02 for the BML model; the AIC of the NBR was slightly lower than that of the BML (1042.72 and 1061.31, respectively).

DISCUSSION

The average chest camp positivity rates were 2.0% for bacteriologically confirmed TB and 3.9% for all forms of TB, reflecting a NNS of 51 and 25 for B+TB and all forms of TB, respectively. Spatial patterns of observed B+TB positivity rates reinforced the predictions made by both models, with hotspots and cold spots located in the same tehsils where BML and spatial lag NBR predictions were highest and lowest respectively. Spatial lag NBR and BML model metrics were similar, with a modestly higher RMSE (accuracy) of the BML model over the spatial lag NBR model, and improved AIC (probability of fit) of the NBR model. The similarity of both models' average RMSE suggests that the level of error is relatively low, while the AIC values suggest a similar likelihood of fit. An RMSE range of 1.02–1.03 suggests an error of 1 B+TB case per 51 people screened.

This level of precision offers a practical degree of discrimination in case-finding efficiency, as the difference of finding 0 vs 1 person with B+TB is substantial in terms of human resources, time and material costs. A key advantage of following the case-finding predictions of the BML approach is its ability to continuously improve these predictions in an unsupervised manner; however, it remains important to compare predictions with those from a supervised modelling framework as a means of cross-validation.

Given the irregular opportunity of having recent cluster-level prevalence survey data accessible, ACF positivity rates might serve as a reasonable proxy for local TB case-finding efficiency. If available, subnational prevalence survey data are preferable to model TB burden; this was done by Alene *et al* using 13 regional prevalence surveys' data in Ethiopia conducted over a 17-year span, which was fitted to a Bayesian geospatial model for TB prevalence.[18] This modelling exercise in Ethiopia is an example of using TB data resourcefully to generate subnational prevalence estimates which may help to spatially target local case-finding efforts.

Results from ACF potentially reflect the TB positivity rates which can be expected in communities where health facilities are harder to reach. However, chest camp positivity rates will be biased by the screening and diagnostic algorithm applied and the stringency with which the algorithms are followed by different mobile screening staff,[10 39] camp timing and location,[40] and pre-camp awareness-raising activities. There are still additional external factors which may affect peoples' decisions to participate, such as ability to travel, weather, government-imposed restrictions (eg, COVID-related), social norms and perceptions around community health events.[5 38 39] While the ACF screened communities in this study were not selected to represent the population, in the absence of resources to conduct population-representative random sampling, this study's dataset has several important features. These include excellent geographic coverage of large districts containing communities where there are likely to be gaps in treatment coverage, i.e. where proxies for prevalence may offer greatest value. Furthermore, the predicted positivity rates can be tested by conducting follow-up chest camps, to both assess external validity and to test the hypothesis that model-based, spatially targeted case-finding is an effective way to optimise NNS, that is, screening efficiency.[40]

Local TB prevalence studies in Pakistan to date are primarily limited to Karachi and areas in Balochistan, while our study's ACF was conducted primarily in Punjab province.[41–44] However, nationwide district level bacteriologically positive TB prevalence estimates were modelled through Pakistan's TB Hackathon for the year of 2018, using bacteriologically confirmed TB cases from the 2010 TB prevalence survey data as the outcome variable and sociodemographic data and TB programme data as predictors. 2018 bacteriologically positive TB notifications were compared with the prevalence estimates at the district level to create 2018 CNR: prevalence ratio estimates, which can be interpreted as proxies for TB treatment coverage.[14] Our study's hotspot tehsils Mandi Bahauddin, Malakwal and Kabirwala are in districts with an average CNR: prevalence ratio of 0.545 as calculated by the Hackathon, while Multan, the single low spatial outlier showed a district CNR: prevalence ratio of 0.891. This stands to reason assuming that case-finding will be 'easier' or more efficient in hotspot tehsils with a lower CNR: prevalence ratio, suggesting a larger treatment coverage gap, whereas less prevalent TB is available to be detected in areas with a higher CNR: prevalence ratio like Multan.

Faccin *et al* modelled TB burden at a similar level of spatial resolution in DRC's South-Kivu province to that of Alene's Ethiopian study.[17] Faccin used a similar set of risk factors to predict incidence rates at community level as Alene and our study; however, they went further to validate their predictions by assessing individual-level risk during ACF and follow-up lab testing.[17] This study demonstrates the potential value of using publicly available data to model local TB prevalence in geographically marginalised populations, where notifications are less likely to reflect burden. Extensive spatial modelling of TB has been done in Brazil; we did not identify any studies using ACF data; however, recently municipality-level TB incidence estimates in Brazil were generated with a mathematical model and Bayesian spatial framework, trained on notifications.[15]

Limitations

While it is plausible that unobserved significant effects may have been eliminated due to the hierarchical structure of the dataset, our findings underscore the need to account for spatial dependencies between units of observation. Because sputum testing was done primarily among symptom positives (and X-rays were primarily done among symptom positives), subclinical TB is largely unaccounted for in these models, which is potentially a problem equal to or larger in magnitude to the number of missed TB diagnoses among symptomatic TB patients.[45]

Although NBR is familiar to epidemiological modellers, using an ML framework allows for unsupervised learning. ML modelling methods are less well understood and less intuitive for interpretation as the direction of covariate effects cannot be discerned, but only compared with those of other covariates. The main limitation of the NBR model was its inability to characterise relationships between potential predictors and chest camp outcomes, likely because of the hierarchical structure of the dataset, that is, having a very local level outcome variable with predictors primarily at tehsil level. Additional local-level covariates may have improved the model, such as public sector screening and laboratory positivity rates, health system resources, and a proxy for private sector activities. Ideally, we could have externally validated one or both models; however, due to the overdispersion of zeros in our dataset, we did not have enough observations to take a hold-out sample for this step, and we did not identify a suitable out-of-sample dataset. This is an important future step to test the models' generalisability in different locations, times or populations.

CONCLUSIONS

Statistical models can be effective in predicting TB hotspots for ACF at tehsil level. The predictions of different modelling approaches were found to be consistent, suggesting that predictions are more driven by covariate selection rather than by the modelling framework. The agreement between model results increases confidence in the utility of models to spatially target case finding activities towards high need and low access areas. However, it is important to test these hypotheses by comparing prospectively collected ACF results in and out of the geographic prediction sample, using different models (ML and non-ML, Bayesian, non-Bayesian, and with and without spatial lag). This study also provides evidence for the possible value of archiving and spatially indexing ACF data to make such studies possible, and to potentially serve as the foundation for spatially targeted TB ACF steering.

Author affiliations

[1]Centre for Applied Spatial Epidemiology, KIT Royal Tropical Institute, Amsterdam, The Netherlands
[2]Department of Global Health, University of Amsterdam, Amsterdam, The Netherlands
[3]EPCON SA Pty (Ltd), Cape Town, South Africa
[4]Mercy Corps Pakistan, Islamabad, Pakistan
[5]Common Management Unit for AIDS Tuberculosis and Malaria, Islamabad, Pakistan
[6]Common Management Unit for AIDS, Tuberculosis and Malaria, Islamabad, Pakistan
[7]KIT Royal Tropical Institute, Amsterdam, The Netherlands

Acknowledgements We gratefully acknowledge Dr. Basharat Javed of Pakistan's CMU for leading the monitoring and evaluation of all Global Fund supported TB activities during the study period. We also acknowledge Dr. Hasan Tahir, Sidra Azmat and Ms. Aamna Rashid of Mercy Corps, for their leadership of the TB active case-finding activities which generated the outcome data, as well as all Mercy Corps staff for their tireless work conducting TB chest camps, and recording and reporting all of the TB chest camp data.

Contributors CM and AWvdM led data analysis, manuscript development and editing, with substantial inputs from SL and JDM on both. ER supported the conceptualisation of the study. AWvdM, JDM, MP, SL and VM curated data together with AL, TA, HT and ZS. DB, FC, ER and MIB provided critical input on the manuscript. All coauthors contributed to manuscript editing. CM accepts full responsibility for the finished work and/or the conduct of the study, had access to the data and controlled the decision to publish. Machine learning technology (a type of AI) was the basis for the machine learning Bayesian model described and compared with a frequentist model in the manuscript.

Funding Funding for this study was provided by KIT Royal Tropical Institute. Grant/award number: not applicable.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the IRB Committee of Pakistan's Common Management Unit for AIDS, TB and Malaria in 2020 with approval ID F-No-IRB-CMU-2020-10. Only aggregate-level data were analysed from chest camp events.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The non-open source data generated and analysed during this study are owned by Pakistan's Common Management Unit for AIDS, Tuberculosis and Malaria and Mercy Corps Pakistan. Requests for analytical code or selected aggregate-level indicators may be possible by approaching one of the respective coauthors. No protocol has been prepared. This study has not been registered. There has been no patient or public involvement in the study.

ORCID iD

Christina Mergenthaler http://orcid.org/0000-0001-9936-9446

REFERENCES

1 World Health Organization GTP. Global tuberculosis report 2023 | 1.2 TB mortality. 2023. Available: https://www.who.int/teams/

global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023/tb-disease-burden/1-2-tb-mortality [Accessed 29 Nov 2023].
2 Global tuberculosis report 2023. Available: https://www.who.int/publications-detail-redirect/9789240083851 [Accessed 05 Dec 2023].
3 Global tuberculosis report 2022. Available: https://www.who.int/publications-detail-redirect/9789240061729 [Accessed 05 Dec 2023].
4 The difference between latent TB infection and active TB disease | national prevention information network | connecting public health professionals with trusted information and each other. Available: https://npin.cdc.gov/publication/difference-between-latent-tb-infection-and-active-tb-disease [Accessed 11 May 2023].
5 KIT Royal Tropical Institute. The KIT match approach for enhancing tb care coverage. Available: https://www.kit.nl/project/the-kit-match-approach-for-enhancing-tb-care-coverage/ [Accessed 18 Jul 2022].
6 Akinyemi AO. Exploring integrated pest management strategies in the control of the fall armyworm in smallholder organic agriculture farms in Africa. *Agro-Sci* 2021;20:65–70.
7 Shewade HD, Gupta V, Satyanarayana S, *et al*. Active versus passive case finding for tuberculosis in marginalised and vulnerable populations in India: comparison of treatment outcomes. *Glob Health Action* 2019;12:1656451.
8 Ho J, Fox GJ, Marais BJ. Passive case finding for tuberculosis is not enough. *Int J Mycobacteriol* 2016;5:374–8.
9 Uplekar M, Creswell J, Ottmani S-E, *et al*. Programmatic approaches to screening for active tuberculosis. *Int J Tuberc Lung Dis* 2013;17:1248–56.
10 Lönnroth K, Corbett E, Golub J, *et al*. Systematic screening for active tuberculosis: rationale, definitions and key considerations. *Int J Tuberc Lung Dis* 2013;17:289–98.
11 Kranzer K, Afnan-Holmes H, Tomlin K, *et al*. The benefits to communities and individuals of screening for active tuberculosis disease: a systematic review. *Int J Tuberc Lung Dis* 2013;17:432–46.
12 Deya RW, Masese LN, Jaoko W, *et al*. Yield and Coverage of Active Case Finding Interventions for Tuberculosis Control:A Systematic Review and Meta-analysis. *Tuberc Res Treat* 2022;2022:9947068.
13 Creswell J, Khan A, Bakker MI, *et al*. The TB REACH Initiative: Supporting TB Elimination Efforts in the Asia-Pacific. *TropicalMed* 2020;5:164.
14 Alba S, Rood E, Mecatti F, *et al*. TB Hackathon: Development and Comparison of Five Models to Predict Subnational Tuberculosis Prevalence in Pakistan. *Trop Med Infect Dis* 2022;7:13.
15 Chitwood MH, Alves LC, Bartholomay P, *et al*. A spatial-mechanistic model to estimate subnational tuberculosis burden with routinely collected data: An application in Brazilian municipalities. *PLOS Glob Public Health* 2022;2:e0000725.
16 Hippner P, Sumner T, Houben RM, *et al*. Application of provincial data in mathematical modelling to inform sub-national tuberculosis program decision-making in South Africa. *PLoS One* 2019;14:e0209320.
17 Faccin M, Rusumba O, Ushindi A, *et al*. Data-driven identification of communities with high levels of tuberculosis infection in the Democratic Republic of Congo. *Sci Rep* 2022;12:3912.
18 Alene KA, Python A, Weiss DJ, *et al*. Mapping tuberculosis prevalence in Ethiopia using geospatial meta-analysis. *Int J Epidemiol* 2023;52:1124–36.
19 Rawson TM, Ahmad R, Toumazou C, *et al*. Artificial intelligence can improve decision-making in infection management. 545. 2019. Available: http://spiral.imperial.ac.uk/handle/10044/1/69553 [Accessed 30 Nov 2023].
20 Lysaght T, Lim HY, Xafis V, *et al*. AI-Assisted Decision-making in Healthcare. *ABR* 2019;11:299–314.
21 EPCON. Healthtech | EPCON | antwerpen. Available: https://www.epcon.ai [Accessed 30 Nov 2023].
22 Shonchoy AS, Mahzab MM, Mahmood TI, *et al*. Data driven contagion risk management in low-income countries using machine learning applications with COVID-19 in South Asia. *Sci Rep* 2023;13:3732.
23 Zhao N, Charland K, Carabali M, *et al*. Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLoS Negl Trop Dis* 2020;14:e0008056.
24 Sukarna, Wahyuni MS, Syam R. Comparison of Bayesian Spatio-temporal Models of Tuberculosis in Makassar, Indonesia. *J Phys: Conf Ser* 2021;2123:012048.
25 Khan MT, Kaushik AC, Ji L, *et al*. Artificial Neural Networks for Prediction of Tuberculosis Disease. *Front Microbiol* 2019;10:395.
26 Hackenberger BK. Bayes or not Bayes, is this the question? *Croat Med J* 2019;60:50–2.
27 Pakistan Bureau of Statistics. Final results of census-2017. Available: https://www.pbs.gov.pk/content/final-results-census-2017-0 [Accessed 01 Dec 2023].
28 NIPS/Pakistan NI of PS, ICF. Pakistan demographic and health survey 2017-18. Available: https://dhsprogram.com/publications/publication-fr354-dhs-final-reports.cfm [Accessed 01 Dec 2023].
29 Pakistan Bureau of Statistics. Pakistan social and living standards measurement. Available: https://www.pbs.gov.pk/content/pakistan-social-and-living-standards-measurement [Accessed 01 Dec 2023].
30 WorldPop :: Population Density. Available: https://hub.worldpop.org/project/categories?id=18 [Accessed 01 Dec 2023].
31 QGIS Documentation. Available: https://qgis.org/en/docs/index.html [Accessed 24 Nov 2021].
32 GADM. Available: https://gadm.org/download_country.html [Accessed 01 Dec 2023].
33 Multiple imputation by chained equations: what is it and how does it work? Available: https://onlinelibrary.wiley.com/doi/10.1002/mpr.329 [Accessed 03 Dec 2023].
34 Stefvanbuuren/mice: vignettes/overview.Rmd. Available: https://rdrr.io/github/stefvanbuuren/mice/f/vignettes/overview.Rmd [Accessed 03 Dec 2023].
35 Regression with stata chapter 2 – regression diagnostics. Available: https://stats.oarc.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/ [Accessed 21 Feb 2024].
36 GeoDa on Github. Available: https://geodacenter.github.io/ [Accessed 19 Apr 2024].
37 EPCON. EPCON | bayesian network approach. Available: https://www.epcon.ai/bayesiannetworkapproach [Accessed 26 Apr 2024].
38 Vatcheva KP, Lee M, McCormick JB, *et al*. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)* 2016;6:227.
39 World Health Organization. Systematic screening for active tuberculosis: an operational guide. Geneva World Health Organization; 2015.66. Available: https://iris.who.int/handle/10665/181164 [accessed 07 Nov 2023]
40 Stop TB field guide 3: finding missing people with TB in communities. Available: https://stoptb-strategicinitiative.org/index.php/publications/ [Accessed 05 Dec 2023].
41 Madhani F, Maniar RA, Burfat A, *et al*. Automated chest radiography and mass systematic screening for tuberculosis. *Int J Tuberc Lung Dis* 2020;24:665–73.
42 Zaidi SMA, Jamal WZ, Mergenthaler C, *et al*. A spatial analysis of TB cases and abnormal X-rays detected through active case-finding in Karachi, Pakistan. *Sci Rep* 2023;13:1336.
43 Fatima R, Harris RJ, Enarson DA, *et al*. Estimating tuberculosis burden and case detection in Pakistan. *Int J Tuberc Lung Dis* 2014;18:55–60.
44 Wali A, Safdar N, Manair R, *et al*. Early TB case detection by community-based mobile X-ray screening and Xpert testing in Balochistan. *Public Health Action* 2021;11:174–9.
45 Stuck L, Klinkenberg E, Abdelgadir Ali N, *et al*. Prevalence of subclinical pulmonary tuberculosis in adults in community settings: an individual participant data meta-analysis. *Lancet Infect Dis* 2024;24:726–36.