**TITLE**: Tuberculosis – who is at highest risk? Derivation and validation of a risk score for predicting tuberculosis disease in adult household contacts.

**AUTHORS:** Matthew J. Saunders1,2,3, Marco A. Tovar1,2, Karine Zevallos1, Tom Wingfield1,2,3,4, Sumona Datta1,2,3,5 Rosario Montoya1, Teresa R. Valencia2, Carlos Santillan1,2, Alejandro Necochea1, Matthew R Baldwin1, Robert E. Black, Robert H. Gilman5, Jon S. Friedland3, Carlton A. Evans1,2,3

**AUTHOR AFFILIATIONS:**

1. Innovación Por la Salud Y Desarrollo (IPSYD), Asociación Benéfica PRISMA, Lima, Perú
2. Innovation For Health And Development (IFHAD), Laboratory of Research and Development, Universidad Peruana Cayetano Heredia, Lima, Perú
3. Innovation For Health And Development (IFHAD), Infectious Diseases & Immunity, Imperial College London, and Wellcome Trust Imperial College Centre for Global Health Research, London, UK
4. The Monsall Infectious Diseases Unit, North Manchester General Hospital, Manchester, UK
5. Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

**Corresponding Author:**

Dr Matthew Saunders, IFHAD Clinical Research Fellow, Laboratory for Research and Development #218, Universidad Peruana Cayetano Heredia, Av. Honorio Delgado 430, San Martín de Porres 15102, Lima, Perú. Email: matthew.saunders@ifhad.org; Tel: +447817455461

**Short running title:**

**Keywords:**

**Word count:**

**References:**

**Figures:**

**Tables:**

**ABSTRACT**

**BACKGROUND.** Household contacts of TB cases are at high risk of developing TB disease. Strategies are required to identify effective practices to deliver TB screening and chemoprophylaxis to high-risk groups. We developed a model to predict risk of TB disease that could be used to target interventions towards those most in need.

**METHODS.** We identified newly diagnosed index-patients with smear-positive, pulmonary tuberculosis (n=714) and recruited their initially healthy household contacts aged ≥15 years (n=2,017) in Callao, Perú. The association between potential risk factors and tuberculosis was tested by Cox proportional-hazards analysis. 1008 contacts were randomly selected to form a derivation cohort and a risk score was created that was then validated in the remaining 1009 contacts.

**RESULTS.** 7.8% of contacts in the derivation and 8.2% in the validation cohort developed tuberculosis during median 11 years follow-up. Eight independent predictors for tuberculosis were identified and each was assigned a point score proportional to its regression coefficient: Aged 15-17 or >50 years; history of previous TB; body mass index, poverty; exposure to indoor air pollution; hours spent with the index, index-sex male and index smear-positivity grade. We calculated risk scores for each contact and defined low, intermediate and high-risk groups. In the development cohort, the proportions of contacts that developed tuberculosis in each group were 2.2% (95%CI: 1.0-4.2), 7.7% (95%CI: 5.3-11) and 20% (95%CI: 15-26) respectively. In the validation cohort the corresponding proportions were 3.8% (95%CI: 2.2-6.2), 7.1% (95%CI: 6.0-12) and 17% (95%CI: 12-24). The C-statistic for the model was 0.74 in the development cohort and 0.70 in the validation cohort.

**CONCLUSION.** A risk score to predict tuberculosis was derived and validated that could be used to target active case-finding and preventative interventions.

**INTRODUCTION**

Tuberculosis (TB) is responsible for persistent human suffering, causing illness in 9 million people in 2014 and killing 1.5 million of them. Infection and disease predominate amongst poor, hungry and marginalised people with social and financial consequences that are often catastrophic for the patient, their family and their community. In order to meet the ambitious elimination targets set out in the World Health Organization’s (WHO) End TB strategy, bold emphasis must be placed on preventions aiming to address both direct and indirect TB risk factors.

Household contacts of “index” TB cases have a high risk of TB disease and represent an accessible group to which active case finding and preventative interventions can be targeted. In low- and middle- income countries such as Peru, the WHO recommends that contact investigation be conducted for household contacts of all index cases with pulmonary TB. This practice aims to: i) facilitate the early diagnosis of individuals with TB disease in order to initiate treatment and prevent onward transmission (active case finding); ii) identify contacts at high risk of developing TB disease in order to provide preventative therapy (chemoprophylaxis). In Peru, the National TB Program (NTP) struggles to treat and cure all TB patients leaving few resources available for active case finding and prevention among household contacts. Furthermore, tests for latent TB infection are unreliable and currently unavailable making it difficult for prescribers to assess who is most likely to benefit from chemoprophylaxis. In its latest annual report, the WHO is clear that intensified efforts are needed to prioritise TB prevention practices to populations at highest risk. In resource-constrained settings, operational evidence is required to identify how these policies can be most effectively and efficiently delivered to maximise their benefit.

Risk of TB infection and subsequent progression to TB disease among household contacts are associated with: *index case factors*: demographics, smear positivity, cavitary disease, drug resistance, strain, symptom duration and smoking; *household factors*: socioeconomic position, food insecurity, crowding, exposure to pollution and poor ventilation; and *contact factors*: demographics, genetics, poor nutrition, previous TB, BCG vaccine, comorbidities, duration of exposure to the index case, smoking, alcohol abuse, drug abuse and exposure to TB outside of the household. As part of a randomised controlled trial evaluating micronutrient supplementation to prevent TB in household contacts we aimed to prospectively characterise risk factors for developing TB disease. We subsequently aimed to create a TB risk score that could be used to target active case finding and chemoprophylaxis to adult household contacts at highest risk.

**METHODS**

**Design.** A prospective cohort study of household contacts of TB patients embedded within a randomised controlled trial.

**Setting.** Recruitment took place between October 2002 and June 2006 in 16 peri-urban shantytowns in Ventanilla, Callao, Perú. The population of Ventanilla was approximately 184,506 in 2002 and grew to 420,805 in 2014 of which a large proportion is comprised of economic migrants and internally displaced persons. Ventanilla is an area of substantial poverty with high rates of TB disease (209/100,000 people/year during the study period) but low rates of HIV prevalence (<2%).

**Participants**. Operational definitions for key study variables outcomes are shown in Box 1. *Index cases:* in conjunction with the NTP, we consecutively invited patients suffering from laboratory confirmed, smear-positive pulmonary TB, to participate in the study. We deliberately aimed to recruit smear-positive patients in order to focus on household contacts with the highest risk of TB disease. Index cases were eligible to be included if they had at least one household contact over the age of 14 years who agreed to participate in the study and were excluded if they were previously a household contact of a case that we had already recruited. If a patient was registered twice during the study period, we only invited them to participate once. *Household contacts:* were defined as individuals who reported being in the same house as the recruited TB patient for over 6 hours/week in the 2 weeks preceding diagnosis. Contacts were eligible to be included if they were aged over 14 years and were able to give their written, informed consent. Contacts were excluded if they were on TB treatment at the time of recruitment.

**Procedures.** *Field:* Interactions between the research team and participants occurred at household visits. All participants completed a locally developed, piloted and refined questionnaire to record baseline demographics and data on risk factors of interest that we considered relevant after literature review and pre-conceptualisation (Figure 1). For index patients, this contained a detailed assessment of their current illness, including defining the date of symptom onset and the nature of symptoms. For contacts, data were collected on the number of hours they had spent per day with the index case and on the number of nights they had slept in the same room as the index in the two weeks prior to the index initiating treatment for TB. For all participants, height and weight were measured and BMI was calculated. A detailed medical history was taken including information on history of previous TB, BCG vaccination, co-morbidities and medications the participant was taking. Socioeconomic position was measured using a composite household poverty index, incorporating 13 variables including education, housing and assets. *Laboratory:*All index patients were invited to give a sputum sample at recruitment. This was tested by Ziehl-Nielsen smear microscopy and by the MODS assay as previously described. The laboratory results of samples taken by the NTP were also collected and the most positive result available used in data analysis. Patients were considered to have multi-drug-resistant TB if they had microbiological evidence of resistance, or had been commenced on 2nd line drug therapy by the NTP.

**Follow up.** In order to monitor the health of contacts, we visited their households every 2-4 weeks for a median of 6 months after recruitment and asked about the presence of symptoms (cough, fever, shortness of breath, weight loss, fatigue, night sweats, loss of appetite, chest pain, rashes, stomach pain, dyspepsia, diarrhoea, nausea/vomiting and headache) and general well being. Any contact that reported symptoms suspicious for TB was referred to the NTP and also offered sputum smear and culture tests and drug sensitivity testing free of charge. We collaborated with the national TB program using treatment records to identify contacts that developed TB disease until September 2015. In addition, we performed two prevalence surveys in 2006-2007 and in 2012 during which we aimed to visit all recruited households to ask about previous TB episodes, check for symptomatic TB disease, and collect sputum samples to diagnose asymptomatic disease.

**Statistical analysis.** All statistical analyses were performed using Stata (version 13, StataCorp) and all p-values that were generated were two-sided with significance tested at the 5% level. Continuous data were tested for normality and summarised by their means, standard deviations (SD) and 95% confidence intervals (CIs) or medians and interquartile ranges. Categorical data were summarised as proportions with 95% CIs and compared using the *z* test for proportions. Sample was size was opportunistic and thus no power calculations were performed.

**Outcomes.** The primary end point was time to development of TB disease in contacts; defined from the date the index case took their first pill. Co-prevalent TB was defined as TB diagnosed in contacts within 60 days of the first pill and incident TB any diagnosis following this date (Box 1).

**Risk score.** We randomly selected half of the population of recruited contacts as the “derivation cohort”. Kaplan-Meier survival analysis was used to examine time to TB disease and Cox-proportional hazards models were generated to evaluate risk factors with hazard ratios and 95% CIs calculated. Initially all variables underwent univariate analysis. Variables with p<0.1 in univariate analysis were then included in multivariate models which were adjusted for household clustering with non significant variables removed from the final model. We tested the proportional hazards assumption by including each significant risk factor as a time-dependent covariate in the model and assessing for significance. To develop a risk score, we assigned each of the risk factors a number of points proportional to its regression coefficient. A risk score was then calculated for each contact and the population divided into three categories: low-risk, intermediate-risk and high-risk. This risk score was then validated on the remaining contacts who are referred to as the “validation cohort”. For both the derivation and validation cohorts we generated histograms to illustrate differences in risk of TB disease at 1, 2, 5 and 10 years. We calculated the C-statistic for models to examine their predictive accuracy. We subsequently performed a sensitivity analysis during which we excluded episodes of TB in contacts if they reported that their symptoms started prior to recruitment and a further analysis where we only considered incident TB.

**Ethical approval:** was granted by the ethics committees of the Callao Ministry of Health, Peru; Asociación Benéfica PRISMA, Peru; and Imperial College London, UK.

**Funding:** the study was funded by IFHAD (Innovation for Health and Development) and The Wellcome Trust.

**RESULTS**