Patients with presumed tuberculosis in sub-Saharan Africa that are not diagnosed with tuberculosis: a systematic review and meta-analysis (statistical appendix)

S Jayasooriya, F Dimambro-Denson, C Beecroft, J Balen, B Awokola, C Mitchell, B Kampmann, F Campbell, PJ Dodd, K Mortimer

August, 2021

Contents

Pre-amble

This document is generated from an R script in literate programming fashion. Some code, output and figures are specified for inclusion of the output document. The script and data are publicly available on GitHub at https://github.com/petedodd/NotTB and once the repository is downloaded, it should be possible to generate this document using R with the command `rmarkdown::render('NotTBmeta.R')` within R, or from a unix-like command line with `R -q -e "rmarkdown::render(\"NotTBmeta.R\",output_dir=\"./output\")"`. Alternatively, the R script can be run in whole or part as a conventional R script.

Dependencies

To compile this document, the rmarkdown & knitr packages must be installed. The other R packages required to run this analysis should be installed if necessary, and loaded, with:

```
pkgs.needed <- c("ggplot2","scales","cowplot","ggpubr", #graphs
                 "data.table","here",                   #data mgt
                 "metafor")                             #metaanalysis
install.packages(setdiff(pkgs.needed, rownames(installed.packages())))
```

```
suppressMessages(
    devnull <- lapply(pkgs.needed, require, character.only = TRUE) #load for use
)
```

This analysis was run using:

```
sI <- sessionInfo()
dI <- data.frame(
    item=c('R version','platform','OS','metafor version'),
    version=c(
        sI$R.version$version.string, #R version
        sI$platform,                 #platofm
        sI$running,                  #OS
        sI$otherPkgs$metafor$Version #metafor version
    )
)
knitr::kable(dI)
```

item	version
R version	R version 4.1.0 (2021-05-18)
platform	x86_64-pc-linux-gnu (64-bit)
OS	Pop!_OS 21.04
metafor version	3.0-2

Main analyses

Approach

We use a random-effects meta-analysis assuming a binomial response and logit link.

$$k_i \sim \text{Binomial}(N_i, p_i)$$
$$\text{logit}(p_i) = \mu + \varepsilon_i$$
$$\varepsilon_i \sim \mathcal{N}(0, \sigma)$$

where $k = 1, \ldots, S$ indexes the numbers of studies.

Use of arcinse or double arcsine transformations has been criticized in this context, with the binomial model above recommended.[1]

check formulae

Read in the data and ensure that factors behave as intended:

```
DD <- fread(file=here('SRMAdata.csv'))
DD[,lab:=factor(lab,levels=rev(DD[order(bac)]$lab),ordered = TRUE)]
```

Create exact binomial confidence intervals:

```
ciz <- function(x,y){
    x <- as.integer(x); y <- as.integer(y)
    list(binom.test(x,y)$conf.int[1],binom.test(x,y)$conf.int[2])
}
```

[1] Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions, by Schwarzer et al.

```
DD[,`NotTB Proportion`:=NnotTB/N]
for(i in 1:nrow(DD)){ DD[i,c('lo','hi'):=ciz(NnotTB,N)]; }
DD[,SE:=(hi-lo)/3.92]
```

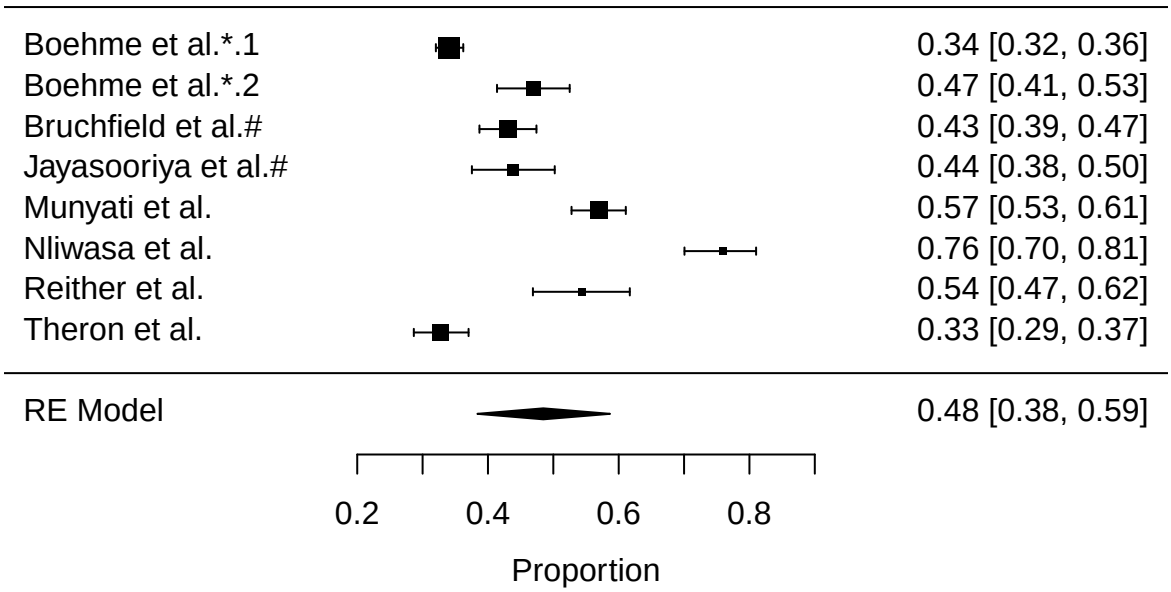
Meta-analyses

Meta-analysis for passively found TB patients with bacteriologically unconfirmed TB included:

```
maPU <- rma(measure = "PLO", #  binomial w/ logit link
            xi = NnotTB,     # numerator
            ni = N,          # denominator
            data = DD[mode=='Passive' &
                      clinical=='(Unconfirmed TB included)'],
            slab = Author)      # what to use as labels on graphs
summary(maPU)
```

```
##
## Random-Effects Model (k = 8; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
##  -6.3265   12.6530   16.6530   16.5448   19.6530
##
## tau^2 (estimated amount of total heterogeneity): 0.3403 (SE = 0.1888)
## tau (square root of estimated tau^2 value):      0.5833
## I^2 (total heterogeneity / total variability):   97.41%
## H^2 (total variability / sampling variability):  38.63
##
## Test for Heterogeneity:
## Q(df = 7) = 221.8886, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb   ci.ub
##  -0.0614  0.2101  -0.2920  0.7703  -0.4732  0.3505
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forest(maPU,transf = transf.ilogit,refline=NA)
```

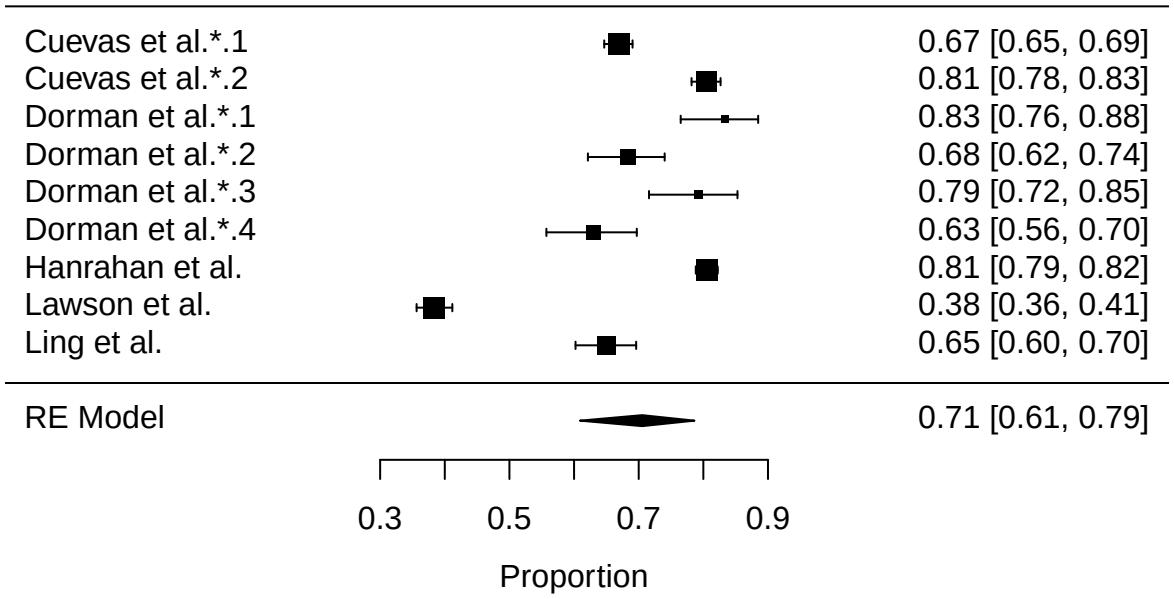


Meta-analysis for passively found TB patients with bacteriologically unconfirmed TB excluded:

```
maPN <- rma(measure = "PLO", #  binomial w/ logit link
            xi = NnotTB,     # numerator
            ni = N,          # denominator
            data = DD[mode=='Passive' &
                      clinical=='(No unconfirmed TB)'],
            slab = Author)      # what to use as labels on graphs
summary(maPN)
```

```
##
## Random-Effects Model (k = 9; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
##  -7.9621   15.9243   19.9243   20.0832   22.3243
##
## tau^2 (estimated amount of total heterogeneity): 0.4153 (SE = 0.2163)
## tau (square root of estimated tau^2 value):      0.6445
## I^2 (total heterogeneity / total variability):   98.34%
## H^2 (total variability / sampling variability):  60.10
##
## Test for Heterogeneity:
## Q(df = 8) = 679.9414, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.8728  0.2193  3.9803  <.0001  0.4430  1.3025  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forest(maPN,transf = transf.ilogit,refline=NA)
```

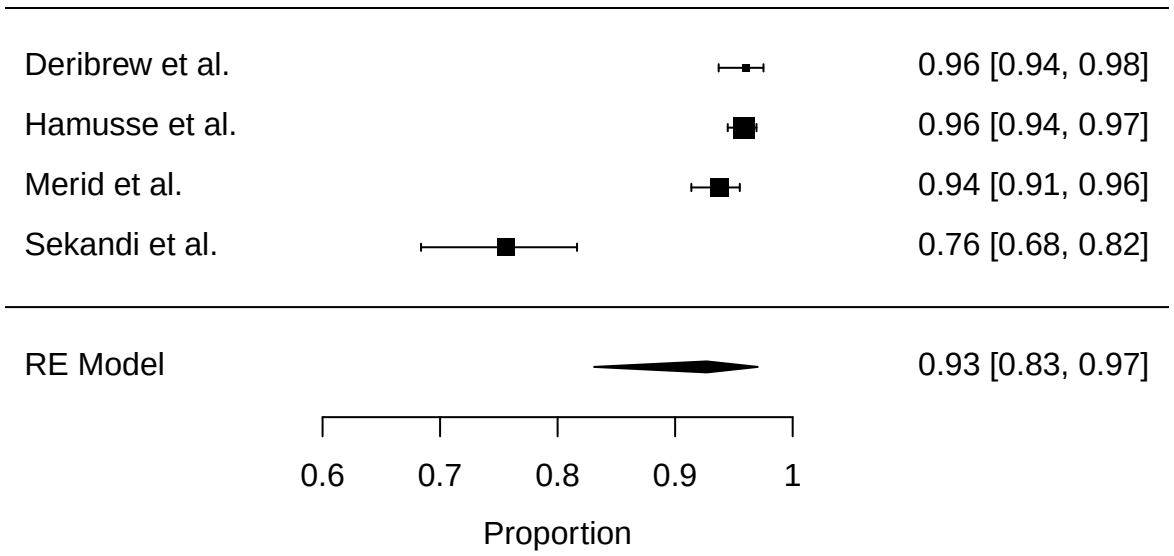


Meta-analysis for actively found TB patients:

```
maA <- rma(measure = "PLO", #  binomial w/ logit link
           xi = NnotTB,     # numerator
           ni = N,          # denominator
           data = DD[mode=='Active'],
           slab = Author)      # what to use as labels on graphs
summary(maA)
```

```
##
## Random-Effects Model (k = 4; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
##  -4.1508    8.3015   12.3015   10.4987   24.3015
##
## tau^2 (estimated amount of total heterogeneity): 0.8952 (SE = 0.7615)
## tau (square root of estimated tau^2 value):      0.9462
## I^2 (total heterogeneity / total variability):   96.27%
## H^2 (total variability / sampling variability):  26.82
##
## Test for Heterogeneity:
## Q(df = 3) = 81.2135, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   2.5396  0.4829  5.2593  <.0001  1.5932  3.4861  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forest(maA,transf = transf.ilogit,refline=NA)
```



Make predictions for plot data:

```
map <- predict(maA,transf = transf.ilogit)
mup <- predict(maPU,transf = transf.ilogit)
mnp <- predict(maPN,transf = transf.ilogit)
```

Creation of combined forest plot

Summary data for combined forest plot:

```
f1 <- function(x)format(round(x,1),nsmall=1)
cnz <- c('(Unconfirmed TB included)',
         '(No unconfirmed TB)',
         '(No unconfirmed TB)')
predz <- data.table(mode=c('Passive','Passive','Active'),
                    clinical=cnz,
                    `NotTB Proportion` = c(mup$pred,mnp$pred,map$pred),
                    lo = c(mup$ci.lb,mnp$ci.lb,map$ci.lb),
                    hi = c(mup$ci.ub,mnp$ci.ub,map$ci.ub),
                    lab=paste0('SUMMARY (',expression(I^2),'=',
                               f1(c(maA$I2,maPN$I2,maPU$I2)),'%)')
                    )
predz[,SE:=(hi-lo)/3.92]
predz[,qty:='summary']
predz[,bac:=0]
predz[,mid:=`NotTB Proportion`]
predz[,CI:=paste0(f1(1e2*mid),' (',f1(1e2*lo),' - ',f1(1e2*hi),')')]
predz[,wt:='100.0%']
predz[,w:=1]
```

Appending plot data to inputs:

```
DD[,qty:='study']
DD[,mid:=`NotTB Proportion`]
DD[,CI:=paste0(f1(1e2*mid),' (',f1(1e2*lo),' - ',f1(1e2*hi),')')]
DD[,wt:=1/SE^2]
DD[,wtt:=sum(wt),by=.(mode,clinical)]
DD[,wt:=1e2*wt/wtt]
```

```
DD[,wt:=paste0(f1(wt),'%')]
DD[,w:=0]
```

Combined plot data:

```
B <- rbind(
    DD[,.(lab,`NotTB Proportion`,lo,hi,SE,mode,clinical,
          qty,bac,CI,wt,w)],
    predz[,.(lab,`NotTB Proportion`,lo,hi,SE,mode,clinical,
             qty,bac,CI,wt,w)]
)
lbz <- as.character(B[order(bac)]$lab)
lbz2 <- c(lbz[1:3],rev(lbz[-c(1:3)]))
B[,lab:=factor(lab,levels=lbz2,ordered = TRUE)]
B[,clinical.g:='Clinically diagnosed tuberculosis included']
B[clinical=='(No unconfirmed TB)',
  clinical.g:='No clinically diagnosed tuberculosis included']
B[mode=='Active',clinical.g:='']
B[,mode:=factor(mode,levels=c('Passive','Active'),ordered = TRUE)]
B[,clinical.g:=factor(clinical.g,levels=unique(clinical.g))]
labdat <- B[1]
labdat[,txt:=' weight (%)']
```

Create publication forest plot figure:

```
SA <- ggplot(B,aes(lab,y=`NotTB Proportion`,
                   ymin=lo,
                   ymax=hi,
                   col=qty)) +
    geom_point(aes(size=1/SE^2,shape=qty)) +
    geom_errorbar(aes(width=w/2)) +
    scale_y_continuous(label=percent,limits = c(0,NA))+
    scale_color_manual(values=c('study'="black",'summary'="blue"))+
    scale_shape_manual(values=c('study'=22,'summary'=23))+
    xlab('') +
    ylab('Proportion of patients with presumptive tuberculosis not diagnosed as tuberculosis')+
    facet_grid(mode + clinical.g ~ .,
               scales = 'free',space='free',
               switch='x'
               )+
    coord_flip() +
    guides(size='none',color='none',shape='none')+
    theme_classic() +
    theme(panel.spacing = unit(2, "lines"), #or 3
          strip.background = element_blank(),
          strip.placement = "outside") +
    geom_text(aes(x=lab,y=1.2,label=CI,hjust='right')) +
    geom_text(aes(x=lab,y=0.0,label=wt))+
    geom_text(data=labdat,aes(x=9.5,y=0,label=txt))+
    ggpubr::grids()

ggsave(SA,file=here('output/ForestPlot.pdf'),h=13,w=12)
ggsave(SA,file=here('output/ForestPlot.eps'),h=13,w=12)
```

Meta-regressions

In this section we consider various potential sources of heterogeneity through scatter plots and meta-regression.

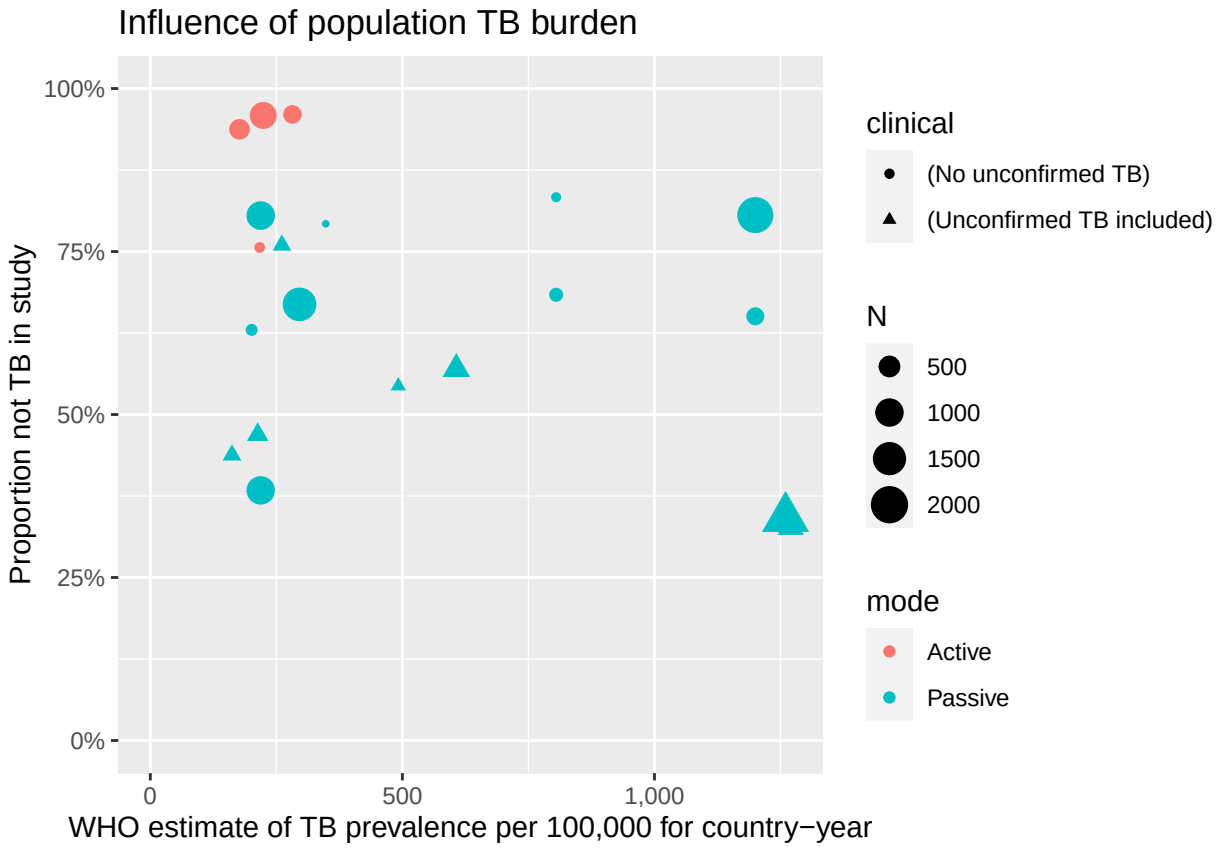
TB prevalence

The burden of TB in a population might reasonably be expected to influence the proportion of presumptive TB that is not TB.

```
DD[,tb:=`WHO TB estimate (per 100 000 year of study)`]

ggplot(DD,aes(tb,`NotTB Proportion`,
              size=N,col=mode,shape=clinical))+
    scale_x_continuous(label=comma,limits=c(0,NA))+
    scale_y_continuous(label=percent,limits=c(0,1))+
    geom_point()+
    xlab('WHO estimate of TB prevalence per 100,000 for country-year')+
    ylab('Proportion not TB in study')+
    ggtitle('Influence of population TB burden')
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



We can formally investigating the influence of TB burden in explaining heterogeneity with a meta-regression:

```
tbmr <-  rma(measure = "PLO",  #binomial w/ logit link
             xi = NnotTB,      # numerator
             ni = N,           # denominator
             data = DD,         # what data to use
             mods = ~mode*clinical + tb)
```

```
## Warning: Studies with NAs omitted from model fitting.
```

```
## Warning: Redundant predictors dropped from the model.
```

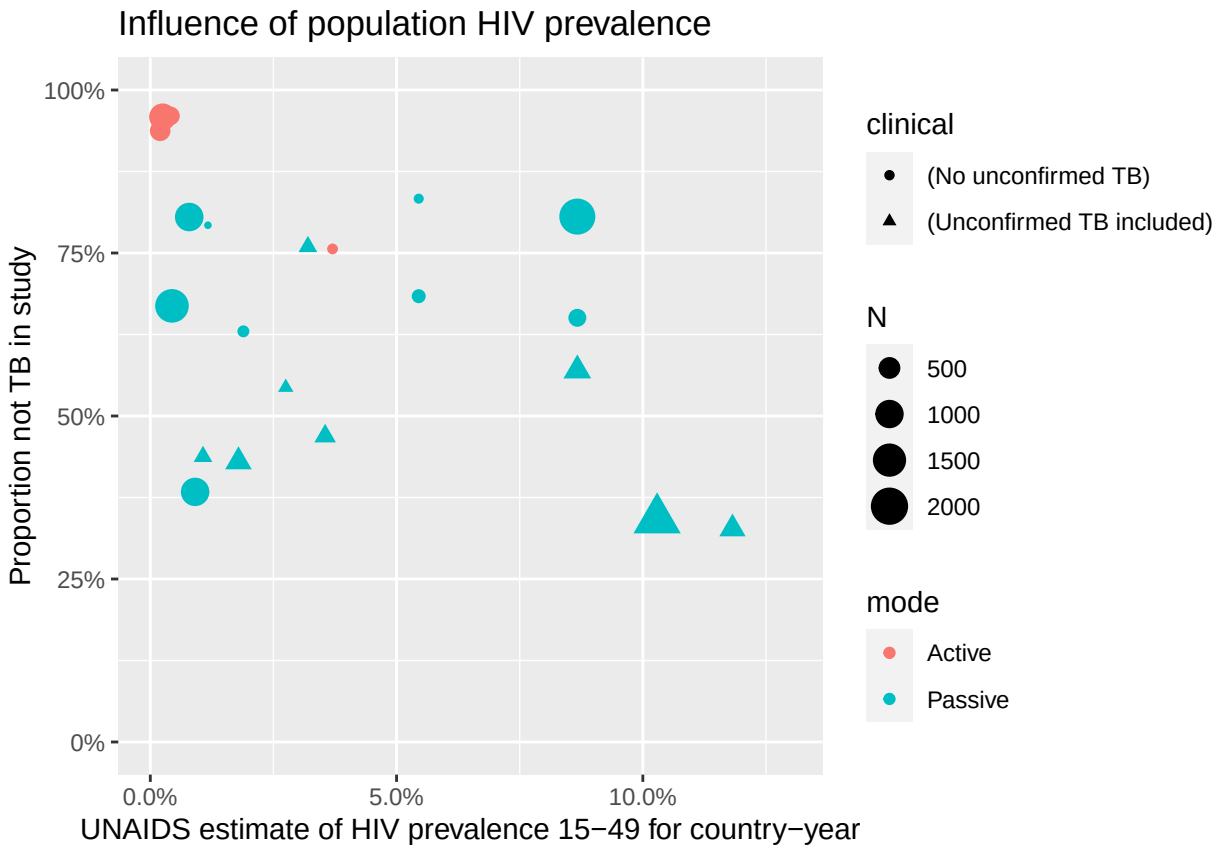
```
summary(tbmr)
```

```
##
## Mixed-Effects Model (k = 20; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
## -17.6992   35.3984   45.3984   49.2614   51.3984
##
## tau^2 (estimated amount of residual heterogeneity):     0.5137 (SE = 0.1887)
## tau (square root of estimated tau^2 value):             0.7167
## I^2 (residual heterogeneity / unaccounted variability): 98.12%
## H^2 (unaccounted variability / sampling variability):   53.20
## R^2 (amount of heterogeneity accounted for):            60.59%
##
## Test for Residual Heterogeneity:
## QE(df = 16) = 973.5088, p-val < .0001
##
## Test of Moderators (coefficients 2:4):
## QM(df = 3) = 30.9942, p-val < .0001
##
## Model Results:
##
##                                    estimate      se     zval    pval    ci.lb
## intrcpt                              2.5734  0.3838   6.7055  <.0001   1.8212
## modePassive                         -1.6053  0.4710  -3.4080  0.0007  -2.5286
## clinical(Unconfirmed TB included)   -0.8972  0.3667  -2.4465  0.0144  -1.6159
## tb                                  -0.0002  0.0004  -0.3650  0.7151  -0.0010
##                                      ci.ub
## intrcpt                             3.3256  ***
## modePassive                        -0.6821  ***
## clinical(Unconfirmed TB included)  -0.1784    *
## tb                                  0.0007
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

HIV prevalence

Population HIV prevalence may plausibly influence the proportion of presumptives not diagnosed with TB both by inluencing TB burden, but also by changing the typical clinical characteristics of TB and most importantly, the burden of other illness that could be designated presumptive TB.

```
ggplot(DD,aes(hiv/1e2,`NotTB Proportion`,
              size=N,col=mode,shape=clinical))+
    scale_x_continuous(label=percent,limits=c(0,0.13))+
    scale_y_continuous(label=percent,limits=c(0,1))+
    geom_point()+
    xlab('UNAIDS estimate of HIV prevalence 15-49 for country-year')+
    ylab('Proportion not TB in study')+
    ggtitle('Influence of population HIV prevalence')
```



We can formally investigating the influence of HIV in explaining heterogeneity with a meta-regression:

```
hivmr <-  rma(measure = "PLO",  #binomial w/ logit link
              xi = NnotTB,      # numerator
              ni = N,           # denominator
              data = DD,         # what data to use
              mods = ~mode*clinical + hiv)
```

```
## Warning: Redundant predictors dropped from the model.
```

```
summary(hivmr)
```

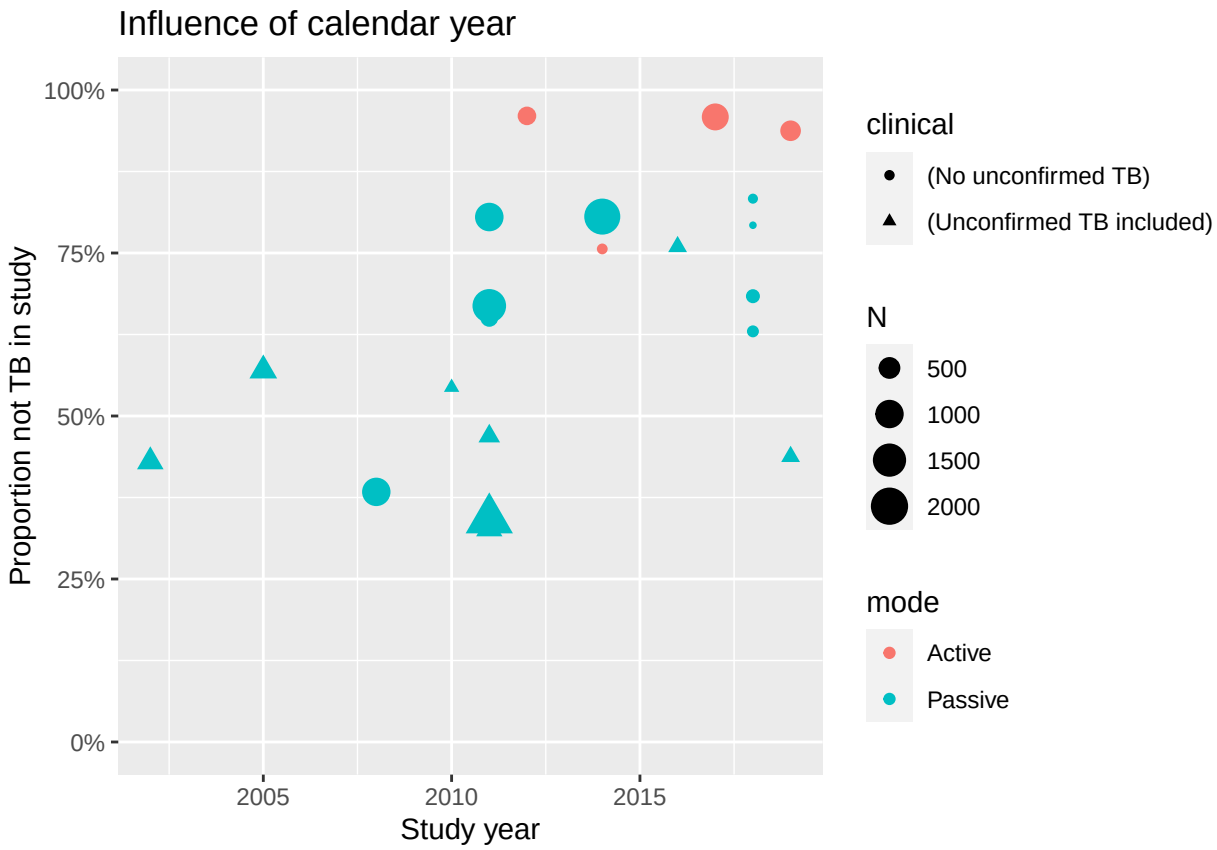
```
##
## Mixed-Effects Model (k = 21; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
## -18.1622   36.3244   46.3244   50.4904   51.7789
##
## tau^2 (estimated amount of residual heterogeneity):     0.4756 (SE = 0.1697)
## tau (square root of estimated tau^2 value):             0.6896
## I^2 (residual heterogeneity / unaccounted variability): 98.02%
## H^2 (unaccounted variability / sampling variability):   50.50
## R^2 (amount of heterogeneity accounted for):            63.48%
##
## Test for Residual Heterogeneity:
## QE(df = 17) = 973.1809, p-val < .0001
##
## Test of Moderators (coefficients 2:4):
```

```
## QM(df = 3) = 36.2039, p-val < .0001
##
## Model Results:
##
##                                    estimate      se     zval    pval    ci.lb
## intrcpt                              2.5749  0.3620   7.1120  <.0001   1.8653
## modePassive                         -1.5794  0.4443  -3.5546  0.0004  -2.4502
## clinical(Unconfirmed TB included)   -0.8771  0.3497  -2.5085  0.0121  -1.5624
## hiv                                 -0.0327  0.0467  -0.7010  0.4833  -0.1242
##                                      ci.ub
## intrcpt                             3.2845  ***
## modePassive                        -0.7085  ***
## clinical(Unconfirmed TB included)  -0.1918    *
## hiv                                 0.0587
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calendar time

To explore whether there has been any change over time, we consider calendar year

```
ggplot(DD,aes(Year,`NotTB Proportion`,
              size=N,col=mode,shape=clinical))+
    ## scale_x_continuous(label=percent,limits=c(0,0.13))+
    scale_y_continuous(label=percent,limits=c(0,1))+
    geom_point()+
    xlab('Study year')+
    ylab('Proportion not TB in study')+
    ggtitle('Influence of calendar year')
```



We can formally investigating the influence of year in explaining heterogeneity with a meta-regression:

```
yearmr <-  rma(measure = "PLO",  #binomial w/ logit link
               xi = NnotTB,     # numerator
               ni = N,          # denominator
               data = DD,        # what data to use
               mods = ~mode*clinical + Year)
```

```
## Warning: Redundant predictors dropped from the model.
```

```
summary(yearmr)
```

```
##
## Mixed-Effects Model (k = 21; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
## -17.6377   35.2753   45.2753   49.4414   50.7299
##
## tau^2 (estimated amount of residual heterogeneity):     0.4449 (SE = 0.1590)
## tau (square root of estimated tau^2 value):             0.6670
## I^2 (residual heterogeneity / unaccounted variability): 97.99%
## H^2 (unaccounted variability / sampling variability):   49.84
## R^2 (amount of heterogeneity accounted for):            65.84%
##
## Test for Residual Heterogeneity:
## QE(df = 17) = 882.4776, p-val < .0001
##
## Test of Moderators (coefficients 2:4):
```

```
## QM(df = 3) = 39.6262, p-val < .0001
##
## Model Results:
##
##                                    estimate       se     zval    pval
## intrcpt                           -88.4098  72.0566  -1.2270  0.2198
## modePassive                        -1.5936   0.4183  -3.8094  0.0001
## clinical(Unconfirmed TB included)  -0.7786   0.3515  -2.2149  0.0268
## Year                                0.0451   0.0357   1.2622  0.2069
##                                       ci.lb    ci.ub
## intrcpt                           -229.6380  52.8185
## modePassive                         -2.4135  -0.7737  ***
## clinical(Unconfirmed TB included)   -1.4676  -0.0896    *
## Year                                -0.0249   0.1152
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sensitivity analyses

Dorman et al. by country only

In the main analysis, we considered the different sites in the 2018 study by Dorman et al to be separate data. This included considering the two sites in South Africa - Cape Town and Johannesburg - as different, which was motivated by the very distinct TB epidemiology in the Western Cape. Here we investigate the impact of aggregating the two South African sites in Dorman et al on the meta-analysis for studies with passive case finding excluding clinically diagnosed TB.

Restrict to relevant data & aggregate over Dorman in South Africa:

```
tmp <- DD[mode=='Passive' & clinical=='(No unconfirmed TB)']
tmp[,Country.Simple:=gsub(" \\-.+$","",Country)]                    #remove cities
tmp[,authorcountry:=paste(gsub("^([A-Za-z]+).*","\\1",Author),Country.Simple,sep = ", ")] #new label
tmp <- tmp[,.(NnotTB=sum(NnotTB),N=sum(N)),by=authorcountry]
knitr::kable(tmp) #check
```

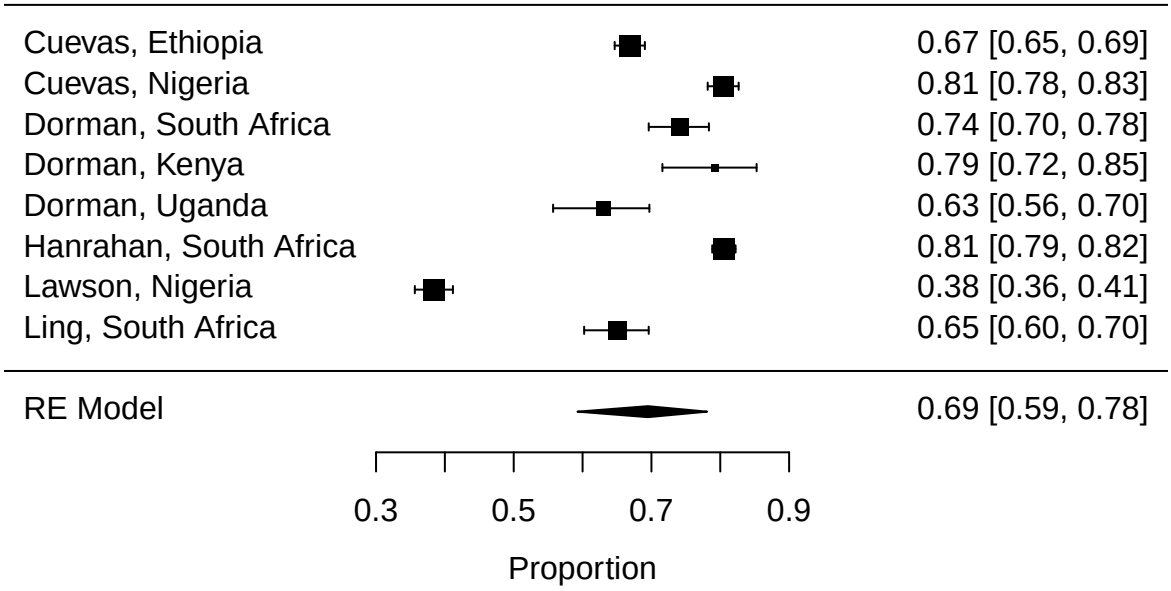
authorcountry	NnotTB	N
Cuevas, Ethiopia	1184	1770
Cuevas, Nigeria	963	1196
Dorman, South Africa	285	384
Dorman, Kenya	107	135
Dorman, Uganda	114	181
Hanrahan, South Africa	1685	2091
Lawson, Nigeria	455	1186
Ling, South Africa	257	395

Rerun this meta-analysis with the new data:

```
maPNsa <- rma(measure = "PLO", # binomial w/ logit link
            xi = NnotTB,     # numerator
            ni = N,          # denominator
            data =tmp,       # new data
            slab = authorcountry)      # what to use as labels on graphs
summary(maPNsa)
```

```
##
## Random-Effects Model (k = 8; tau^2 estimator: REML)
##
##   logLik  deviance       AIC       BIC      AICc
##  -6.8441   13.6881   17.6881   17.5800   20.6881
##
## tau^2 (estimated amount of total heterogeneity): 0.4057 (SE = 0.2238)
## tau (square root of estimated tau^2 value):      0.6370
## I^2 (total heterogeneity / total variability):   98.51%
## H^2 (total variability / sampling variability):  67.02
##
## Test for Heterogeneity:
## Q(df = 7) = 671.4861, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval   ci.lb   ci.ub
##   0.8231  0.2288  3.5974  0.0003  0.3746  1.2715  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
forest(maPNsa,transf = transf.ilogit,refline=NA)
```



This is very similar to the main analysis above.

Regional groupings

TODO