

## **The effect of *M. tuberculosis* lineage on clinical phenotype**

Duc Hong Du<sup>1</sup>, Ronald B Geskus<sup>1,2</sup>, Yanlin Zhao<sup>3</sup>, Luigi Ruffo Codecasa<sup>4</sup>, Daniela Maria Cirillo<sup>5</sup>,  
Reinout van Crevel<sup>2,6</sup>, Dyshelly Nurkartika Pascapurnama<sup>7</sup>, Lidya Chaidir<sup>8</sup>, Stefan Niemann<sup>9,10</sup>,  
Roland Diel<sup>11,12</sup>, Shaheed Vally Omar<sup>13</sup>, Louis Grandjean<sup>14</sup>, Sakib Rokadiya<sup>14</sup>, Arturo Torres Ortiz<sup>14</sup>,  
Nguyễn Hữu Lâm<sup>15</sup>, Đặng Thị Minh Hà<sup>15</sup>, E. Grace Smith<sup>16</sup>, Esther Robinson<sup>16</sup>, Martin Dedicoat<sup>16,17</sup>,  
Le Thanh Hoang Nhat <sup>1</sup>, Guy E Thwaites<sup>1,2</sup>, Le Hong Van<sup>1</sup>, Nguyen Thuy Thuong Thuong<sup>1,2\*</sup>, Timothy  
M Walker<sup>1,2\*</sup>

\*Contributed equally

<sup>1</sup> Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam

<sup>2</sup> Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford,  
UK

<sup>3</sup> CDC China, Beijing, China

<sup>4</sup> Regional TB Reference Centre/ Istituto Villa Marelli- ASST Grande Ospedale Metropolitano Niguarda, Milano,  
Italy

<sup>5</sup> IRCCS San Raffaele Scientific Institute, Milano, Italy

<sup>6</sup> Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical  
Center, Nijmegen, the Netherlands

<sup>7</sup> Research Center for Care and Control of Infectious Diseases, Padjadjaran University, Bandung, West Java,  
Indonesia.

<sup>8</sup> Division of Microbiology, Department of Biomedical Science, Faculty of Medicine, Padjadjaran University,  
Bandung, West Java, Indonesia

<sup>9</sup> Research Center Borstel, Germany

<sup>10</sup> German Center for Infection Research, partner site Hamburg-Lübeck-Borstel-Riems, Germany

<sup>11</sup> University Hospital Schleswig-Holstein, Campus Kiel, Germany

<sup>12</sup> Lung Clinic Grosshansdorf, Germany, Airway Disease Center North (ARCN), German Center for  
Lung Research (DZL)

<sup>13</sup> NICD, Johannesburg, South Africa

<sup>14</sup> University College London Hospital, London, UK

<sup>15</sup> Pham Ngoc Thach Hospital, Ho Chi Minh City, Vietnam

<sup>16</sup> TB Unit and National Mycobacterial Reference Service, UK Health Security Agency, UK

<sup>17</sup> University Hospitals Birmingham NHS Foundation Trust, UK

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

33 **Abstract**

34 Eight lineages of *Mycobacterium tuberculosis sensu stricto* are described. Single-country or  
35 small observational data suggest differences in clinical phenotype between lineages. We  
36 present strain lineage and clinical phenotype data from 12,246 patients from 3 low-incidence  
37 and 5 high-incidence countries. We used multivariable logistic regression to explore the  
38 effect of lineage on site of disease and on cavities on chest radiography, given pulmonary  
39 TB; multivariable multinomial logistic regression to investigate types of extra-pulmonary TB,  
40 given lineage; and accelerated failure time and Cox proportional-hazards models to explore  
41 the effect of lineage on time to smear and culture-conversion. Mediation analyses quantified  
42 the direct effects of lineage on outcomes. Pulmonary disease was more likely among  
43 patients with lineage(L) 2, L3 or L4, than L1 (adjusted odds ratio (aOR) 1.79, (95%  
44 confidence interval 1.49-2.15),  $p < 0.001$ ; aOR=1.40(1.09-1.79),  $p = 0.007$ ; aOR=2.04(1.65-  
45 2.53),  $p < 0.001$ , respectively). Among patients with pulmonary TB, those with L1 had greater  
46 risk of cavities on chest radiography versus those with L2 (aOR=0.69(0.57-0.83),  $p < 0.001$ )  
47 and L4 strains (aOR=0.73(0.59-0.90),  $p = 0.002$ ). L1 strains were more likely to cause  
48 osteomyelitis among patients with extra-pulmonary TB, versus L2-4 ( $p = 0.033$ ,  $p = 0.008$  and  
49  $p = 0.049$  respectively). Patients with L1 strains showed shorter time-to-sputum smear  
50 conversion than for L2. Causal mediation analysis showed the effect of lineage in each case  
51 was largely direct. The pattern of clinical phenotypes seen with L1 strains differed from  
52 modern lineages (L2-4). This has implications for clinical management and could influence  
53 clinical trial selection strategies.

54

55

56

57

58

59

60

## 61 **Introduction**

62 *Mycobacterium tuberculosis* kills more people each year than any other pathogen  
63 besides SARS-CoV-2 (1). Pulmonary disease is its most common form, but tuberculosis can  
64 disseminate throughout the host and manifest anywhere. Nine separate human adapted  
65 *Mycobacterium tuberculosis* lineages are described(2,3). Each lineage has emerged in its own  
66 geographical niche with some so-called specialist lineages having adapted to the host  
67 population with which they co-evolved, and other more generalist lineages having successfully  
68 spread throughout the world (4–6).

69 For the purpose of clinical trials, most diagnostics, and treatment, TB is still considered  
70 one disease rather than a family of different but related entities. However, data suggesting  
71 genuine differences exist are accumulating. The minimum inhibitory concentrations for  
72 pyrazinamide and pretomanid were recently found to be higher for lineage 1 than for other  
73 tested lineages (7,8). Epidemiological studies into the relationship between lineage and  
74 clinical phenotype have in the past indicated that lineage 1 in particular is more strongly  
75 associated with extra-pulmonary disease than lineages 2-4 (9). Of late an association and  
76 proposed mechanism have even been proposed for how lineage 1 is more likely to result in  
77 TB osteomyelitis than other lineages (10). However, observational studies are fraught with  
78 potential confounders, such as immigration patterns, and only two small studies have so far  
79 simultaneously looked at data from both endemic countries and at countries where the majority  
80 of patients can be linked to another country of origin (11,12).

81 After decades of limited progress, there are now a wealth of new drugs and even  
82 vaccines in the pipeline. Were differences between lineages thought to be major, this would  
83 need to be considered when designing clinical studies assessing new interventions. Here we  
84 present data from eight countries across four continents to explore the effect of *M. tuberculosis*  
85 lineage on, and risk factors for, clinical manifestations from the focus of disease to radiological  
86 markers of severity to treatment response.

87

## 88 **Methods**

## 89 Sample selection

90 We analysed existing datasets from three low-incidence countries where most patients  
91 with TB can be linked to another country of origin, and 5 high-incidence countries where  
92 immigration patterns are less relevant to local TB patterns. In each case data had originally  
93 been generated for other studies, independent of this one. Data from Germany and the UK  
94 are from metropolitan population studies, and include isolates from all patients for whom a  
95 culture was available and a lineage determinable over defined time periods. In Germany in  
96 particular this resulted in enrichment for pulmonary TB (90% vs. 71% background rate in the  
97 country(13)). Data from Italy were representative of all patients with pulmonary TB attending  
98 a clinic in Milan. Data from China were from a national pulmonary TB prevalence study, and  
99 those from South Africa were from a surveillance study of bedaquiline resistance among  
100 patients with rifampicin resistant pulmonary TB. Data from Peru are from a locally  
101 representative observational study on pulmonary TB. Data from Vietnam and Indonesia  
102 originated both from observational studies on pulmonary TB, and from clinical trials on TB  
103 meningitis. See Table 1 for details.

104

## 105 Variables

106 All isolates had been assigned a lineage based on their whole-genome sequence  
107 before being shared for the purpose of this analysis. No further WGS analysis was conducted  
108 here. Outcome variables included locus of disease, cavities on chest radiographs (CXR),  
109 Timika scores (the higher the score, the more severe the disease on CXR),(14) and time to  
110 smear and culture conversion. These were all obtained from local clinicians, radiologists and  
111 laboratories. We included other variables plausibly effecting outcome, including sex, diabetes,  
112 HIV infection, age, isoniazid or rifampicin resistance, country of origin of the data and whether  
113 the patient was born in that same country or not. Supplementary table S1 shows which  
114 datasets contained which variables.

115

## 116 Statistical analysis

117 In order to understand which variables should be included as potential confounders,  
118 we developed Directed Acyclic Graphs (DAGs) identifying the hypothesized causal  
119 relationships between individual variables (Supplementary figures 1a-c). Logistic regression  
120 was used to relate the exposure (lineage) to (A) pulmonary vs. extra-pulmonary TB, and (B)  
121 the presence of pulmonary cavities in patients with pulmonary TB. In each case age, country  
122 of data origin, birth in that country, diabetes and HIV infection were included in the model  
123 where these variables were available. As not all datasets included all the same variables, we  
124 first analysed each country separately to assess risk factors for the outcome, before  
125 performing a causal analysis after pooling data from a subset of countries adjusting for  
126 covariates. Confidence intervals and p-values were computed based on the likelihood ratio  
127 test and result reported with Firth's correction where numbers were small. Age was included  
128 in the models using restricted cubic splines with three knots (the knots were chosen at the  
129 10%, 50% and 90% percentile of the values of the variable) to allow for potential non-linear  
130 relationships where applicable. Linear regression was used to examine the effect of lineage  
131 on Timika CXR score. As the score had a skewed distribution, we used the Box-Cox procedure  
132 to find a suitable transformation (i.e. a square root transformation) that makes the outcome  
133 variable more symmetrically distributed. Multivariable multinomial logistic regression was used  
134 to analyse the association between lineage and different forms of extra-pulmonary TB with a  
135 Wald test to obtain confidence intervals and p-values.

136 We investigated the effect of lineage on time to smear and culture conversion using  
137 an accelerated failure time model under the assumption of a log-logistic baseline distribution  
138 based on the Akaike information criterion (AIC) of different parametric fitted models including  
139 log-logistic (lowest AIC), exponential, Weibull, gamma and log-normal baseline distribution.  
140 The nature of the time-to-event data varied by study so these had to be standardised to  
141 achieve a lower and upper bound for the window of time in which smear or culture conversion  
142 is likely to have occurred. Where only one date – that of a first negative sample – was  
143 available, this was set as the upper bound of the interval and the lower bound set according  
144 to the expected monthly sampling date prior to that (day 0, 30, 60, 90 etc.). Where results from

145 multiple sampling dates were available for a patient, the upper bound was defined as the first  
146 of the first two consecutive negatives, and the last preceding positive set as the lower bound.  
147 Where the first negative sample was also the last sample documented, this was treated as the  
148 conversion date because we assume that there were more samples but they were all negative  
149 and therefore not reported. See appendix for further details. We analysed the same data using  
150 a Cox proportional-hazards model allowing for interval censored data to explore whether the  
151 findings were consistent.

152 We conducted causal mediation analyses to investigate the pathways of the effects of  
153 lineage on outcome. Intermediary and confounding variables were pre-determined from the  
154 DAGs (supplementary figures 1a-c). Effects of lineage on outcomes were divided into natural  
155 direct and indirect effects via the regression-based approach with closed-form parameter  
156 function estimation or direct counterfactual imputation estimation or via the g-formula  
157 approach where applicable (15,16). Exposure-mediator interaction was considered. We also  
158 assessed the contribution of the effect of lineage on outcome as operating through mediators,  
159 using the “proportion mediated” where applicable, i.e. when direct and indirect effects are in  
160 the same direction (17). The proportion mediated is defined as the ratio of the natural indirect  
161 effect (NIE) to the total effect (TE) (on the logit scale), that is,

$$PM = \frac{NIE}{TE}$$

163

164 **Supplementary Figure 1:**

165 Directed Acyclic Graph (DAG) on the causal assumptions underlying the effect of  
166 Lineage on (A) Pulmonary tuberculosis (TB); (B) the presence of Cavity; and (C)  
167 Time to culture/smear conversion. Arrows indicate the direction of the effect.  
168 Exposure, Mediator, Outcome and Confounders listed below each graph.

169

170

171 With binary outcomes, the proportion mediated can be calculated from the natural  
172 direct effects odds ratio  $OR^{NDE}$  and the natural indirect effect odds ratio  $OR^{NIE}$  when the  
173 outcome is rare and a logistic model was used (18):

$$174 \frac{OR^{NDE} (OR^{NIE} - 1)}{(OR^{NDE} \times OR^{NIE} - 1)}$$

175 When the outcome is not rare, the relative risk or rate ratio (RR) from a log-linear model  
176 was reported instead of an OR to prevent bias if implemented via the regression-based  
177 approach with closed-form parameter function estimation. Using the direct counterfactual  
178 imputation estimation will ignore this problem (17).

179 We reported estimated effects (ORs and RRs) for both natural direct and indirect  
180 effects with 95% CIs and p-values obtained by two methods: delta and bootstrapping (17). We  
181 reported estimated ORs using the direct counterfactual imputation estimation and 95%CIs and  
182 p-values obtained by bootstrapping method as the main results.

183 Analyses were performed in R version 4.2.0, and the packages ‘ggplot2’, ‘icenReg’,  
184 and ‘CMAverse’(19–22).

185

#### 186 Ethics, funding and data statements

187 IRB approvals were obtained from the University of Lübeck (Germany); Universidad  
188 Peruana Cayetano Heredia via the Peruvian Ministry of Health (ref. 100252) (Peru); Pham  
189 Ngoc Thach Hospital Ho Chi Minh City and the University of Oxford (OxTREC) (Vietnam);  
190 China CDC (China); University of Witwatersrand Human Research Ethics Committee (ref.  
191 M160667) (South Africa); Hasan Sadikin Hospital/Faculty of Medicine of Universitas  
192 Padjadjaran (clinical trial ref. NCT02169882) (Indonesia); Observational/Interventions  
193 Research Ethics Committee, London School of Hygiene and Tropical Medicine (ref: 6449)  
194 (Indonesia); institutional review boards in Indonesia in the context of the TANDEM study  
195 (Indonesia); Ospedale San Raffaele, Milan (ref: OSR 82/DG 26/2/10, amended 11/12/14)  
196 (Italy). Data from the UK data were obtained entirely from the published literature.(23) Those

197 data had been collected under public health law with no need for further IRB approval or  
198 individual patient consent.

199 All data are available in Supplementary Table S8. R code is available here:  
200 <https://github.com/duhongduc/lineage>. Raw WGS data are not provided as were not part of  
201 this analysis.

202

### 203 **Results**

204 Data on 12,547 patients were obtained from eight countries. Lineage 1 accounted for  
205 1,024 (8%), lineage 2 for 6,477 (52%), lineage 3 for 796 (6%) and lineage 4 for 3,955 (32%)  
206 observations. Data on lineage were missing for 254 patients, 40 labelled as a mixture of  
207 lineages were excluded, and 7 represented other lineages that were too rare to justify inclusion  
208 in this analysis. Tables 1 and supplementary table S1 show the numbers of patients and  
209 isolates from each country, and how these were sampled. The largest collections came from  
210 China, Vietnam, and the UK, with each of the other countries contributing data from fewer than  
211 1,000 patients.

212



**Table 1: Descriptions of datasets**

<b><u>Country</u></b>	<b><u>Data set description</u></b>	<b><u>Number of patients / isolates</u></b>	<b><u>Time interval</u></b>	<b><u>Bias</u></b>
Germany	Population study of TB in Hamburg	673	2008-2017	Enriched for pulmonary TB due to ease of sampling
UK	Population study of TB in Birmingham	1653	2009-2019	
Italy	All pulmonary TB isolates from patients in Milan	90	2018-2019	Pulmonary TB only
South Africa	Bedaquiline resistance prevalence study among patients with MDR-TB	175	2015-2019	Enriched for rifampicin resistant; pulmonary TB only
Vietnam 1	Observational studies of all pulmonary TB	1536	2008-2011	Pulmonary TB only
Vietnam 2	Observational study of MDR pulmonary TB	257	2017-2020	Pulmonary MDR-TB only
Vietnam 3	Observational study of rifampicin susceptible pulmonary TB	537	2017-2020	Pulmonary Rifampicin-susceptible TB only
Vietnam 4	Clinical trial of TB meningitis	370	2011-2014	TB meningitis only
Vietnam 5	Clinical trial of TB meningitis in HIV infected individuals	111	2017-2020	TB meningitis only; HIV+
Vietnam 6	Clinical trial of TB meningitis in HIV uninfected individuals	105	2018-2020	TB meningitis only; HIV-
Indonesia 1	Clinical trial of TB meningitis	106	2006-2016	TB meningitis only
Indonesia 2	Observational study of pulmonary TB, enriched for MDR-TB	765	2006-2016	Pulmonary TB only; enriched for MDR-TB
China	Population prevalence study of pulmonary TB	5445	2013-2017	Pulmonary TB only
Peru	Observational study of pulmonary TB	429	2018-2019	Pulmonary TB only

## 213 Pulmonary vs. extra-pulmonary disease

214 Datasets from Vietnam, Indonesia, Germany and the UK contained data on patients  
215 with pulmonary TB and on patients with extra-pulmonary TB so could be used to assess  
216 whether lineage influences the spread of TB beyond the lung. For this purpose, patients known  
217 to have both pulmonary and extra-pulmonary TB were classified as 'extra-pulmonary TB'.

218 Data on patient sex were available for all countries, and on patient age and HIV  
219 infection from Germany, Indonesia and Vietnam. However, as data on HIV from Vietnam were  
220 from two TB meningitis clinical trials on HIV infected and uninfected individuals respectively,  
221 this variable was excluded from Vietnam to avoid bias. Data on diabetes were available from  
222 Germany and Indonesia only.

223 We first analysed each country individually to explore risk factors for each outcome. In  
224 Germany and Vietnam lineages 2 and 4 were more likely than lineage 1 to associate with  
225 pulmonary TB after controlling for immigration, and in the case of Germany, HIV infection and  
226 diabetes as well (aOR=4.88 (95% CI 1.41-19.8), p=0.016 and aOR=3.16 (1.29-7.23), p=0.008  
227 respectively for Germany; aOR=1.7 (1.38-2.09), p<0.001 and aOR=2.43 (1.66-3.63), p<0.001  
228 respectively for Vietnam). No difference was seen by lineage for Indonesia where the trend  
229 was in the opposite direction. Increasing age was a risk factor for pulmonary TB in Vietnam  
230 and Indonesia (supplementary Table S2a).

231 We next performed a causal analysis of each country controlling for immigration, and  
232 age where available. In Germany, Vietnam and Indonesia we saw the same patterns as in the  
233 analysis of risk factors. A similar pattern was also seen in the UK where lineages 2, 3 and 4  
234 were all more likely to cause pulmonary TB as compared to lineage 1 (aOR=2.21 (1.31-3.81),  
235 p=0.003; aOR=1.51 (1.11-2.07), p=0.009; aOR=2.19 (1.59-3.04), p<0.001, respectively). To  
236 combine data from the 4 countries and still control for age, we needed to impute age for the  
237 UK patients. We based this on the mean age in Germany, which a sensitivity analysis showed  
238 did not bias the results. In the consequent combined analysis, lineages 2, 3 and 4 were again  
239 all more likely to cause pulmonary TB than lineage 1 (aOR=1.79 (1.49-2.15), p<0.001;

240 aOR=1.40 (1.09-1.79), p=0.007; aOR=2.04 (1.65-2.53), p<0.001 respectively) (Figure 1a,  
241 supplementary table S2b).

242

243 **Figure 1:**

244 **A:** Multivariable logistic regression model on the association between lineage and  
245 pulmonary versus extra-pulmonary tuberculosis (TB) controlling for age and  
246 immigration. Estimated odd ratios (ORs) and bars representing 95% confidence  
247 intervals (CIs) are shown on the x-axis for lineage 2, 3 and 4, compared to lineage 1  
248 as reference, for each country as well as for all these countries combined. P-values  
249 denote evidence of the associations of lineage and pulmonary TB.

250 **B:** Causal mediation analysis (CMA) on the effect of lineage on pulmonary TB,  
251 mediated by drug resistance. Estimated odds ratio (ORs) and bars representing 95%  
252 confidence intervals (CIs) are shown on the y-axis for each decomposition effect  
253 including NDE: natural direct effect odds ratio; NIE: natural indirect effect odds ratio;  
254 and TE: total effect odds ratio for lineage 2, lineage 3, and lineage 4, compared to  
255 lineage 1 as reference. All multivariable models adjusted for country, immigration,  
256 and age are shown. P-values denote evidence of natural indirect effect of lineage on  
257 pulmonary TB mediated through drug resistance. The red horizontal lines indicate the  
258 thresholds of the results (ORs) of interest.

259

260 A causal mediation analysis adjusting for country (Germany, Indonesia and Vietnam),  
261 immigration, and age indicated that the effect of lineage on pulmonary TB is largely  
262 independent (direct effect) but that some of the effect is also mediated through drug  
263 resistance. Lineage 2 had a higher probability of drug resistance than lineage 1, and drug  
264 resistance in turn led to a higher probability of pulmonary TB (Figure 1b, supplementary table  
265 S2c).

266

267 Pulmonary cavity vs. no cavity

268           Given the association between lineages 2, 3 and 4 and pulmonary disease, we  
269 explored whether these lineages are also more likely than lineage 1 to lead to lung cavities as  
270 seen on CXR. Cavities are a marker of severity of disease, and are considered a risk factor  
271 for onward transmission (24). Data on the presence of cavities were available from Germany,  
272 Italy, Vietnam, Indonesia, China and Peru, and any missing data was understood to be missing  
273 completely at random.

274           Logistic regression was used to assess each country's data for risk factors for cavity  
275 formation. Diabetes was a risk factor for cavities on CXR in China and Peru (aOR=1.27 (1.02-  
276 1.56), p=0.028 and aOR=4.04 (1.37-17.3), p=0.026 respectively, supplementary table S3b).  
277 As age and immigration were identified as potential confounders in the DAG, further analyses  
278 were restricted to include only these co-variables.

279           In Vietnam, where 490 lineage 1 isolates were present, both lineages 2 and 4 were  
280 found to be less likely to associate with cavity formation than lineage 1 (aOR=0.66 (0.53-0.81),  
281 p<0.001; aOR=0.47 (0.33-0.67), p<0.001, respectively). No difference was seen between  
282 lineages for data from Germany, Italy, Indonesia, China or Peru (supplementary table S3a).  
283 However, the maximum number of lineage 1 isolates in any one of these countries was just  
284 29. In Peru there were zero lineage 1 isolates, leaving L2 as the reference. For the causal  
285 analysis we pooled the data from all five countries. Controlling for age, immigration and  
286 country we found that lineage 1 was more likely to cause cavities than lineage 2 and 4. The  
287 odds ratio for lineage 3 was similar to the other two modern lineages but the confidence  
288 intervals allowed for the possibility that it might behave similar to lineage 1 (Figure 2a,  
289 supplementary table S3a).

290

291           **Figure 2:**

292           **A:** Multivariable logistic regression model on the association between lineage and the  
293 presence of cavity versus no cavity among patients with pulmonary TB, controlling for  
294 age and immigration. Estimated odd ratios (ORs) and bars representing 95%  
295 confidence intervals (CIs) are shown on the x-axis for lineage 2, 3 and 4, compared

296 to lineage 1 as reference, for each country as well as for all these countries  
297 combined. P-values denote evidence of the associations of lineage and cavity.  
298 **B:** Causal mediation analysis (CMA) on the effect of lineage on cavity, mediated by  
299 drug resistance. Estimated odds ratio (ORs) and bars representing 95% confidence  
300 intervals (CIs) are shown on the y-axis for each of the decomposition effect including  
301 NDE: natural direct effect odds ratio; NIE: natural indirect effect odds ratio; and TE:  
302 total effect odds ratio of lineage 2, lineage 3, and lineage 4, compared to lineage 1 as  
303 reference. All multivariable models adjusted for country, immigration, and age are  
304 shown. P-values denote evidence of natural indirect effect of lineage on cavity  
305 mediated through drug resistance. The red horizontal lines indicate the thresholds of  
306 the results (ORs) of interest.

307

308 A causal mediation analysis was performed for Germany, Italy, Indonesia, China and  
309 Vietnam as there was data on drug resistance for these. This indicated a protective natural  
310 direct effect of lineage from cavity formation for lineages 2 and 4, compared to lineage 1. Here  
311 however the natural indirect effect of drug resistance increased the odds of cavity for lineage  
312 2 over lineage 1. Lineage 2 has higher odds of drug resistance versus lineage 1, and drug  
313 resistance in turn leads to higher odds of cavity formation (Figure 2b, supplementary table  
314 S3c).

315 Severity of diseases on CXR had been assessed for data from Peru, Italy, Indonesia  
316 and Vietnam using the Timika scoring system to which the presence of cavities on the CXR  
317 contributes (14). As the score had a skewed distribution, we used the Box-Cox procedure to  
318 find a suitable transformation (a square root transformation) to generate a more symmetrical  
319 distribution. Although lineage 1 was associated with a higher Timika score than lineage 4 in  
320 Vietnam, this effect was not observed elsewhere (supplementary table S4).

321

322 Extra-pulmonary disease

323 Having found that in these data lineage 1 was associated with extra-pulmonary TB  
324 relative to other lineages, we investigated whether there was a particular focus of extra-  
325 pulmonary diseases which lineage 1 favours. The population-based data from Germany and  
326 the UK were most suited to assess this. Data from the UK reported pulmonary TB; TB  
327 meningitis; TB osteomyelitis; or other types of TB. Data from Germany were more detailed on  
328 other forms of TB, including lymph node and pleural disease, although there were only 3  
329 patients with TB meningitis and 10 with osteomyelitis across all 4 lineages (supplementary  
330 table S1). Data from the two countries were therefore pooled, coding the sites of disease for  
331 Germany as for the UK. One case with two sites of extra-pulmonary TB was dropped from this  
332 analysis. Using multivariable multinomial logistic regression we took pulmonary TB as the  
333 reference and assessed the odds ratio of TB meningitis, osteomyelitis or other forms of extra-  
334 pulmonary TB, given lineage. TB osteomyelitis was significantly more likely than non-  
335 meningeal foci of extra-pulmonary TB (“other”) for lineage 1 as compared to lineages 2, 3 and  
336 4 ( $p=0.032$ ,  $p=0.01$ , and  $p=0.049$  respectively; supplementary tables S5a and S5b; Figure 3  
337 and supplementary Figure 2).

338

339 **Figure 3:**

340 Multinomial, multivariable regression for pulmonary tuberculosis (TB) vs. three types  
341 of extra-pulmonary TB (TB meningitis, TB osteomyelitis, and other forms of extra-  
342 pulmonary TB), by lineage, controlling for country and immigration. Estimated odd  
343 ratios (ORs) and bars representing 95% confidence intervals (CIs) are shown on the  
344 x-axis for the odds ratios of being lineage 2, 3 and 4, compared to lineage 1 as  
345 reference, comparing pulmonary TB to each of form of extra-pulmonary TB (“TBM” -  
346 TB meningitis, “Osteo” - TB osteomyelitis, and “Other” - other forms of extra-  
347 pulmonary TB) on the y-axis. P-values denote evidence of the association between  
348 lineages and different forms of extra-pulmonary TB compared to pulmonary TB.

349 **Supplementary Figure 2:**

350

351 Predicted probability for pulmonary tuberculosis (TB) vs. three types of extra-  
352 pulmonary TB (TB meningitis, TB osteomyelitis, and other forms of extra-pulmonary  
353 TB), by lineage “lin1234” (Lineage 1, 2, 3, and 4), country (Germany and UK) and  
354 immigration (“immi” (0 – born local and 1 – born overseas) from multinomial,  
355 multivariable regression.

356

357

#### 358 Time to smear and culture conversion

359 To explore a different manifestation of the clinical phenotype by lineage, we related  
360 lineage to the time to sputum smear and culture conversion. Data from Indonesia, Italy and  
361 Vietnam were available to model time to smear and culture conversion whilst correcting for  
362 country, immigration, and age as potential confounders.

363 In the accelerated failure time (AFT) model, time to both smear and culture conversion  
364 increased with lineage 2 compared to lineage 1 (Figure 4a; supplementary tables S6a-b). The  
365 Cox proportional-hazards model produced similar results, showing that the ‘hazards’ of having  
366 culture or smear conversion at any given time was lower for lineages 2 and 4 than for lineage  
367 1 (i.e. longer time to conversion), after adjusting for age, country and immigration  
368 (supplementary Figure 3; supplementary tables S7a-b).

369

#### 370 **Figure 4:**

371 **A:** Interval censored regression using accelerated failure time models on the  
372 association between lineage and time to culture (i) and smear (ii) conversion  
373 controlling for age, country and immigration status. Estimated time ratios and bars  
374 representing 95% confidence intervals (CIs) are shown on the x-axis. Data from  
375 Indonesia, Italy and South Africa all had interval censored data whereas the data  
376 from Vietnam were binary ( $\leq 60$  days or  $> 60$  days). The Vietnamese data were  
377 therefore converted to interval data (“0 to 60” if  $\leq 60$ ; and “61 to  $\infty$ ” if  $> 60$ ). P-values

378 denote evidence of the associations of lineage and time to culture or smear  
379 conversion.  
380 **B:** Causal mediation analysis (CMA) on the effect of lineage on time to culture  
381 conversion mediated by drug resistance and cavity. Estimated time ratios and bars  
382 representing 95% confidence intervals (CIs) are shown on the y-axis for each of the  
383 decomposition effect including NDE: natural direct effect odds ratio; NIE: natural  
384 indirect effect odds ratio; and TE: total effect odds ratio of lineage 2 and lineage 4,  
385 compared to lineage 1 as reference. All multivariable models adjusted for country,  
386 immigration, and age are shown. P-values denote evidence of natural indirect effect  
387 of lineage on time to culture conversion mediated through drug resistance and cavity.  
388 The red horizontal lines indicate the thread holds of the results (ORs) of interest.

389 **Supplementary Figure 3:**

390 Interval censored regression using proportional hazards models on the association  
391 between lineage and time to culture (A) and smear (B) conversion controlling for age,  
392 country and immigration. Estimated hazards ratios and bars representing 95%  
393 confidence intervals (CIs) are shown on the x-axis. Data from Indonesia, Italy and  
394 South Africa all had interval censored data whereas the data from Vietnam were  
395 binary ( $\leq 60$  days or  $> 60$  days). The Vietnamese data were therefore converted to  
396 interval data ("0 to 60" if  $\leq 60$ ; and "61 to  $\infty$ " if  $> 60$ ). P-values denote evidence of the  
397 associations of lineage and time to culture or smear conversion.

398  
399 Causal mediation analysis looking at cavity and drug resistance showed no evidence  
400 of a natural indirect effect mediated through drug resistance and cavity on time to culture  
401 conversion (Figure 4b, supplementary table S6c). We found moderate evidence that lineage  
402 4 shortened time to smear conversion, compared to lineage 1, as a natural indirect effect  
403 mediated through drug resistance and cavity, in the same direction as the natural direct effect  
404 (supplementary Figure 4, supplementary table S6d).

405



406 **Supplementary Figure 4:**

407 Causal mediation analysis (CMA) on the effect of lineage on time to smear  
408 conversion mediated by drug resistance and cavity. Estimated time ratios and bars  
409 representing 95% confidence intervals (CIs) are shown on the y-axis for each of the  
410 decomposition effect including NDE: natural direct effect odds ratio; NIE: natural  
411 indirect effect odds ratio; and TE: total effect odds ratio of lineage 2 and lineage 4,  
412 compared to lineage 1 as reference. All multivariable models adjusted for country,  
413 immigration, and age are shown. P-values denote evidence of natural indirect effect  
414 of lineage on time to smear conversion mediated through drug resistance and cavity.  
415 The red horizontal lines indicate the thread holds of the results (ORs) of interest.

416

417 **Discussion**

418 We explore causal relationship between *M. tuberculosis* lineage and TB clinical  
419 phenotypes using a large dataset derived from 3 low-incidence and 5 high-incidence countries.  
420 In these data lineage 1 is a risk factor for extra-pulmonary TB, with a suggestion that it favours  
421 TB osteomyelitis in particular. When lineage 1 does cause pulmonary TB, we find that it is a  
422 risk factor for pulmonary cavities.

423 Lineage 1 is ancestral to the other 3 lineages assessed in this study (25). It is most  
424 prevalent in south Asia and has been linked to both extra-pulmonary TB and to more severe  
425 disease phenotypes (9,26,27). It is plausible that the modern lineages (2, 3 and 4) are more  
426 adapted to manifesting as pulmonary TB, thereby favouring more transmission(10). Indeed,  
427 there is evidence from Vietnam and elsewhere that lineage 1 is being out-competed by the  
428 other three global lineages (28,29). Our finding that lineage 2 has a longer time to smear and  
429 culture conversion is consistent with this (11,30), although we find no difference between  
430 lineages 1 and 4. It may be that if lineage 1 does have a preponderance to cause extra-  
431 pulmonary TB, that it compensates by forming cavities when it does manifest as pulmonary  
432 TB, boosting opportunities for onward transmission when it can (24). We however know of few  
433 other data to support this hypothesis.

434           The results of the causal mediation analyses indicate that the effect of lineage on locus  
435 of disease and cavity formation is only minimally mediated through drug resistance. That some  
436 effect is mediated through cavity is an intuitive finding given that drug resistance could lead to  
437 worse disease where it is more difficult to treat. Although we might have expected to find some  
438 effect on time to smear and culture conversion mediated through drug resistance and cavity,  
439 this was only observed for time to smear, but not for time to culture conversion.

440           Results from observational studies on the association between *M. tuberculosis* lineage  
441 and clinical phenotype must be treated with caution. Other potential confounders, include host  
442 susceptibility, co-morbidities, and local socio-economic circumstances. Most previous studies  
443 that we are aware of have focussed only on data from single countries (9), and those that have  
444 focussed on more than one country have been small (11,12). We have gathered data from 8  
445 countries and over 12,000 patients. Our data represent both endemic settings and countries  
446 where the majority of TB is most likely acquired elsewhere. Although we could not control for  
447 all potential confounders, controlling for country of data origin may have captured at least  
448 some. Indeed, the recent identification of a potential mechanism for the association between  
449 lineage 1 and TB osteomyelitis provides support for our approach and findings.(10)

450           Limitations of our study are that our data were derived from a wide range of studies,  
451 each primarily designed for another purpose. Our sampling frames ranged from population  
452 prevalence studies to clinical trials, and from studies that included all clinical samples over  
453 defined periods of time and space to studies that selected only for one manifestation of  
454 disease. These intrinsic biases need to be recognised when constructing each of the  
455 presented models. However, none of the studies we pooled data from selected on the basis  
456 of lineage as such data could not be known until after the isolates had been analysed. As such  
457 each collection had the opportunity to inform on the distribution of lineages within it.

458           Other limitations include heterogeneity of measurement for individual variables. For  
459 example, cavities were reported from CXRs by attending physicians or by radiologists in the  
460 different countries without centralised, standardised reporting across the study. Time to smear  
461 and culture conversion was reported differently for different studies, with some data sets

462 requiring more assumptions to define time intervals than others. No data on patient treatment  
463 were available to control for whether they were on effective regimens. As all data were  
464 gathered from pre-existing studies, there was no opportunity to remedy this. Despite the  
465 inevitable noise in our data, the large size of the combined data set will have helped overcome  
466 a degree of random error by increasing power.

467         As our understanding of the phenotypic differences between *M. tuberculosis* lineages  
468 increases, we need to keep re-assessing whether to factor differences into study designs (11).  
469 Increased MICs for pyrazinamide and pretomanid for lineage 1 are clearly relevant to clinical  
470 trials of those drugs. Whether different clinical phenotypes should be considered as well  
471 remains an open question, but it could be of interest if certain lineages are more prone to  
472 cavity formation or different durations to smear or culture conversion as these variables could  
473 plausibly impact outcome. Multi-centre, multi-country clinical studies are usually designed to  
474 capture diversity across host populations, but it may be that incorporating the full genetic  
475 diversity of the pathogen in such studies is equally important.

### Funding:

This research was funded in whole, or in part, by the Wellcome Trust [214560/Z/18/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This work was supported by Wellcome Trust (grants 226007/Z/22/Z to LG) and the National Institute of Allergy and Infectious Diseases, National Institutes of Health (grant 1R01AI146338 to LG). Parts of this work have been funded by the Leibniz Science Campus Evolutionary Medicine of the LUNG (EvoLUNG), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2167 Precision Medicine in Inflammation as well as the Research Training Group 2501 TransEvo, and the German Center for Infection Research (DZIF). TMW is a Wellcome Trust Clinical Career Development Fellow (214560/Z/18/Z).

### Acknowledgements:

We would like to acknowledge the help of Dr. Stefania Torri of the Regional Tb Reference Laboratory, Grande Ospedale Metropolitano Niguarda, Milan Italy, and of Dr. Simone Villa, of the Centre for Multidisciplinary Research in Health Science University of Milan, Italy.

## References

1. World Health Organization. Global tuberculosis report 2021 [Internet]. Geneva: World Health Organization; 2021 [cited 2021 Oct 14]. Available from: <https://apps.who.int/iris/handle/10665/346387>
2. Ngabonziza JCS, Loiseau C, Marceau M, Jouet A, Menardo F, Tzfidia O, et al. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. *Nat Commun.* 2020 Dec;11(1):2917.
3. Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of Mycobacterium africanum reveals a new lineage and a complex evolutionary history. *Microbial Genomics.* 2021 February 1;7(2).
4. Gagneux S, DeRiemer K, Van T, Kato-Maeda M, de Jong BC, Narayanan S, et al. Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci USA.* 2006 Feb 21;103(8):2869–73.
5. Pasipanodya JG, Moonan PK, Vecino E, Miller TL, Fernandez M, Slocum P, et al. Allopatric tuberculosis host–pathogen relationships are associated with greater pulmonary impairment. *Infection, Genetics and Evolution.* 2013 Jun;16:433–40.
6. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, et al. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet.* 2016 Dec;48(12):1535–43.
7. Modlin SJ, Marbach T, Werngren J, Mansjö M, Hoffner SE, Valafar F. Atypical Genetic Basis of Pyrazinamide Resistance in Mono-resistant Mycobacterium tuberculosis. *Antimicrob Agents Chemother.* 2021 May 18;65(6):e01916-20.

8. Bateson A, Ortiz Canseco J, McHugh TD, Witney AA, Feuerriegel S, Merker M, et al. Ancient and recent differences in the intrinsic susceptibility of *Mycobacterium tuberculosis* complex to pretomanid. *Journal of Antimicrobial Chemotherapy*. 2022 May 29;77(6):1685–93.
9. Click ES, Moonan PK, Winston CA, Cowan LS, Oeltmann JE. Relationship Between *Mycobacterium tuberculosis* Phylogenetic Lineage and Clinical Site of Tuberculosis. *Clinical Infectious Diseases*. 2012 Jan 15;54(2):211–9.
10. Saelens JW, Sweeney MI, Viswanathan G, Xet-Mull AM, Jurcic Smith KL, Sisk DM, et al. An ancestral mycobacterial effector promotes dissemination of infection. *Cell*. 2022 Nov; S0092867422013617.
11. Nahid P, Bliven EE, Kim EY, Mac Kenzie WR, Stout JE, Diem L, et al. Influence of *M. tuberculosis* Lineage Variability within a Clinical Trial for Pulmonary Tuberculosis. Marais B, editor. *PLoS ONE*. 2010 May 20;5(5):e10753.
12. Negrete-Paz AM, Vázquez-Marrufo G, Vázquez-Garcidueñas MaS. Whole-genome comparative analysis at the lineage/sublineage level discloses relationships between *Mycobacterium tuberculosis* genotype and clinical phenotype. *PeerJ*. 2021 Sep 8;9:e12128.
13. Robert Koch Institute. Report on the Epidemiology of Tuberculosis in Germany - 2020 [Internet]. 2021 Dec [cited 2022 May 30]. Available from: [https://www.rki.de/EN/Content/infections/epidemiology/inf\\_dis\\_Germany/TB/summary\\_2020.html](https://www.rki.de/EN/Content/infections/epidemiology/inf_dis_Germany/TB/summary_2020.html)
14. Ralph AP, Ardian M, Wiguna A, Maguire GP, Becker NG, Drogumuller G, et al. A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. *Thorax*. 2010 Oct 1;65(10):863–9.

15. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*. 2013 Jun;18(2):137–50.
16. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. *American Journal of Epidemiology*. 1986 Mar;123(3):392–402.
17. Van der Weele TJ. Explanation in causal inference: developments in mediation and interaction. *Int J Epidemiol*. 2016 Nov 17;dyw277.
18. Van der Weele TJ, Vansteelandt S. Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *American Journal of Epidemiology*. 2010 Dec 15;172(12):1339–48.
19. Shi B, Choirat C, Coull BA, VanderWeele TJ, Valeri L. CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses. *Epidemiology*. 2021 Sep;32(5):e20–2.
20. R Core Development Team. A language and environment for statistical computing. R Foundation for Statistical Computing 2022 [Internet]. Available from: Available from: <https://www.R-project.org/>
21. Wickham H. *Ggplot2: elegant graphics for data analysis*. New York: Springer; 2009. 212 p. (Use R!).
22. Anderson-Bergman C. *icenReg : Regression Models for Interval Censored Data in R*. J Stat Soft [Internet]. 2017 [cited 2022 Aug 24];81(12). Available from: <http://www.jstatsoft.org/v81/i12/>
23. Walker TM, Choisy M, Dedicoat M, Drennan PG, Wyllie D, Yang-Turner F, et al. *Mycobacterium tuberculosis* transmission in Birmingham, UK, 2009–19: An observational study. *The Lancet Regional Health - Europe*. 2022 Jun;17:100361.

24. Urbanowski ME, Ordonez AA, Ruiz-Bedoya CA, Jain SK, Bishai WR. Cavitory tuberculosis: the gateway of disease transmission. *The Lancet Infectious Diseases*. 2020 Jun;20(6):e117–28.
25. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat Genet*. 2013 Sep 1;45(10):1176–82.
26. Manson AL, Abeel T, Galagan JE, Sundaramurthi JC, Salazar A, Gehrman T, et al. *Mycobacterium tuberculosis* Whole Genome Sequences From Southern India Suggest Novel Resistance Mechanisms and the Need for Region-Specific Diagnostics. *Clinical Infectious Diseases*. 2017 Jun 1;64(11):1494–501.
27. Smittipat N, Miyahara R, Juthayothin T, Billamas P, Dokladda K, Imsanguan W, et al. Indo-Oceanic *Mycobacterium tuberculosis* strains from Thailand associated with higher mortality. *Int J Tuberc Lung Dis*. 2019 Sep 1;23(9):972–9.
28. Holt KE, McAdam P, Thai PVK, Thuong NTT, Ha DTM, Lan NN, et al. Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat Genet*. 2018 Jun;50(6):849–56.
29. Freschi L, Vargas R, Husain A, Kamal SMM, Skrahina A, Tahseen S, et al. Population structure, biogeography and transmissibility of *Mycobacterium tuberculosis*. *Nat Commun*. 2021 Dec;12(1):6099.
30. Click ES, Winston CA, Oeltmann JE, Moonan PK, Mac Kenzie WR. Association between *Mycobacterium tuberculosis* lineage and time to sputum culture conversion. *Int J Tuberc Lung Dis*. 2013 Jul;17(7):878–84.



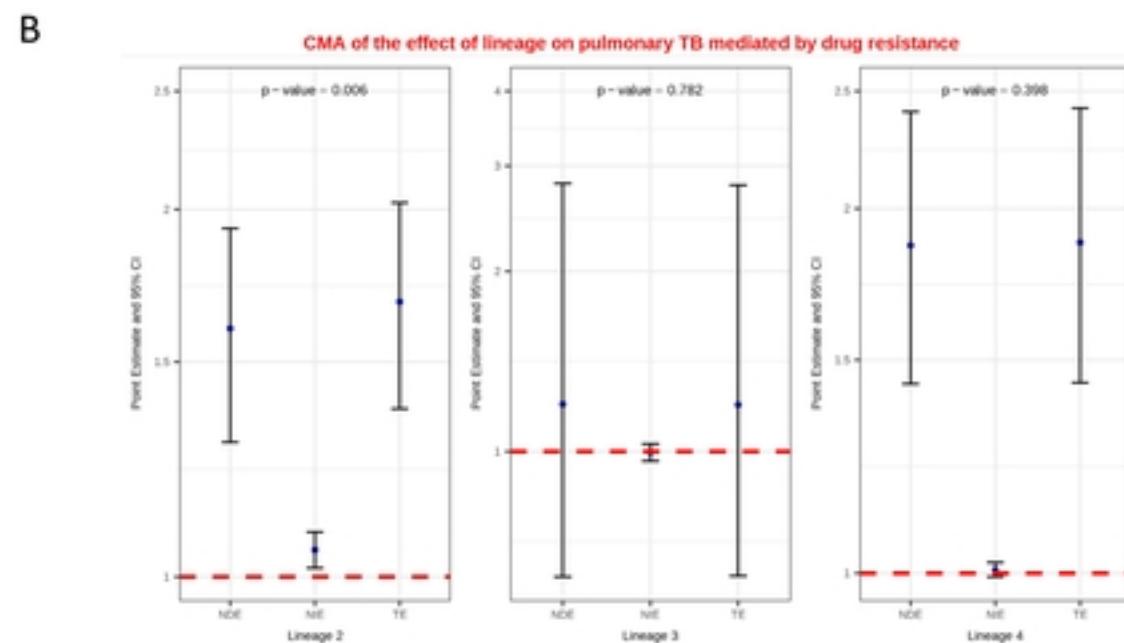
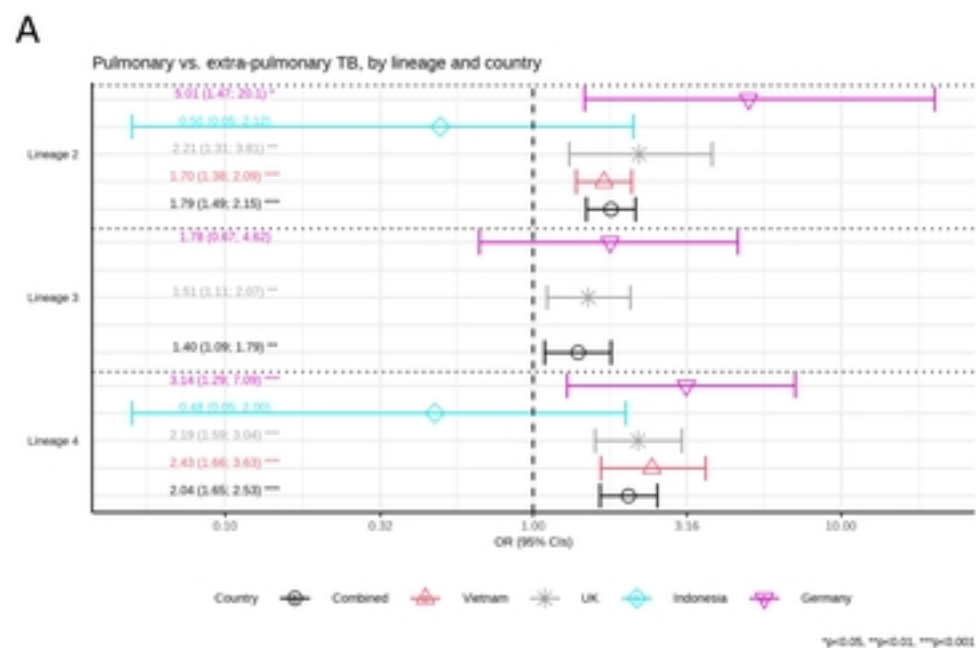


Figure 1

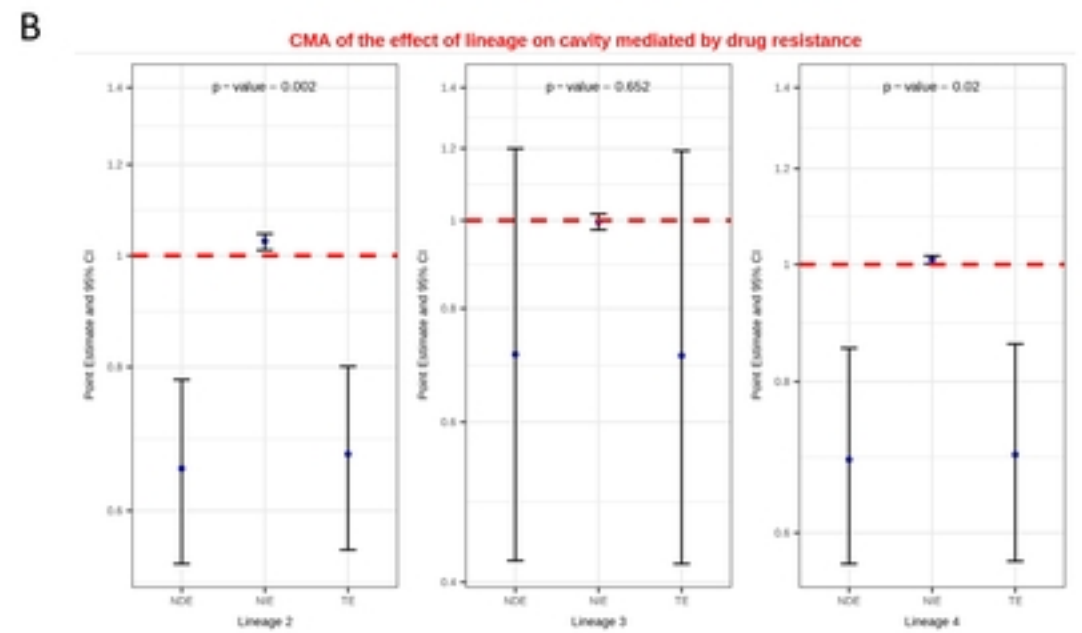
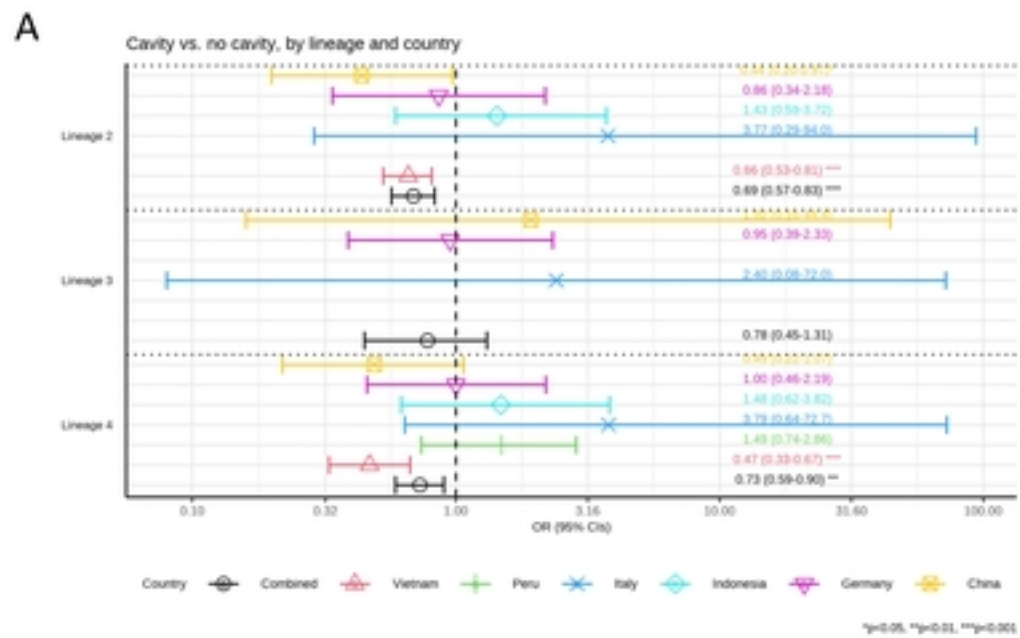
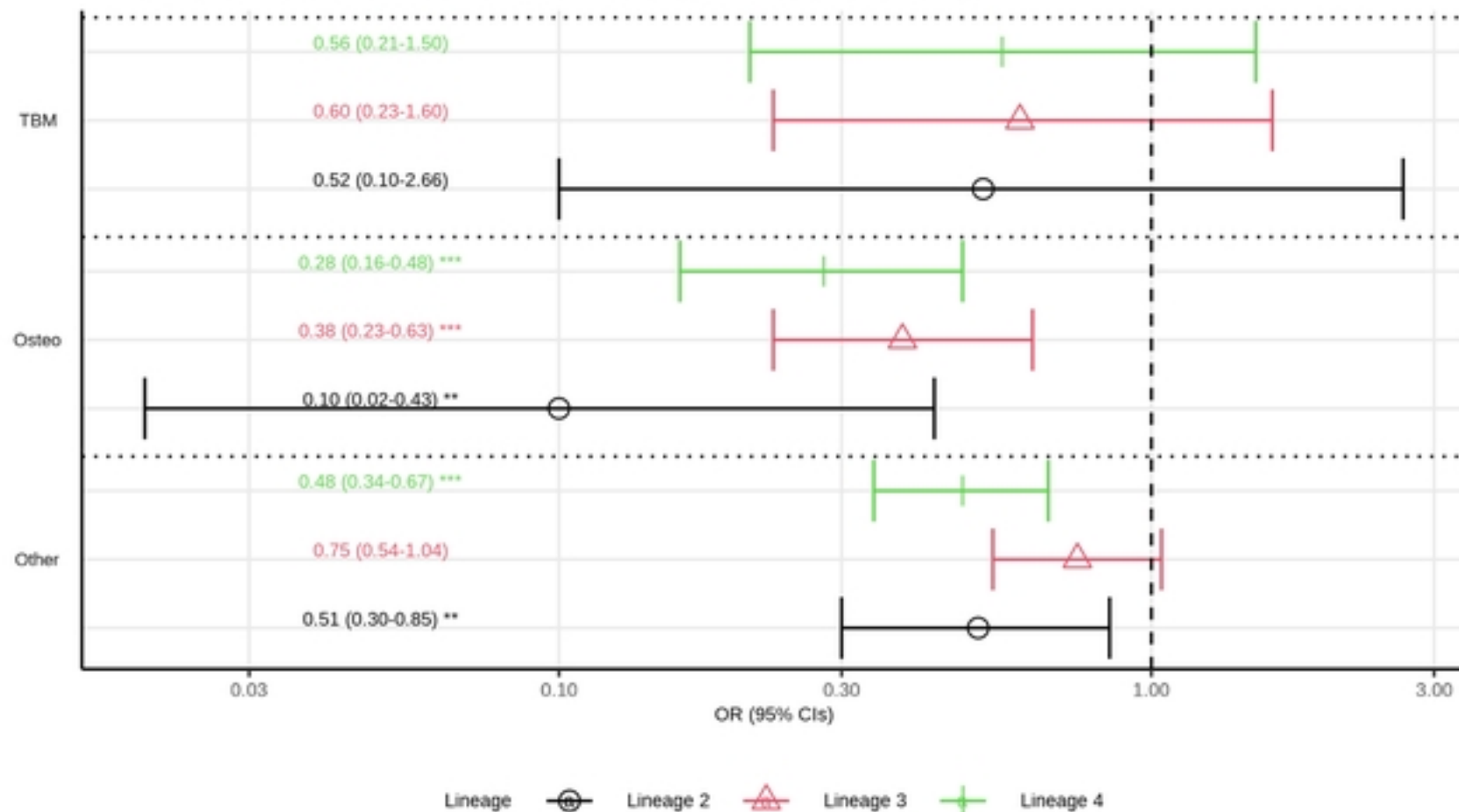


Figure 2

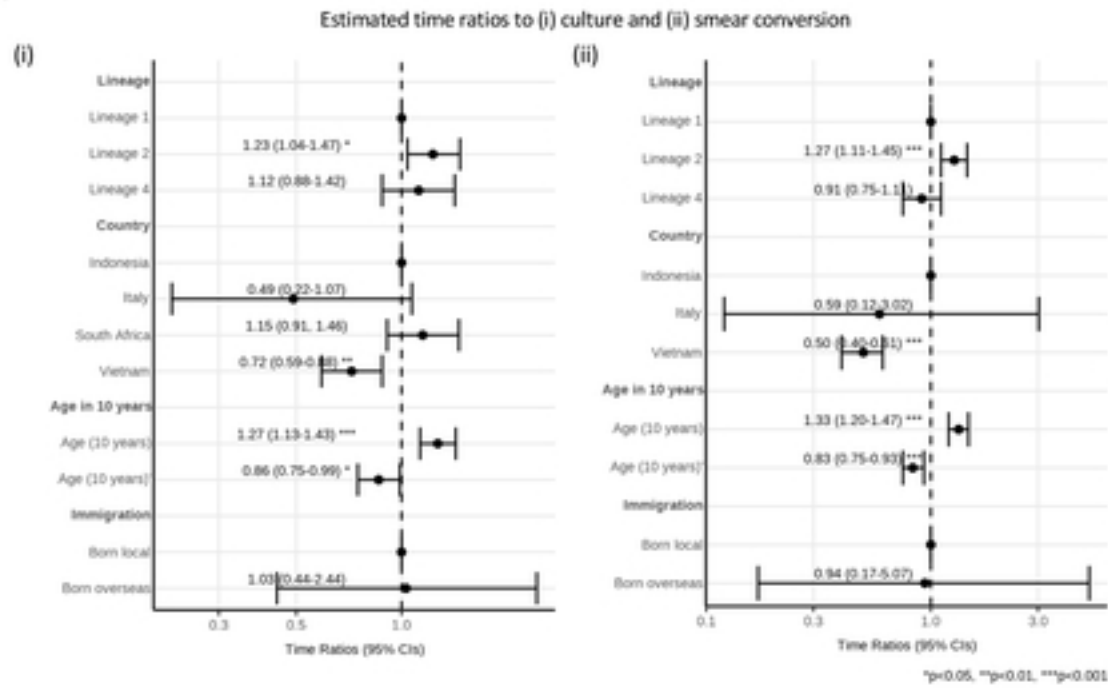
Estimated odds ratio of different forms of EPTB compared to PTB, by lineage



\*p<0.05, \*\*p<0.01, \*\*\*p<0.001

Figure3

A



B

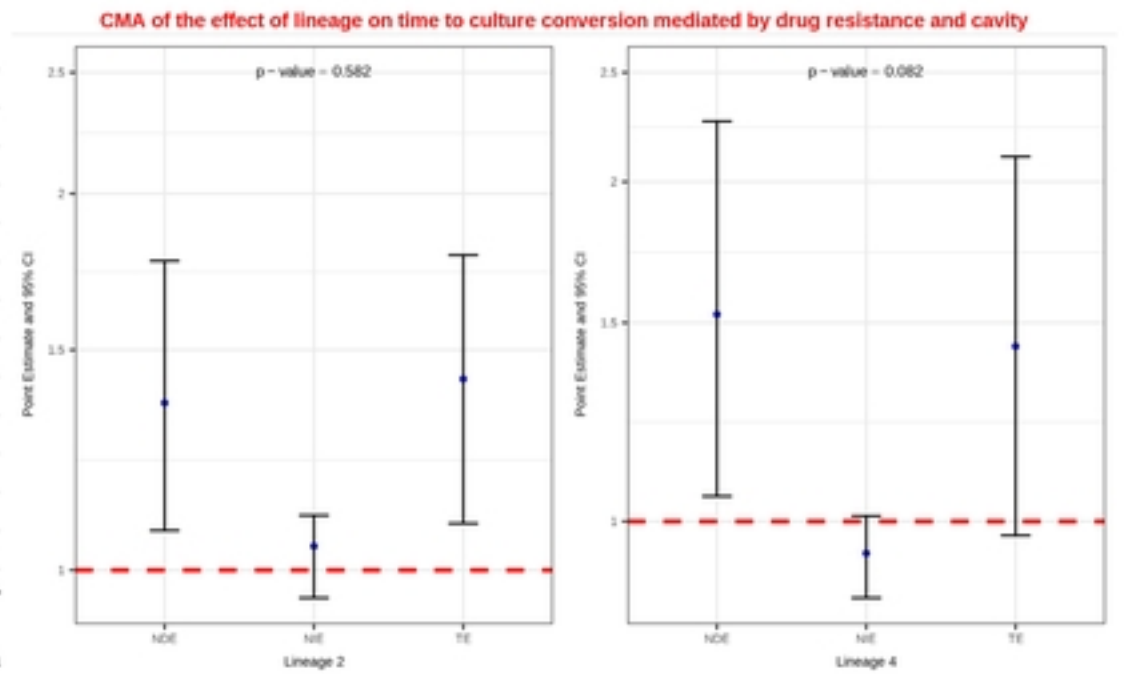
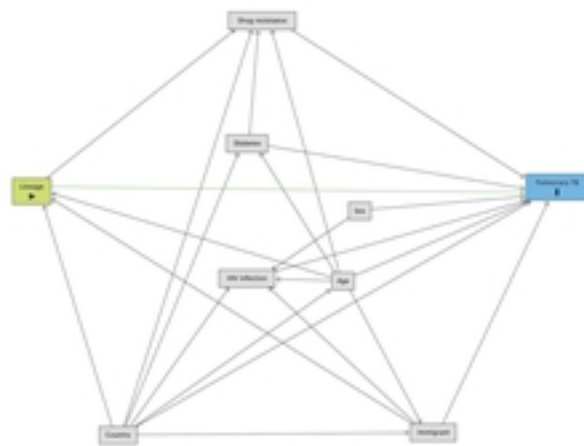


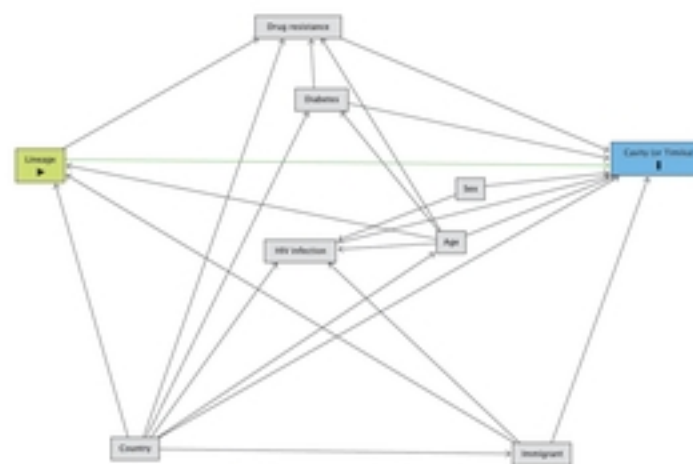
Figure4

A



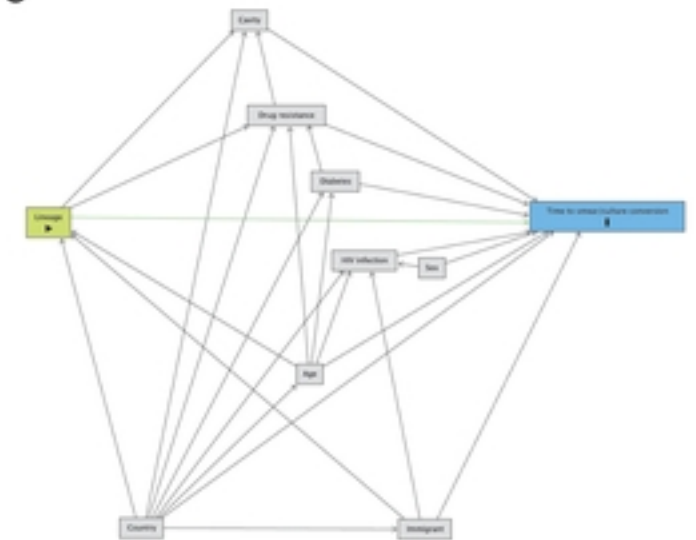
**Exposure: Lineage**  
**Mediator: Drug resistance**  
**Outcome: Pulmonary TB**  
**Confounders: Country, Immigration, Age**

B

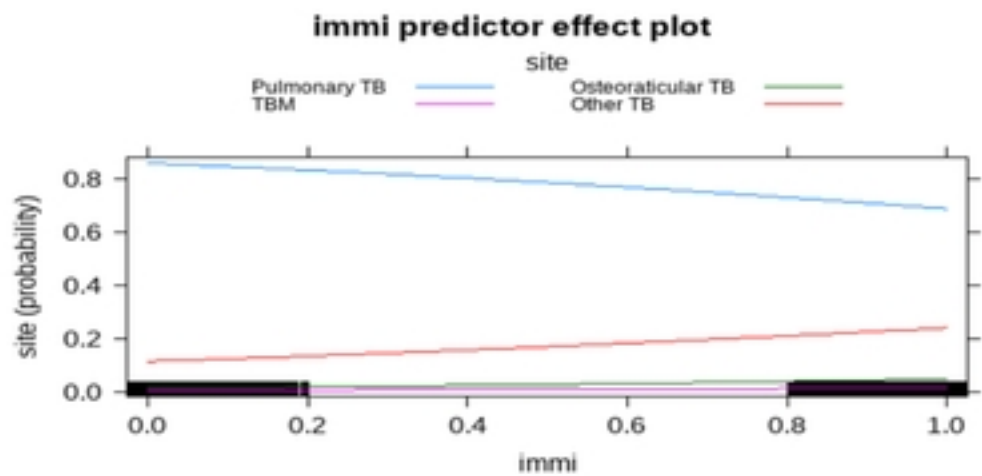
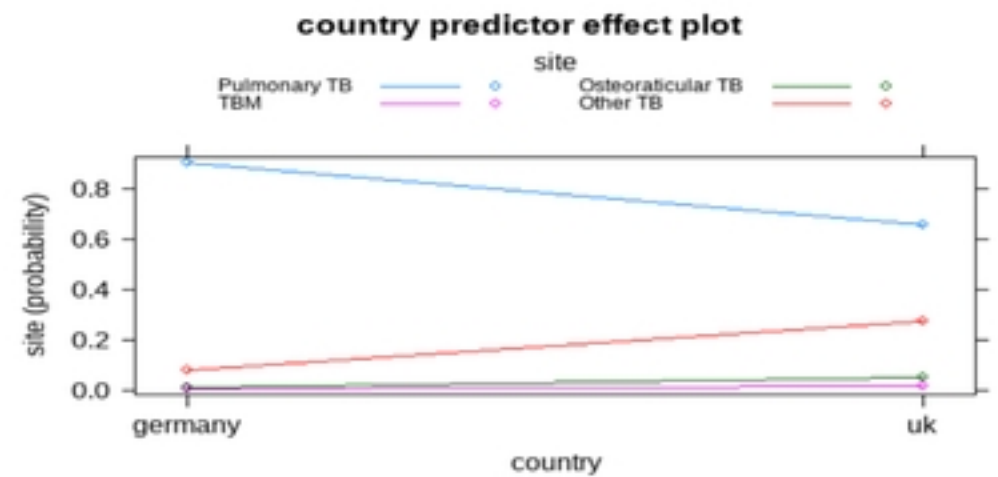
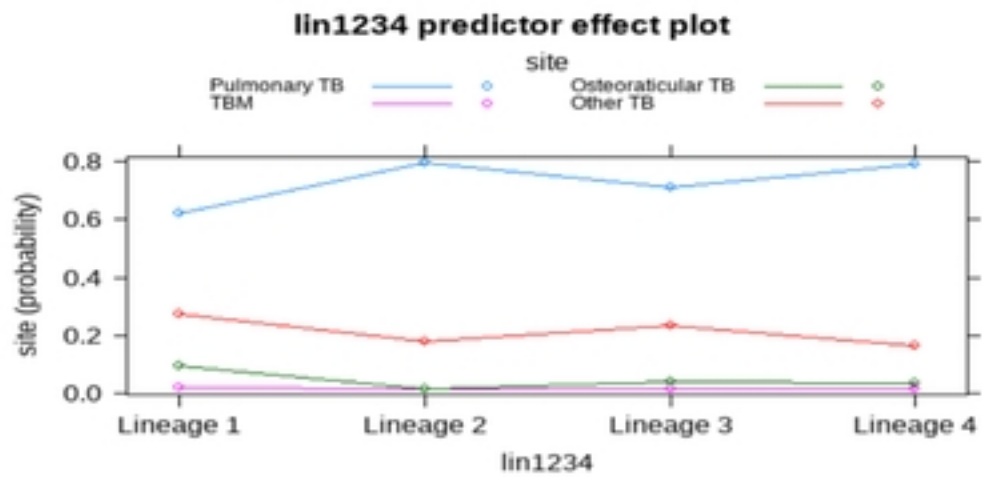


**Exposure: Lineage**  
**Mediator: Drug resistance**  
**Outcome: Cavity**  
**Confounders: Country, Immigration, Age**

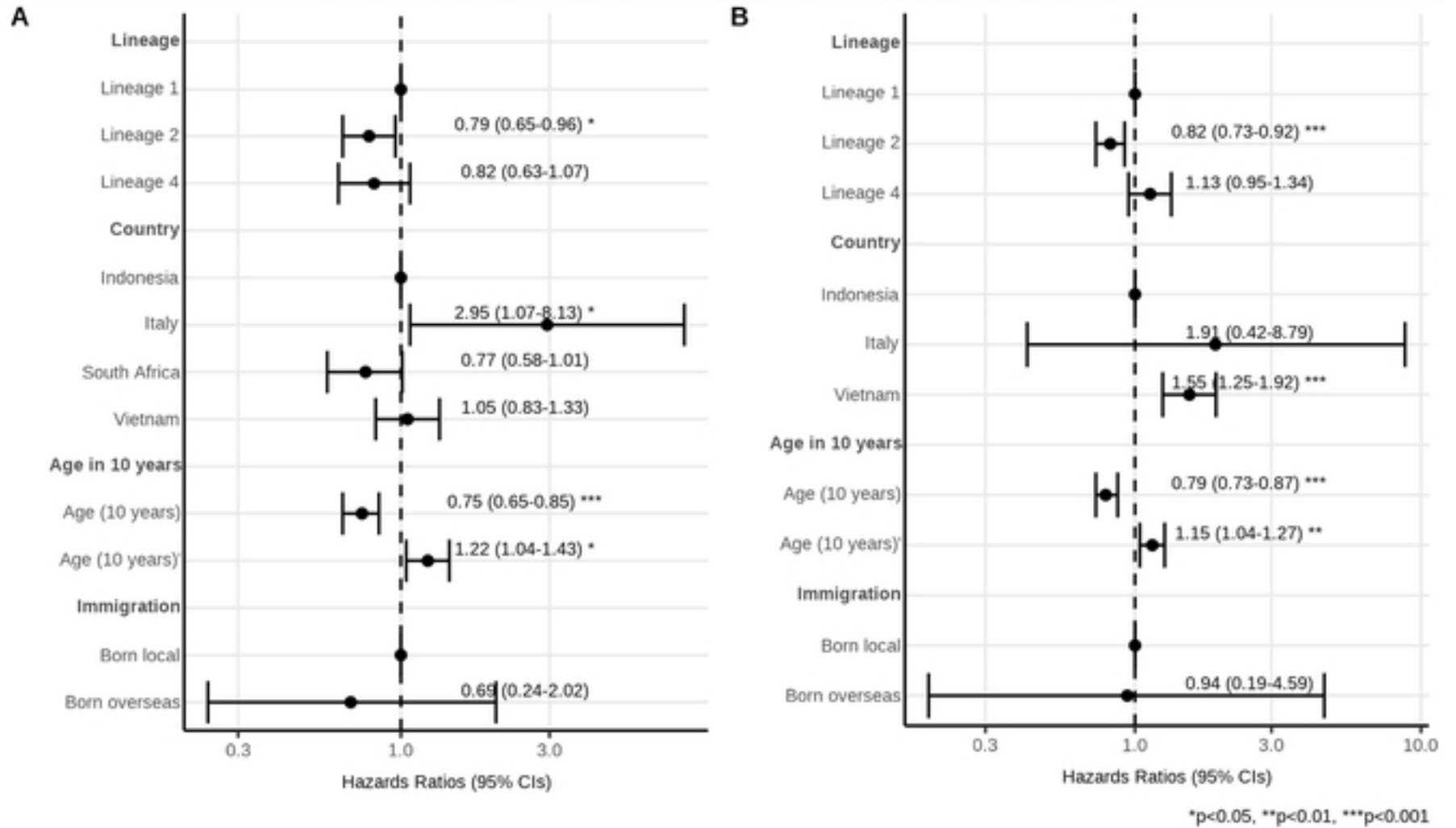
C



**Exposure: Lineage**  
**Mediator: Drug resistance, Cavity**  
**Outcome: Time to culture/smear conversion**  
**Confounders: Country, Immigration, Age**



Estimated hazards ratios to culture (A) and smear (B) conversion



**CMA of the effect of lineage on time to smear conversion mediated by drug resistance and cavity**

