STUDY PROTOCOL

REVISED The diagnosis of tuberculous meningitis in adults and adolescents: protocol for a systematic review and individual patient data meta-analysis to inform a multivariable prediction model [version 3; peer review: 2 approved]

Tom Boyles[1,2], Anna Stadelman[3], Jayne P. Ellis[4], Fiona V. Cresswell[5,6], Vittoria Lutje[7], Sean Wasserman[8], Nicki Tiffin[8,9], Robert Wilkinson[8,10,11]

[1]Wits Reproductive Health and HIV Institute, University of the Witwatersrand, Johannesburg, Gauteng, 2001, South Africa
[2]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
[3]School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA
[4]Hospital for Tropical Diseases, University College London Hospitals NHS Foundation Trust, London, UK
[5]Clinical Research Department, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
[6]Research Department, Infectious Diseases Institute, Kampala, Uganda
[7]Cochrane Infectious Diseases Group, University of Liverpool, Liverpool, UK
[8]Wellcome Centre for Infectious Disease Research in Africa, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, South Africa
[9]Division of Computational Biology, Integrative Biomedical Sciences, University of Cape Town, University of Cape, South Africa
[10]Department of Medicine, Imperial College London, London, UK
[11]The Francis Crick Institute, London, UK

v3 **First published:** 31 Jan 2019, **4**:19
https://doi.org/10.12688/wellcomeopenres.15056.1
Second version: 01 Oct 2019, **4**:19
https://doi.org/10.12688/wellcomeopenres.15056.2
Latest published: 04 Jan 2021, **4**:19
https://doi.org/10.12688/wellcomeopenres.15056.3

Abstract

Background: Tuberculous meningitis (TBM) is the most lethal and disabling form of tuberculosis. Delayed diagnosis and treatment, which is a risk factor for poor outcome, is caused in part by lack of availability of diagnostic tests that are both rapid and accurate. Several attempts have been made to develop clinical scoring systems to fill this gap, but none have performed sufficiently well to be broadly implemented. We aim to identify and validate a set of clinical predictors that accurately classify TBM using individual patient data (IPD) from published studies.

Methods: We will perform a systematic review and obtain IPD from studies published from the year 1990 which undertook diagnostic testing for TBM in adolescents or adults using at least one of, microscopy for acid-fast bacilli, commercial nucleic acid amplification test for *Mycobacterium tuberculosis* or mycobacterial culture of cerebrospinal fluid. Clinical data that have previously been shown to

Open Peer Review

Approval Status ✓ ✓

	1	2
version 3 (revision) 04 Jan 2021		✓ view
version 2 (revision) 01 Oct 2019		? view
version 1 31 Jan 2019	✓ view	? view

1. **Ravindra Kumar Garg**, King George's Medical University, Lucknow, India

2. **Kym I.E. Snell**, Keele University, Keele, UK

be associated with TBM, and can inform the final diagnosis, will be requested. The data-set will be divided into training and test/validation data-sets for model building. A predictive logistic model will be built using a training set with patients with definite TBM and no TBM. Should it be warranted, factor analysis may be employed, depending on evidence for multicollinearity or the case for including latent variables in the model.
Discussion: We will systematically identify and extract key clinical parameters associated with TBM from published studies and use a 'big data' approach to develop and validate a clinical prediction model with enhanced generalisability. The final model will be made available through a smartphone application. Further work will be external validation of the model and test of efficacy in a randomised controlled trial.

Keywords
Tuberculous meningitis, multivariable prediction rule, machine learning, diagnostics

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the The Francis Crick Institute gateway.

This article is included in the Wellcome Centre for Infectious Diseases Research in Africa (CIDRI-Africa) gateway.

Corresponding author: Tom Boyles (tboyles@wrhi.ac.za)

Author roles: Boyles T: Conceptualization, Methodology, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Stadelman A**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Ellis JP**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Cresswell FV**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Lutje V**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Wasserman S**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Tiffin N**: Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Wilkinson R**: Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome [210772; 104803; 203135; FC0010218].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Boyles T, Stadelman A, Ellis JP *et al.* **The diagnosis of tuberculous meningitis in adults and adolescents: protocol for a systematic review and individual patient data meta-analysis to inform a multivariable prediction model [version 3; peer review: 2 approved]** Wellcome Open Research 2021, **4**:19 https://doi.org/10.12688/wellcomeopenres.15056.3

First published: 31 Jan 2019, **4**:19 https://doi.org/10.12688/wellcomeopenres.15056.1

REVISED **Amendments from Version 2**

The data synthesis section has been completely re-written based on reviewer comments.

Any further responses from the reviewers can be found at the end of the article

Introduction

Tuberculosis remains a major global health problem, with the most lethal and disabling form being tuberculous meningitis (TBM), of which there are more than 100,000 new cases each year[1]. Mortality is high, particularly in children and patients who are co-infected with HIV-1[2]. The diagnosis is often delayed by the insensitive and lengthy culture technique required for disease confirmation, with delayed diagnosis and treatment being important risk factors for poor outcome[1]. Recently introduced nucleic acid amplification tests (NAATs) allow more rapid detection of TBM. Pooled specificity of 98.0% and 90% for Xpert MTB/RIF and Xpert MTB/RIF Ultra respectively, suggest that they are effective rule-in tests with the potential to speed up diagnosis and reduce unnecessary treatments for alternative conditions in some patients. However, the pooled sensitivity is 71.1% and 90% respectively, which is even lower for patients with HIV (58% to 81%)[3]. Given the extremely high mortality if treatment is withheld from patients with TBM, these values are unlikely to be sufficient evidence to withhold treatment when negative in most patients. Improved strategies to rapidly and accurately diagnose TBM are urgently needed[1].

A major stumbling block in TBM research had been the absence of a single reference standard test or standardised diagnostic criteria. In 2010, a committee of 41 international experts in the field developed consensus case definitions for TBM for use in clinical research[4]. These case definitions have helped to standardise research but are not appropriate for use in routine clinical care as they depend on variables such as cerebrospinal fluid (CSF) culture results, which can take up to 6 weeks to become positive and may include brain imaging, which is not available in many resource constrained settings.

Another approach to improving rapid diagnosis in TBM, particularly in resource-limited settings where the majority of cases occur, is to develop and validate multivariable prediction models. At least 10 models have been published for the diagnosis of TBM, but a major limitation is that their performance is variable in different populations and settings[1]. A major reason for heterogeneous model performance across different settings and populations is case mix variation, which refers to the distribution of important predictor variables such as HIV status and age, and the prevalence of TBM. Case mix variation across different settings or populations can lead to genuine differences in the performance of a prediction model, even when the true predictor effects are consistent (that is, when the effect of a particular predictor on outcome risk is the same regardless of the study population)[5].

Recent studies have shown how big datasets can be used to examine heterogeneity and improve the predictive performance of a model across different populations, settings, and subgroups[6–8]. Individual patient data meta-analysis is preferred to aggregate data meta-analysis, as risk scores can be generated and validated, and multiple individual level factors can be examined in combination[9].

Objectives

1. Conduct a systematic review to identify studies that applied systematic diagnostic strategies for TBM in adolescents and adults presenting with meningitis

2. Establish an international collaboration among TBM research groups who are willing to provide individual patient data (IPD)

3. Use IPD to develop a clinical prediction model that estimates the probability of TBM in adolescent and adults, based on clinical and laboratory data that is routinely available within 48 hours of initial evaluation

Secondary objectives include an assessment of the number and quality of studies addressing the diagnosis of TBM, as well as an analysis of demographic and clinical characteristics of cases and non-cases of TBM.

Protocol

A systematic review and IPD meta-analysis will be performed according to Preferred Reporting Items for Systematic review and Meta-Analysis of IPD (PRISMA-IPD) guidelines[10].

Identification of studies

Potentially eligible studies will be identified by an extensive search of electronic databases, manual search of reference lists and by contacting researchers with interest and expertise in meningitis who may have access to unpublished studies.

We have designed a broad search strategy to maximise sensitivity. We will combine medical subject heading (MeSH) and free text terms to identify relevant studies, see Table 1. We will search Medline (accessed via PubMed), Africa-Wide Information and CINAHL (both accessed via EBSCO Host). We will not limit our searches by geographical location. The search will be restricted to studies published after 01 January 1990 and in English. The detailed search strategies will be presented in an online supplementary appendix. Reference lists of the selected articles and reviews will be searched manually to identify additional relevant studies.

Types of studies

Inclusion criteria

- Randomized controlled trials, cross-sectional studies, and observational cohort studies

- Participants presenting to care with clinical meningitis

- Use of at least 1 of microscopy for acid-fast bacilli, commercial nucleic acid amplification test (NAAT) for *Mycobacterium tuberculosis* or mycobacterial culture of CSF to diagnose TBM

- Study includes a minimum of 10 participants aged ≥ 13 years

Table 1. Proposed search terms.

Search	Query
#1	Search tuberculosis meningitis Field: Title/Abstract
#2	Search "tuberculosis, meningeal"[MeSH]
#3	Search cerebral tuberculosis Field: Title/Abstract
#4	Search "brain tuberculosis" Field: Title/Abstract
#5	Search TBM Field: Title/Abstract
#6	Search ((((tuberculosis meningitis) OR "tuberculosis, meningeal"[MeSH Terms]) OR "cerebral tuberculosis") OR "brain tuberculosis") OR TBM
#7	Search "Diagnosis"[Majr]
#8	Search diagnosis or diagnostic Field: Title/Abstract
#9	Search "clinical scores" or "clinical scoring" Field: Title/Abstract
#10	Search "Research Design"[Mesh]
#11	Search predictor* or predictive Filters: Field: Title/Abstract
#12	Search "clinical predict*" Field: Title/Abstract
#13	Search "clinical feature*" Field: Title/Abstract
#14	Search (((#13 OR ((#12) OR ((#11) OR ((#10) OR ((#9) OR #8 OR #7 Filters: Humans
#15	Search #14 AND #6 Filters: Humans

Exclusion criteria

- Case-control studies and case reports/series of patients with confirmed TBM
- Participants taking anti-TB drugs at the time of their evaluation
- Non-English articles
- Studies published before 1990
- Full text unable to be located
- Studies not in humans

Screening and study selection

Duplicate studies will be removed. Study selection will follow the process described in the Cochrane Handbook of Systematic Reviews and PRISMA-IPD statements[10]. Two investigators will independently screen titles and abstracts to remove irrelevant studies. Full text review will be performed on the remaining studies to determine eligibility. Any disagreements will be resolved by consensus or in consultation with a third reviewer.

Data extraction

Data will be extracted on a proforma, independently by two review authors on study level variables: study setting and dates; contact details; inclusion criteria and exclusion criteria, and number of patients. Corresponding authors of studies identified as eligible after full text review will be contacted with a request to provide anonymised individual patient data. IPD for variables that have previously been shown to be predictive of TBM[1] and competing diagnoses will be requested, Table 2. Investigators will be requested to share their anonymised data after obtaining a signed agreement.

Data management

Investigators will be asked to share anonymised individual patient data, preferably electronically using encrypted files and other secure data transfer technologies using standardised data collection forms. Only study collaborators will have access to the combined IPD data available in Box. Box Secure Storage is a cloud storage and collaboration service configured to meet the security standards for HIPAA data. Data will remain stored in Box for the duration of the study and will not be used or sold for any commercial purpose.

Authorship

Authors providing IPD will be asked to nominate co-authors to expand the expertise of the review group, including review of preliminary findings and manuscript authorship. The number of co-authors will depend on the amount of data supplied, 1 author for <100 patients, 2 authors for >100 and <250 patients, and 3 authors for >250 patients.

Quality assessment

Quality assessment in terms of risk of bias and applicability for each included study will be performed according the QUADAS-2 tool for diagnostic accuracy studies[11]. This tool comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and the first 3 domains are also assessed in terms of

Table 2. Individual patient data that will be requested from authors. LAM= lipoarabinomannan NAAT= nucleic acid amplification test.

Clinical data at presentation	Laboratory results (blood)	Laboratory results (CSF)
• Age* • Sex* • Presence of extrapyramidal movements* • Presence of neck stiffness* • Duration of symptoms* • Focal neurological deficit (including cranial nerve palsy)* • Temperature* • Glasgow Coma Scale* • AVPU score*	• HIV sero-status* • Total leukocytes* • CD4 count* • Glucose*	• Appearance* • Total leukocytes* • Total neutrophils* • Total lymphocytes* • Protein* • Glucose* • Gram stain* • Adenosine deaminase activity* • Bacterial culture • India ink stain* • Cryptococcal antigen* and culture • Microscopy for acid-fast bacilli • Mycobacterial culture • NAAT for *Mycobacterium tuberculosis* • NAAT for any virus • Syphilis serology* • Any other test informing an alternative diagnosis
Laboratory results (urine, sputum and serous effusions)	**Radiological investigations**	**Autopsy**
• Urine LAM* • Microscopy for acid-fast bacilli* • Mycobacterial culture • NAAT for *Mycobacterium tuberculosis**	• Chest X-ray* • Abdominal ultrasound scan • CT brain • MRI brain	• Histological results from autopsy

*Factors chosen *a priori* to be used to develop the initial model.

concerns regarding applicability. Signalling questions are included to help judge risk of bias.

Data synthesis

1. Descriptive analysis of available parameters, data completeness check, and IPD meta-analysis.

The contributing datasets will be reviewed for sample size, available variables and data completeness, to inform the selection of a modelling approach. A descriptive analysis will be undertaken to understand similarities and differences between the contributing datasets. Participant characteristics and clinical features (Table 1) will be summarized for each contributing dataset and compared across datasets using chi-square, t-tests, or non-parametric methods as warranted.

2. Selection of Candidate Predictors

The objective of this step will be to reduce the number of variables that go into the development of the TBM prediction model. Prior studies have indicated several predictive variables such as

- Symptom duration prior to presentation at the hospital (days)
- CSF leukocytes
- CSF neutrophil (%)
- CSF glucose
- CSF protein

We aim to include these variables in the predictive model as "primary predictors" in an effort to retain the variables that are the most predictive of TBM diagnosis as well as easily acquired in low resource settings. Primary predictors will be assessed for missingness, imputed if missing (see Step 3 for more detail), and will be used in predictive model development. Other variables that we would like to include, as "secondary predictors" are,

- Age
- Sex
- Blood glucose
- Blood leukocytes
- Country
- HIV status

We aim to include the above secondary predictors in an effort to explore their predictive value of diagnosing TBM. The secondary predictors have been selected based on prior published diagnostic algorithms. They will be assessed for missingness and imputed if missing (see Step 3 for more detail)

but may or may not be included in the final algorithm if their addition to the algorithm does not result in better predictive performance. It is also possible that age, sex, country, and HIV status could explain case-mix variation.

We will also consider employing methods such as principal component analysis and joint individual variation explained[12] to identify the variables that explain most of the variation in TBM diagnosis to retain in the final model(s).

3. Multiple Imputation for missing data

Multiple imputation for this study will be carried out within contributing datasets that have <65% missing data for the primary and select secondary predictors; blood glucose and blood leukocytes. Characteristics of participants with 10–64.99% missing data will be summarized and compared to those with 'complete' data (<10% missing data) to explore the nature of missingness and identify auxiliary variables that could later inform imputation. Comparison characteristics include sex, age, survival time (days), outcome, diagnosis, and TBM case status (definite, probable, possible, not-TBM). If there is no clear pattern of missingness, the data will be assumed missing at random and imputed. After imputation, the fraction of missing information (FMI) statistic will be estimated in the modeling step to ensure that the imputation model is well specified[13]. It is not always reasonable to assume predictors missing ≥65% of data are missing at random. Therefore, we will attempt to determine the underlying mechanism of missingness and not impute for these predictors if the missingness cannot be accounted for by another variable with complete information or if the missing data appear to be missing not at random.

Auxiliary variables such as sex and age, which are typically predictive of most biological values, will be used to help inform imputation. Further auxiliary variables for each missing predictor will be selected based on biologic plausibility. For example, missing CSF glucose will be informed by blood glucose. After auxiliary variables are identified for each missing primary or secondary predictor, missing data will be imputed within each contributing dataset (i.e. not informed with data from other contributing datasets.

Multiple imputation will be carried out in R using the MICE package. Patients with TBM and other types of meningitis are typically acutely ill, therefore we are expecting skewed values for all the biologic metrics and will be utilizing the chained equations approach in the MICE package. We will impute 20 datasets for each missing variable.

4. Developing a Predictive Model

After that we will build a predictive algorithm via IPD meta-analysis using a logistic regression model for the diagnosis of TBM[7]. The first step is to estimate the predictor–outcome associations from the available studies in order to assess heterogeneity in predictors across studies. Predictors that have homogenous predictor-outcome associations will be prioritized in model inclusion, but we will not exclude variables that have heterogenous predictor-outcome relationships across studies. All predictors will be included in a model with a stratified intercept for each study to underscore the baseline predictor-outcome value of each of the predictors in the different contributing datasets[7]. We will also develop a model with a stratified intercept for each country (pooling datasets from the same country) for the purposes of external validation and implementation after the model is developed.

All the data collected from the systematic review will be used in the development of the clinical prediction tool. The objective of this step is to develop three different prediction models. The first model will be developed using logistic regression with backward stepwise variable selection (p-value threshold of 0.1)[14,15]. This is the modeling approach used in prior clinical prediction tools developed for TBM. The second model will be developed using the IPD meta-analysis framework with a stratified intercept for country[7]. As discussed earlier, this approach is appropriate for these data as it encompasses information from multiple contributing datasets in the development of the tool. The final model will be developed using machine learning techniques such as classification and regression trees or random forest classifier analysis.

5. Testing the model for internal validity

The model(s) will be internally validated using the bootstrap and internal-external cross validation approaches. Bootstrap validation is the process for which observations from within each contributing dataset are sampled with replacement to go into the development of the model, the model development analyses are repeated, and then this model is internally validated in the original datasets[16,17]. This validation step will give information of optimism and overfitting. Internal-external cross validation approach is a multiple validation approach that accounts for multiple datasets by rotating which are used toward model development and validation[7]. Each contributing dataset will be excluded from the available set, and the remainder will be used to develop the diagnostic model; the excluded study will then be used to validate the model externally. This process will be repeated with each study being omitted in turn, allowing the consistency of the developed model and its performance to be examined on multiple occasions.

Performance of the developed model(s) will be assessed using calibration and discrimination, metrics for model fit. Calibration is defined as the agreement between observed outcomes and predictions[18]. We will use the ratio of predicted (expected) to observed outcomes, otherwise known as E/O, to assess model calibration. Ideally, the ratio should be close to 1, which represents a calibrated model[7]. Calibration is also related to goodness-of-fit, which relates to the ability of a model to fit a given set of data[18]. The Hosmer-Lemeshow goodness-of-fit test is often used to assess goodness-of-fit with binary outcome data, which can be graphically displayed in a calibration plot. Usually, patients are grouped by decile of predicted probability. A better calibrated model will have the average prediction value

within each decile falling along a 45 degree line in the plot, where the true probability in each decile (y-axis) is equal to the average predicted probability for that group (x-axis)[18].

Discrimination is defined as the ability to generate predictions that discriminate between those with and those without the outcome (e.g. TBM vs not-TBM diagnosis)[18]. We can assess discrimination using a receiver operating characteristic (ROC) curve, which plots the sensitivity (true positive rate) against 1 – specificity (false-positive rate) for consecutive cutoffs for the probability of an outcome (i.e. TBM diagnosis). For the bootstrap approach, ROC curves will be produced as an average for the models that are bootstrapped, and discrimination will be compared using the area under the curve (AUC) c-statistic. For the internal-external cross validation approach, discrimination will be estimated in each study that is excluded from development. The AUC can be interpreted as the probability that a patient with the outcome is given a higher probability of the outcome by the model than a randomly chosen patient without the outcome[18]. The higher the AUC, represented by the curve closer to the upper left-hand corner (higher sensitivity and specificity), the better the model is at predicting the outcome, in this case TBM diagnosis. Furthermore, ROC curves are often used in diagnostic research to quantify the diagnostic value of a test over its whole range of possible cutoffs for classifying patients as positive vs. negative[18]. The curves trend to the upper left corner when the distributions of predictions are more separate between those with and without the outcome.

Optimism or "overfitting" is where the model fits the data so well that it is not valid for new subjects, a key threat to internal validity that needs to be addressed throughout the internal validation process. Model performance statistics are generated and then optimism-corrected by taking the apparent performance and subtracting the optimism. The optimism for a particular statistic is calculated by repeating the model development in each bootstrap sample, calculating the performance in the bootstrap sample (where it will be optimistic), and then applying that model back to the original dataset (acting as a validation dataset). After repeating this process 100–200 times, the average of the differences between these model performance statistics is the estimate of optimism.

As discussed, high optimism is indicative of overfitting, which can be corrected via shrinkage. Shrinkage is defined as the regression of coefficients towards zero as a way to improve model performance[18]. Although we do not anticipate that we will need to employ this method to correct for overfitting because we will have completed the data reduction step described in step 3 above, we will validate the necessity of the shrinkage approach when assessing model optimism.

In addition to assessing model calibration and discrimination, the Brier score will be used to assess overall model performance. The Brier score measures the accuracy of probabilistic predictions, which is a combination of both calibration and discrimination[18,19]. The Brier score calculates the squared differences between actual outcomes and predictions. For a model, the Brier score can range from 0% for a perfect model to 0.25 for a non-informative model. A Brier score will be generated from the model developed in the original dataset.

After the internal validation process, we will select model(s) that perform well and calculate the overall sensitivity, specificity, positive predictive value, and negative predictive to assess the accuracy of the model(s) in predicting TBM. Due to the high mortality among unstable TBM patients a 10% predictive probability will be used as a threshold of a positive case status[20]. Any predictive probability equal to or above 10% will be sufficient to treat a patient for tuberculosis since the outcome of not treating TBM is almost always death[20].

6. Sensitivity analysis

We will perform the following sensitivity analyses to explore the contributions of risk of bias on the final model(s):

- Exclude studies that investigated all patients for TBM regardless of other CSF findings (co-infection). This is important for discerning which clinical characteristics are predictive of TBM or other meningitis-causing diseases.
- Develop the prediction model(s) with different TBM-case status groupings. Confirmed TBM and non-TBM cases will remain as such in model development with probable and possible cases shifting to either category. The model will be developed with the following case status groupings:
    - Confirmed TBM vs. probable, possible, and non-TBM
    - Confirmed and probable TBM vs. possible and non-TBM
    - Confirmed, probable, and possible TBM vs. non-TBM

  We will compare the observed heterogeneity, clustering, predictor selection, and model performance between the models developed with the above case groupings. This is important for two reasons. First, shifting the threshold of inclusion or TBM "caseness" could provide further insight into the misclassification of cases as a result of poor diagnostic strategies for TBM. Second, including the information from probable and possible TBM cases in model development would result in a more applicable model for probable and possible cases, with the intention that the model could better predict TBM for these persons.
- Conduct a misclassification bias analysis. This step is important since there is no "gold-standard" diagnostic criteria for TBM diagnosis so there is likely misclassification bias. Many of the diagnostic methods used to ascertain TBM, such as TB culture in CSF, have known sensitivities and specificities that will be used towards reclassifying cases. Model(s) will be

developed with the reclassified case statuses and compared to the model(s) developed with the original case classification.

We will also assess risk of bias for each included study using the QUADAS-2 tool[11]. This tool comprises four domains: participant selection, predictor measurement, and outcome definition and measurement. Each domain is assessed in terms of risk of bias and are also assessed in terms of concerns regarding applicability. Signaling questions are included to help judge risk of bias.

Registration

This review is registered with PROSPERO, number CRD42018110501.

Presenting and reporting of results

We will report the results according to the Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data Statement (PRISMA-IPD)[10]. This will include a flow diagram to summarise the study selection process and detail the reasons for exclusion of studies screened as full text. We will publish our search strategy and quality-scoring tool as supplementary documents. Quantitative data will be presented in evidence tables of individual studies as well as in summary tables. We plan to report on quality scores and risk of bias for each eligible study. This may be tabulated and accompanied by narrative summaries. A descriptive analysis of the strength of evidence assessment will be reported. The final prediction model(s), that is, the variable-selected model(s) with the highest area under the receiver operating characteristic curve (AUC), will be implemented in a Smart phone application and a Web-based calculator and graphically depicted using nomograms.

Discussion

TBM is a serious public health concern with delayed diagnosis and treatment being important risk factors for poor outcome[1]. At least 10 attempts have been made to develop clinical prediction models to aid the rapid diagnosis of TBM but none have been broadly successful. The aim of this project is to combine data from multiple sources to develop and internally validate a novel clinical prediction model, which will be made easily available as a smart phone application and a Web-based calculator. By combining data from multiple geographical locations and using advanced machine learning techniques it is hoped that we can develop a model that is broadly generalizable around the world. Further work will involve external validation of the model(s) and testing in randomised controlled trials.

Ethics

No specific ethical approval has been sought for this systematic review. Authors who submit IPD will be asked to confirm that the dissemination of anonymised data was included in the original patient consent document.

Data availability

Underlying data

No data is associated with this article.

Reporting guidelines

Figshare: PRISMA-P checklist for The diagnosis of tuberculous meningitis in adults and adolescents: protocol for a systematic review and individual patient data meta-analysis to inform a multivariable prediction model, https://doi.org/10.6084/m9.figshare.7628639.v1

References

1. Wilkinson RJ, Rohlwink U, Misra UK, *et al.*: **Tuberculous meningitis.** *Nat Rev Neurol.* 2017; **13**(10): 581–98.
PubMed Abstract | Publisher Full Text
2. Thwaites GE, van Toorn R, Schoeman J: **Tuberculous meningitis: more questions, still too few answers.** *Lancet Neurol.* 2013; **12**(10): 999–1010.
PubMed Abstract | Publisher Full Text
3. Boyles TH, Thwaites GE: **Appropriate use of the Xpert® MTB/RIF assay in suspected tuberculous meningitis.** *Int J Tuberc Lung Dis.* 2015; **19**(3): 276–7.
PubMed Abstract | Publisher Full Text
4. Marais S, Thwaites G, Schoeman JF, *et al.*: **Tuberculous meningitis: a uniform case definition for use in clinical research.** *Lancet Infect Dis.* 2010; **10**(11): 803–12.
PubMed Abstract | Publisher Full Text
5. Riley RD, Ensor J, Snell KI, *et al.*: **External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges.** *BMJ.* 2016; **353**: i3140.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Ahmed I, Debray TPA, Moons KGM, *et al.*: **Developing and validating risk prediction models in an individual participant data meta-analysis.** *BMC Med Res Methodol.* 2014; **14**: 3.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Debray TPA, Moons KGM, Ahmed I, *et al.*: **A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis.** *Stat Med.* 2013; **32**(18): 3158–80.
PubMed Abstract | Publisher Full Text
8. Jolani S, Debray TPA, Koffijberg H, *et al.*: **Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE.** *Stat Med.* 2015; **34**(11): 1841–63.
PubMed Abstract | Publisher Full Text
9. Riley RD, Lambert PC, Abo-Zaid G: **Meta-analysis of individual participant data: rationale, conduct, and reporting.** *BMJ.* 2010; **340**: c221.
PubMed Abstract | Publisher Full Text
10. Stewart LA, Clarke M, Rovers M, *et al.*: **Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement.** *JAMA.* 2015; **313**(16): 1657–65.
PubMed Abstract | Publisher Full Text
11. Whiting PF, Rutjes AWS, Westwood ME, *et al.*: **QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.** *Ann Intern Med.* 2011; **155**(8): 529–36.
PubMed Abstract | Publisher Full Text
12. Kaplan A, Lock EF: **Prediction With Dimension Reduction of Multiple Molecular Data Sources for Patient Survival.** *Cancer Inform.* 2017; **16**: 1176935117718517.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Madley-Dowd P, Hughes R, Tilling K, *et al.*: **The proportion of missing data should not be used to guide decisions on multiple imputation.** *J Clin*

Epidemiol. 2019; **110**: 63–73.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Moons KGM, Altman DG, Reitsma JB, *et al.*: **Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.** *Ann Intern Med.* 2015; **162**(1): W1–73.
PubMed Abstract | Publisher Full Text

15. Sauerbrei W, Royston P, Binder H: **Selection of important variables and determination of functional form for continuous predictors in multivariable model building.** *Stat Med.* 2007; **26**(30): 5512–28.
PubMed Abstract | Publisher Full Text

16. Steyerberg EW, Bleeker SE, Moll HA, *et al.*: **Internal and external validation of predictive models: a simulation study of bias and precision in small samples.** *J Clin Epidemiol.* 2003; **56**(5): 441–7.
PubMed Abstract | Publisher Full Text

17. Kim JH: **Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap.** *Comput Stat Data An.* 2009; **53**(11): 3735–45.
Publisher Full Text

18. Steyerberg EW: **Clinical prediction models: a practical approach to development, validation, and updating**. New York: Springer; 1 online resource. SpringerLink (Online service), 2009; xxviii, 497.
Publisher Full Text

19. Brier GW: **The Statistical Theory of Turbulence and the Problem of Diffusion in the Atmosphere.** *J Meteorol.* 1950; **7**(4): 283–90.
Publisher Full Text

20. Boyles T, Locatelli I, Senn N, *et al.*: **Determining clinical decision thresholds for HIV-positive patients suspected of having tuberculosis.** *Evid Based Med.* 2017; **22**(4): 132–8.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 3

Reviewer Report 04 February 2021

https://doi.org/10.21956/wellcomeopenres.18198.r41935

✔ **Kym I.E. Snell**
Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

I thank the authors for taking my previous comments on board and addressing many of my concerns. The method section is vastly improved in this latest version. I hope that the authors will consider using risk of bias tools and reporting guidelines developed specifically for diagnostic (and prognostic) model studies such as PROBAST and TRIPOD. An extension for TRIPOD covering clustered data (including IPDMAs) is also underway and likely to be published prior to this study being written up.

I wish the group the best of luck with their research.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, prediction modelling and IPD-MA

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 25 October 2019

https://doi.org/10.21956/wellcomeopenres.16866.r36627

?

Kym I.E. Snell
Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

The authors have addressed some of my previous comments. However, there are several points that I have to still query.

Firstly, I think there has been some confusion in the use of the terms 'case-mix variation' and heterogeneity. As the project includes an IPD-MA, I was referring to differences in case-mix across studies and heterogeneity in performance across studies in line with text in the introduction. Therefore, some of the changes to the protocol are rather confusing.

If the authors intend to look at heterogeneity in predictor effects, their approach (a stratified intercept model with fixed effect for the predictor effect) will not allow them to investigate this as the predictor effect is fixed. If heterogeneity in predictor effects is of interest, then I would suggest placing a random effect on the predictor effect allowing the slope to change by study and reporting estimates of heterogeneity. If this is not of interest, then please remove mention of investigating heterogeneity in predictor effects.

It is not clear what is meant by 'We will also use this method to assess predictor heterogeneity by HIV status, WHO region, and TB burden within each country'? Do you mean evaluate predictors within subgroups of these factors, fit interactions between predictors and these effects, or something else?

I'm also rather confused by the suggestion of classification and regression trees, supervised and unsupervised ML, LCA in relation to case-mix variation. These are just different statistical techniques that can be used to investigate clustering of individuals (by characteristics) and doesn't explicitly say how different studies will be used or what the objective is.

The methods for model development and validation are still vague and suggest to me that the data from multiple studies will be combined and treated as a single dataset rather than as an IPD-MA (accounting for clustering by study). IPD-MA can be very challenging, however there is a lot of guidance on how to conduct them, particularly by T. Debray and R. Riley. Sadly, much of the guidance has either not been considered or been ignored. If using regression modelling, clustering by study could be accounted for by using a random intercept or stratified intercept. It is then also important to decide how the model would be implemented i.e. what intercept would be used for validation in a new study or for new patients in practice (Debray *et al.*, 2013[1])? I am not aware of how clustering can be accounted for when using methods such as CART.

I have some concern about the decision not to impute for missing data, especially given the list of 28 predictors of interest and no minimum sample size that they aim to achieve. In my experience, expecting to receive datasets with 28 predictors of interest is incredibly optimistic. It is more likely that both systematically missing data and partially missing data will be a challenge. If every dataset was missing just one of those variables, there would be no data left to use. It often ends up being a compromise between the number of predictors (and which ones are) considered and the number of datasets included. If then also not imputing for partially missing data, further data is lost and predictor effects may be biased (e.g. only severe cases get certain tests). This could affect your target population and result in models that are not applicable to the population of interest.

Regarding the decision to develop the model in only confirmed TBM and non-TBM patients, I would think that the target population would include probable and possible TBM and therefore there would be an applicability issue if the model was developed using an unrepresentative subgroup of patients (Moons *et al.*, 2019[2]). These probable and possible cases are also often the individuals in which a prediction model has the most potential to be useful. Other options to consider are to include them with one or the other of confirmed TBM and non-TBM so as not to exclude them. Sensitivity may be to change which outcome they're grouped with.

Finally the description of bootstrap validation is not correct. Optimism is not a statistic as such but can be estimated for each performance statistic of interest. The 'optimism-corrected' statistics are then calculated as apparent minus optimism. The optimism for a particular statistic is calculated by repeating the model development in each bootstrap sample, calculating the performance in the bootstrap sample (where it will be optimistic) and also applying that model back to the original dataset (acting as a validation dataset). Repeat lots of times and the average of the differences between these is the estimate of optimism. Estimating optimism by bootstrapping is sufficient, no need to calculate for other validation approaches like IECV.

References

1. Debray TP, Moons KG, Ahmed I, Koffijberg H, et al.: A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis.*Stat Med*. 2013; **32** (18): 3158-80 PubMed Abstract | Publisher Full Text
2. Moons K, Wolff R, Riley R, Whiting P, et al.: PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine*. 2019; **170** (1). Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, prediction modelling and IPD-MA

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 08 March 2019

https://doi.org/10.21956/wellcomeopenres.16426.r34913

?

Kym I.E. Snell

Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK

This is a clear and well-written protocol for a systematic review, collection of IPD and IPD meta-analysis for a new diagnosis model for TBM. I think the aim of the study and details relating to the systematic review part are clear, however I have a few comments/questions for clarity and reproducibility, mostly regarding what happens once IPD has been collected.

1. In the introduction the authors mention that case-mix can affect the predictive performance of a model and that big datasets can be used to examine the heterogeneity and improve the predictive performance. However, I don't think they really address this issue in the methods or say how they will use the IPD to try improve the performance. Heterogeneity in performance if the predictor effects are consistent suggests differences in case-mix that are not being captured by the predictors in the model. Unless additional variables that are thought to improve the model are included, how will this be addressed? Will the authors consider recalibrating the baseline risk to different populations for example?

2. For the risk of bias assessment, I suggest using items from PROBAST too (excluding the analysis domain) which has recently been published and relates to prediction modelling studies (Wolff *et al.*, 2019[1]).

3. Have the authors considered how much data they would need to acquire to develop new prediction models for TBM e.g. any sample size calculations based on likely event rate and expected number of candidate predictors for consideration in the models, as a target to aim for?

4. In my experience, one of the biggest difficulties with IPD-MA like these is how different studies record different combinations of variables. Therefore, combining studies for model development can be very difficult and it may be necessary to prioritise certain variables (or combinations of variables) and use a subset of studies with those variables, hence my previous comment regarding sample size. Have the authors considered which variables are of particular interest and what they will do if these are not recorded in individual studies? How will IPD be selected for developing new models as it is unlikely to all be used?

5. How will missing data be handled? If imputing, will this be done within or across datasets?

6. It's not clear if one model will be developed or multiple models (using each of the different modelling approaches). If aiming for a single model, how will it be selected?

7. Bottom of page 5: I'm not sure what is meant by "Model development will initially be carried out using participants with either definite TBM or definitely not TBM. The model will then be applied to participants with possible TBM." Can the authors please clarify? Do they mean that possible TBM will be included in the definition of TBM?

8. I'm also not sure what is meant by "the training set will be calibrated to optimise the model coefficients for best predictive accuracy using AUC-ROC score" (Data synthesis, part 2)? By definition, the model will be calibrated to the development data and is therefore optimised to the data, which can lead to overfitting.

9. Will clustering by dataset be accounted for in the model development e.g. using a random

intercept?

10. Will calibration of the model be assessed? This is also likely to be heterogeneous in different populations and therefore may need tailoring to different populations. In contrast, the AUC depends on the case-mix and will be lower in more homogeneous populations which doesn't mean the model doesn't work well.

11. I don't see the point in splitting each dataset for development and validation, especially when some studies are likely to be small (min. sample size of 10 so even fewer events). The internal-external cross-validation is a better approach as it still retains the external validation element and will help evaluate the heterogeneity in performance across datasets.

12. Have the authors considered the potential for optimism in model development, particularly if they have few events and small sample size overall? Will they consider shrinking the coefficients (in a regression modelling approach) to correct for optimism?

13. Data synthesis, part 3: What threshold will be selected to calculate measures of diagnostic test accuracy – will this be based on a predicted probability and pre-specified to avoid bias in using 'optimal' thresholds? I would also suggest evaluating calibration and discrimination as part of the internal validation.

14. I would suggest reporting according to the TRIPOD guidelines for the multivariable modelling (Collins *et al.*, 2015[2]).

15. I would caution against simply developing smart phone apps and web-based calculators unless the model demonstrates good predictive ability. Ideally it should be externally validated first before considering it as a tool for use in practice.

References

1. Wolff RF, Moons KGM, Riley RD, Whiting PF, et al.: PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.*Ann Intern Med*. 2019; **170** (1): 51-58 PubMed Abstract | Publisher Full Text
2. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.*BMJ*. 2015; **350**: g7594 PubMed Abstract | Publisher Full Text

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics, prediction modelling and IPD-MA

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 21 Aug 2019

Tom Boyles, University of the Witwatersrand, Johannesburg, South Africa

We would like to thank Professor Snell for her very helpful comments.

Responses drafted by Anna Stadelman.

READ ME: **Comments from Professor Snell are in bold-type font**. My responses are below each comment.

This is a clear and well-written protocol for a systematic review, collection of IPD and IPD meta-analysis for a new diagnosis model for TBM. I think the aim of the study and details relating to the systematic review part are clear, however I have a few comments/questions for clarity and reproducibility, mostly regarding what happens once IPD has been collected.

1. In the introduction the authors mention that case-mix can affect the predictive performance of a model and that big datasets can be used to examine the heterogeneity and improve the predictive performance. However, I don't think they really address this issue in the methods or say how they will use the IPD to try [to] improve the performance. Heterogeneity in performance if the predictor effects are consistent suggests differences in case-mix that are not being captured by the predictors in the model. Unless additional variables that are thought to improve the model are included, how will this be addressed? Will the authors consider recalibrating the baseline risk to different populations for example?

The objective of assessing heterogeneity will be to ascertain case-mix variation among TBM cases and non-cases to inform model development. To begin, the contributing datasets will be reviewed for sample size, available predictors, and data completeness to inform the selection of a modelling approach. A descriptive analysis will be undertaken to understand similarities and differences between the contributing datasets. Participant characteristics and clinical features will be summarized for each contributing dataset and compared across datasets using chi-square, t-tests, or non-parametric methods as warranted. Then, we will formally evaluate case-mix variation and predictor heterogeneity via IPD meta-analysis using a logistic regression model with stratified intercepts for each study (99). Each TBM predictor will be rotated into the model individually to underscore the baseline predictive value of each in the different contributing datasets. We will also use this method to assess

predictor heterogeneity by HIV status, WHO region, and TB burden within each country. For both the informal (descriptive statistics) and formal (IPD meta-analysis) assessment of heterogeneity, predictor estimates and their uncertainty intervals will be used to determine relative significance as opposed to p-values. Uncertainty intervals for each predictor will indicate how reliable the predictor is in terms of its prediction value. Furthermore, looking at p-values will only assess statistical significance, which may not be clinically meaningful.

Subsequently, we will employ methods of model development that take into account the heterogeneity observed in the IPD meta-analysis. These methods include, but are not limited to, classification and regression trees, supervised and unsupervised machine learning, Latent component analysis, etc. There may be other sources of heterogeneity that become evident during model development which will also be included in the development of the clinical prediction rule.

2. For the risk of bias assessment, I suggest using items from PROBAST too (excluding the analysis domain) which has recently been published and relates to prediction modelling studies (Wolff *et al.*, 2019[1]).

Thanks for the recommendation.

3. Have the authors considered how much data they would need to acquire to develop new prediction models for TBM e.g. any sample size calculations based on likely event rate and expected number of candidate predictors for consideration in the models, as a target to aim for?

Sample size is difficult to calculate in the context of developing a prediction model. However, the size of the development dataset and number of predictors in the final model have an impact on the statistical power to detect a difference in TBM case vs. non-TBM case. The greater the number of individual participants the more information we have to inform the development of the model, specifically the parameterization of the predictors and variability explained. More data (i.e. individual participants) better optimizes the individual predictors and has a better chance of capturing the variability in TBM case presentations. Ultimately, we will make every effort to acquire as many datasets as possible and limit the number of predictors to the ones that explain the most variability in TBM diagnosis.

4. In my experience, one of the biggest difficulties with IPD-MA like these is how different studies record different combinations of variables. Therefore, combining studies for model development can be very difficult and it may be necessary to prioritise certain variables (or combinations of variables) and use a subset of studies with those variables, hence my previous comment regarding sample size. Have the authors considered which variables are of particular interest and what they will do if these are not recorded in individual studies? How will IPD be selected for developing new models as it is unlikely to all be used?

We have included in the table which variables are of interest in the development of the model(s) (marked with an *). We consider these variables to be the most important for model development and inclusion of data into model development will be contingent on the representation of these variables in the individual contributing datasets. We will conduct a sensitivity analysis with all the individual contributing datasets, regardless of variable inclusion, so as to assess any bias introduced into the model by excluding certain datasets.

5. How will missing data be handled? If imputing, will this be done within or across datasets?

We will not impute any missing data. We will request all the diagnostic data available from investigators and any missingness on an individual level may ultimately end up excluding that particular individual from model development.

6. It's not clear if one model will be developed or multiple models (using each of the different modelling approaches). If aiming for a single model, how will it be selected?

We will create multiple models with the development dataset and compare the fit across models via bootstrap, k-fold cross validation, and internal-external cross-validation.

7. Bottom of page 5: I'm not sure what is meant by "Model development will initially be carried out using participants with either definite TBM or definitely not TBM. The model will then be applied to participants with possible TBM." Can the authors please clarify? Do they mean that possible TBM will be included in the definition of TBM?

The model(s) will be developed with confirmed TBM and non-TBM cases, and we will test the model(s) on probable and possible TBM cases as part of the sensitivity analysis.

8. I'm also not sure what is meant by "the training set will be calibrated to optimise the model coefficients for best predictive accuracy using AUC-ROC score" (Data synthesis, part 2)? By definition, the model will be calibrated to the development data and is therefore optimised to the data, which can lead to overfitting.

Sorry for the confusion. Suggested revision in Version 2.0 of the protocol.

9. Will clustering by dataset be accounted for in the model development e.g. using a random intercept?

Yes, this will be the aim of the IPD meta-analysis. Each predictor of interest will be rotated into a model predicting TBM that has a random intercept for each contributing dataset. The aim of this will be to ascertain heterogeneity in predictor strength, which will be accounted for in the final model(s). However, the overall aim is to develop a model(s) that accounts for

heterogeneity by region, HIV status, and other known causes of heterogeneity in TBM cases. Therefore, we are hoping that including these predictors in the final model(s) will account for most of the variation in predictor strength that may be introduced by individual contributing datasets.

10. Will calibration of the model be assessed? This is also likely to be heterogeneous in different populations and therefore may need tailoring to different populations. In contrast, the AUC depends on the case-mix and will be lower in more homogeneous populations which doesn't mean the model doesn't work well.

Yes, model(s) calibration will be assessed and you bring up important points about the metrics for calibration and discrimination.

11. I don't see the point in splitting each dataset for development and validation, especially when some studies are likely to be small (min. sample size of 10 so even fewer events). The internal-external cross-validation is a better approach as it still retains the external validation element and will help evaluate the heterogeneity in performance across datasets.

Agreed. We will revise our internal validation approach to include bootstrap, cross-validation (k-fold), and internal-external validation. Bootstrap validation tells us more about the validity of predictor variable selection in algorithm development, which is useful for assessing how well our predictors assess TBM diagnosis within different samples. Simulations have demonstrated that bootstrap is the best approach to internal validation as it appropriately reflects all sources of model uncertainty, especially in predictor variable selection (113). We will then utilize a k-fold cross-validation approach to assess the validation of the model approach and accuracy of model fit. The resulting c-statistic will convey overall model optimism and accuracy of model fit. The predictive model(s) will be further validated using 'internal-external cross-validation', which is a multiple validation approach that accounts for multiple studies by rotating which are used toward model development and validation (99).

12. Have the authors considered the potential for optimism in model development, particularly if they have few events and small sample size overall? Will they consider shrinking the coefficients (in a regression modelling approach) to correct for optimism?

High optimism, indicative of overfitting, can be corrected via shrinkage and we will consider this approach if overfitting is evident. However, we do not anticipate that we will encounter overfitting due to the data reduction step described in Version 2.0 of the protocol.

13. Data synthesis, part 3: What threshold will be selected to calculate measures of diagnostic test accuracy – will this be based on a predicted probability and pre-

specified to avoid bias in using 'optimal' thresholds? I would also suggest evaluating calibration and discrimination as part of the internal validation.

We will assess optimism, calibration, and discrimination as part of the internal validation approach and have discussed this process further in Version 2.0 of the protocol. As for determining a pre-specified predictive threshold for defining TBM versus not, there is little information in the literature to inform an appropriate cutoff for TBM. Prior prediction models have used ROC curve to determine an optimal cutoff. Furthermore, this is the first study to include data from different populations world-wide. As such it is difficult to pre-specify the optimal predictive cutoff.

14. I would suggest reporting according to the TRIPOD guidelines for the multivariable modelling (Collins *et al.*, 2015[2]).

Great! Thanks for the recommendation.

15. I would caution against simply developing smart phone apps and web-based calculators unless the model demonstrates good predictive ability. Ideally it should be externally validated first before considering it as a tool for use in practice.

Absolutely agree. Developing an application and/or website calculator is our end goal, but the step to getting there includes further external validation.

References

1. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group†: PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.*Ann Intern Med*. 2019; **170** (1): 51-58 PubMed Abstract | Publisher Full Text
2. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.*BMJ*. 2015; **350**: g7594 PubMed Abstract |Publisher Full Text

Competing Interests: None

Reviewer Report 05 February 2019

https://doi.org/10.21956/wellcomeopenres.16426.r34749

Ravindra Kumar Garg
Department of Neurology, King George's Medical University, Lucknow, Uttar Pradesh, India

I read with interest the protocol that aims to identify and validate a set of clinical predictors that accurately identify patients with definite tuberculous meningitis and absence of tuberculous meningitis. Conventionally, microscopy for acid-fast bacilli, commercial nucleic acid amplification test for *Mycobacterium tuberculosis* or mycobacterial culture of cerebrospinal fluid, are the tests that are used to bacteriologically confirm the diagnosis of tuberculous meningitis.

In developing countries and countries with a very high tuberculosis burden, tuberculous meningitis is encountered very frequently. Tuberculous meningitis is the commonest CNS infection seen in Neurology and Medicine indoors. Facing resource constrains, we always have to rely on clinical, imaging and cerebrospinal fluid parameters. Despite constrains, we are able to make reliable diagnosis of tuberculous meningitis most of the time. Our classical teaching points, to diagnose tuberculous meningitis, are often accurate. With a clinical diagnosis of meningitis along with characteristic cerebrospinal fluid findings help in making reasonable and prompt diagnosis enabling to start antituberculosis treatment with confidence. Raised cerebrospinal fluid lymphocyte count and markedly raised protein are characteristically seen in tuberculous meningitis.

Certain clinical signs are very specific to tuberculous meningitis. For example, sixth nerve involvement and vision loss in points towards a basal meningeal involvement and tuberculous meningitis. Other cranial nerve involvements are very infrequent. In patients with multiple cranial nerve palsies, fungal infection and a malignancy are more likely possibilities. As per observation, headache and fever are often not dominant features, and they are never presenting features. Similarly, neck rigidity may not be present in many patients. In cryptococcal meningitis, severe and dominant headache may be a presenting feature. Presence of extrapyramidal movements is a rare manifestation of tuberculous meningitis in adults. Extrapyramidal movements are more frequent in children.

Computed tomographic findings, if present, are quite characteristic of tuberculous meningitis. Basal exudates along with hydrocephalus with or without tuberculoma and periventricular infarcts indicates tuberculous meningitis and differential diagnosis option are then limited. A search for spinal cord involvement, we believe, if present, add to the diagnostic accuracy. A combination of optochiasmatic arachnoiditis and spinal lumbo-sacral arachnoiditis, in my opinion, is probably as accurate as bacteriological confirmation. Demonstration of paradoxical reaction, if present, also helps us in substantiating the reliable diagnosis of tuberculous meningitis.

Tuberculous meningitis, frequently, is a manifestation of more disseminated tuberculosis. Search for other sites of involvement often help us establishing clinical diagnosis. For example, ordinary X-ray chest shows additional pulmonary involvement. Many cases surprisingly show asymptomatic military tuberculosis. Lymph adenopathy and spinal vertebral tuberculosis are also seen in many cases.

Diagnostic caution is exercised in elderly patients and HIV infected patients. In these two groups, there are high chances of alternative diagnosis. We routinely perform tests with India ink preparation and detection of malignant cells. Still, distinctive features of tuberculous meningitis help in diagnosis of tuberculous meningitis in these two populations as well. Aspergillosis has a

more aggressive course and large vessel involvement is more common. In tuberculous meningitis infarcts are usually small and periventricular.

Another issue that need to be addressed is diagnosis of drug-resistant tuberculous meningitis. XpertMTB/RIF test, which is now readily available, start discovering drug-resistant tuberculous meningitis in increasing number. This is not surprising because India harbors the major portion of global drug-resistant tuberculosis problems. This issue also needs to be given due emphasis.

I greatly appreciate the investigators efforts to evolve a predictive logistic model to accurately diagnose definite tuberculous meningitis. There are certain points that I highlighted that need to be re-looked and can be incorporated in this protocol.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Yes

Are sufficient details of the methods provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: CNS tuberculosis and other CNS infections

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.