Supplementary information

Integrating serological and genetic data to quantify cross-species transmission dynamics: Brucellosis as a case study.

Mafalda Viana[1], Gabriel M. Shirima[2], Kunda S. John[3], Julie Fitzpatrick[4], Rudovick R. Kazwala[5],Joram J. Buza[2], Sarah Cleaveland[1], Daniel T. Haydon[1*] & Jo E.B. Halliday[1]

[1]*Boyd Orr Centre for Population and Ecosystem Health, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*
[2]*Nelson Mandela African Institution of Science and Technology, School of Life Sciences and Bioengineering, Arusha, Tanzania*
[3]*National Institute of Medical Research, PO Box 9653, 11101 Dar es Salaam, Tanzania*
[4]*Moredun Research Institute, Pentlands Science Park. Penicuik, Midlothian EH26 0PZ, UK*
[5]*Department of Veterinary Medicine and Public Health, Sokoine University of Agriculture, P.O. Box 3021, Morogoro, Tanzania*

[*]*Corresponding author:* daniel.haydon@glasgow.ac.uk

The main goal of the model presented in this paper is to identify the source of human *Brucella* infection and quantify the contribution of different potential source populations to the probability of infection in humans (p_h). We do this by integrating serological and genetic data. Using serology data alone we can identify which animal population (cows (c) or caprids (s)) drives human (h) infection. However, only its integration with genetic typing data will allow us to distinguish which animal population is infecting humans with which *Brucella* species, i.e. *Brucella abortus* (a) or *melitensis* (m).

Human infection probability

Ultimately we are interested in estimating the probability of infection with *Brucella abortus* (a) or *melitensis* (m) in humans in household i ($pType_h(i)$), which is defined through a logit transformation such that:

$$pType_h(i) = \frac{exp(lnType_h(i))}{1 + exp(lnType_h(i))} \tag{1}$$

where $lnType_h(i)$ is the linear predictor of the logit transformed ($pType_h(i)$), formulated as a function of covariates that describe the *Brucella* transmission process in humans as a function of the infection in cows and caprids. In this case, the linear predictor is:

$$lnType_h(i) = \beta_{0_h} + \beta_{1_h} Y_{c,a,i} + \beta_{2_h} Y_{s,a,i} + \beta_{3_h} Y_{c,m,i} + \beta_{4_h} Y_{s,m,i} \tag{2}$$

The coefficient β_{0_h} corresponds to the intercept, while β_{1_h} corresponds to the coefficient governing the effect of the estimated number of cows infected with *B. abortus* in the i^{th} household ($Y_{c,a,i}$), β_{2_h} to the estimated number of caprids infected with *B. abortus* ($Y_{s,a,i}$), β_{3_h} to the estimated number of cows infected with *B. melitensis* ($Y_{c,m,i}$) and β_{4_h} to the estimated number of caprids infected with *B. melitensis* ($Y_{s,m,i}$). For further details on the estimation of these Y values see the subsection "Livestock infection with *B. abortus* & *melitensis*". The priors for all β coefficients are provided in Table S2.

In turn, $pType_h(i)$ is the mean probability of an individual from household i being seropositive. Assuming that the j^{th} individual was sampled in household i, the probability $r_h(j, i)$ that, at sampling, the j^{th} individual was seropositive is:

$$pType_h(i) = mean(r_h(, i)) \tag{3}$$

This probability - $r_h(j,i)$ - translates the individual level probability of infection to the household level - $pType_h(i)$ - generating the link to integrate serology data (collected at the individual level) and genetics (estimated at the household level). Furthermore, $r_h(j,i)$ corresponds to the probability of an individual being truly seropositive after taking account of diagnostic test performance and the potential for misclassification of the serostatus of an individual (see below).

Serological Test Performance

To account for the potential for misclassification of individual serostatus (considering both human and animal serological test data) we introduce probabilities of Type I and Type II errors in serological diagnostic testing. Based on the serological literature, we expect high probability of correct classification (q^+ & q^-) and low probability of false classification ($1-q^+$ & $1-q^-$) (Table S1 & S2).

Table S1: Probabilities associated with misclassification of *Brucella* serostatus.

		True state	
		+	-
Test	+	q^+	$1-q^-$
result	-	$1-q^+$	q^-

The likelihood that an individual j is classified as seropositive ($P(y_j = 1)$) or seronegative ($P(y_j = 0)$) is based on serological test data y and was defined in our model as:

$$P(y_j = 1) = r_h(j,i)q^+ + (1 - r_h(j,i))(1 - q^-) \quad (4)$$
$$P(y_j = 0) = r_h(j,i)(1 - q^+) + (1 - r_h(j,i))q^- \quad (5)$$

The likelihood that the data y from individual j was generated from a Bernoulli distribution with success probability P, i.e. probability of classifying an individual as positive upon testing, is:

$$y_{i,h} \sim Bernoulli(P(y_{j,h} = 1)) \quad (6)$$

Where $y_{i,h} = 1$ corresponds to a *Brucella* positive titer and $y_{i,h} = 0$ to a *Brucella* negative titer. If both realisations are equally likely, $P(y_h = 1) = P(y_h = 0) = 0.5$.

Livestock infection with *B. abortus* & *B. melitensis*

A similar approach to that described for human infection was used to estimate the probability of cattle and caprid infection and the number of animals in each household infected with *B. abortus* and/or *B. melitensis*. However, in contrast to the serology model where we included the potential for generation of false positive or false negative test results, a similar individual level mechanism was not used for genetic typing data. In this step of the model we estimate the proportion of seropositive animals that are positive for each *Brucella* species. We consider that all seropositive individuals are genetically positive and therefore did not include evaluation of the performance of the genetic diagnostic tests (see manuscript for further evaluation and handling of this assumption). In this way, the integration of the serological and genetic diagnostic test data enable estimation of more robust estimates of the proportion of individuals infected with each genetic species than would be possible using just the genetic data alone.

We describe the probabilities that a seropositive cow or caprid in the i^{th} household is infected with *B. melitensis* or *B. abortus* through logit transformations of the linear predictors (i.e. $lnType_{c,m}(i)$, $lnType_{s,m}(i)$, $lnType_{c,a}(i)$ or $lnType_{s,a}(i)$). These are described by the number of cattle and sheep present at the i^{th} household as covariates.

$$lnType_{c,m}(i) = \beta_{0_{c,m}} + \beta_{1_{c,m}} N_{c,i} + \beta_{2_{c,m}} N_{s,i} \tag{7}$$

$$lnType_{s,m}(i) = \beta_{0_{s,m}} + \beta_{1_{s,m}} N_{s,i} + \beta_{2_{s,m}} N_{c,i} \tag{8}$$

$$lnType_{c,a}(i) = \beta_{0_{c,a}} + \beta_{1_{c,a}} N_{c,i} + \beta_{2_{c,a}} N_{s,i} \tag{9}$$

$$lnType_{s,a}(i) = \beta_{0_{s,a}} + \beta_{1_{s,a}} N_{s,i} + \beta_{2_{s,a}} N_{c,i} \tag{10}$$

The coefficient β_0 corresponds to the intercept, while β_1 and β_2 correspond to the coefficients describing the effect of the number of cows ($N_{c,i}$) and caprids ($N_{s,i}$) in each household.

Finally, we estimate the expected number of *B. abortus* ($Y_{c,a}(i)$ and $Y_{s,a}(i)$) and *B. melitensis* ($Y_{c,m}(i)$ and $Y_{s,m}(i)$) infected individuals in each household, from a binomial distribution in which the chance of an individual being genetically positive for each *Brucella* species is raised by the chance of being seropositive in each the household (see Serology section below):

$$Y_{c,m}(i) \sim Bin(lnType_{c,m}(i) * lnSer_c, N_c(i)) \quad (11)$$
$$Y_{c,a}(i) \sim Bin(lnType_{c,a}(i) * lnSer_c, N_c(i)) \quad (12)$$
$$Y_{s,m}(i) \sim Bin(lnType_{s,m}(i) * lnSer_s, N_s(i)) \quad (13)$$
$$Y_{s,a}(i) \sim Bin(lnType_{s,a}(i) * lnSer_s, N_s(i)) \quad (14)$$

The estimated number of animals per household infected with *B. abortus* and *B. melitensis* are then used to estimate the probability of human infection with *Brucella* as in Equation (2).

Livestock & human infection with *Brucella*

In order to determine what can be inferred from serology alone and what additional understanding can be gained through integration of the genetic data both the animal and human probabilities of infection were also estimated using serology data only. The approach used was similar to that described for human infection using genetics data. For each host population, a Bernoulli process was used to describe the probability of being seropositive given a probability of missclassification, as in Equations (3) to (6) while the linear predictor of the logit transformed probability (e.g. $lnSer_c(i)$) of individual infection at the household level, that mirrors equation (2), was defined:

for cattle,
$$lnSer_c(i) = \theta_{0,c} + \theta_{1c} * N_c(i) + \theta_{2c} * N_s(i) \quad (15)$$

for caprids,
$$lnSer_s(i) = \theta_{0,s} + \theta_{1s} * N_s(i) + \theta_{2s} * N_c(i) \quad (16)$$

and for humans,
$$lnSer_h(i) = \theta_{0,h} + \theta_{1h} * Y_c(i) + \theta_{2h} * Y_s(i) \quad (17)$$

where the coefficient θ_{0c} and θ_{0s} correspond to the intercepts, θ_{1c}, θ_{2c}, θ_{1s}, θ_{2s} correspond to the effect of the number of cows ($N_c(i)$) and caprids ($N_s(i)$) on the probability of infection of cattle and caprids, respectively. The coefficents θ_{1h} and θ_{2h} correspond to the effect of the estimated number of *Brucella* infected cows ($Y_c(i)$) and caprids ($Y_s(i)$) in each household on the probability of human infection, estimated as:

$$Y_c(i) \sim Bin(lnSer_c(i), N_c(i)) \qquad and \qquad Y_s(i) \sim Bin(lnSer_s(i), N_s(i)) \quad (18)$$

The priors for these coefficients are defined in Table S2.

Table S2: Prior distributions for the coefficients used to model human infection risk with *B. abortus* and *B. melitensis*. The coefficient denomination is shown as presented in the model description (*Coef.*) with corresponding notation in the JAGS code (*Par. JAGS*).

Variable	**Coef.**	**Par. JAGS**	**Distribution**	**Prior**
Human genetics				
Intercept	β_{0h}	bhgen0	Normal	$\sim dnorm(0, 0.001)$
Cattle with abortus	β_{1h}	bha1	Normal	$\sim dnorm(0, 0.001)$
caprids with abortus	β_{2h}	bha2	Normal	$\sim dnorm(0, 0.001)$
Cattle with melitensis	β_{3h}	bhm1	Normal	$\sim dnorm(0, 0.001)$
caprids with melitensis	β_{4h}	bhm2	Normal	$\sim dnorm(0, 0.001)$
Animal genetics				
Intercept (*melitensis*)	$\beta_{0c,m}/\beta_{0s,m}$	bcm0/bsm0	Normal	$\sim dnorm(0, 0.001)$
Intercept (*abortus*)	$\beta_{0c,a}/\beta_{0,as}$	bca0/bsa0	Normal	$\sim dnorm(0, 0.001)$
No. cattle on cattle	$\beta_{1c,m}/\beta_{1c,a}$	bcm1/bca1	Normal	$\sim dnorm(0, 0.001)$
No. caprids on cattle	$\beta_{2c,m}/\beta_{2c,a}$	bcm2/bca2	Normal	$\sim dnorm(0, 0.001)$
No. caprids on caprids	$\beta_{1s,m}/\beta_{1s,a}$	bsm1/bsa1	Normal	$\sim dnorm(0, 0.001)$
No. cattle on caprids	$\beta_{2s,m}/\beta_{2s,a}$	bsm2/bsa2	Normal	$\sim dnorm(0, 0.001)$
Human serology				
Intercept	θ_{0h}	bh0	Normal	$\sim dnorm(0, 0.001)$
No. cattle on humans	θ_{1h}	bh1	Normal	$\sim dnorm(0, 0.001)$
No. caprids on humans	θ_{2h}	bh2	Normal	$\sim dnorm(0, 0.001)$
Animal serology				
Intercept	θ_{0c}/θ_{0s}	bc0/bs0	Normal	$\sim dnorm(0, 0.001)$
No. cattle on cattle	θ_{1c}	bc1	Normal	$\sim dnorm(0, 0.001)$
No. caprids on cattle	θ_{2c}	bc2	Normal	$\sim dnorm(0, 0.001)$
No. caprids on caprids	θ_{1s}	bs1	Normal	$\sim dnorm(0, 0.001)$
No. cattle on caprids	θ_{2s}	bs2	Normal	$\sim dnorm(0, 0.001)$
Sensitivity serology				
Correct detection	q^+	Fp_	Beta	$\sim dbeta(25, 0.5)$ or 1
False detection	q^-	Fn_	Beta	$\sim dbeta(0.5, 25)$ or 0

*Normal distributions are expressed in JAGS notation, i.e. in terms of mean and precision.

Model fit and diagnostics

All models were fitted using JAGS software which uses Gibbs sampling to generate pos-

terior distributions of the parameters given the likelihood, prior distributions and the data itself. The JAGS code used to run our model is given in the subsequent section. We ran our models for $3 \cdot 10^5$ iterations with burn-in of $1.5 \cdot 10^5$ to achieve convergence. Convergence was assessed by visual inspection of the chains and posterior distributions, as well as Gelman-Rubin diagnosis. We further evaluated model fit by comparing the posterior distributions with the data and, where appropriate, the true values used to generate the data. We further assess model fit and the impact of the uninformative priors by ensuring that we can recover the simulated coefficients, that are similar to those of the field data. In addition to showing that the model is capable of capturing the necessary dynamics, the recovery of the original simulated parameters also indicates that the magnitude of the priors is not influencing the ability of the model to converge to correct values.

JAGS code

```
model{
  for(i in 1:Nhh){
                                    #Serology
#Cattle
    for(j in 1:Ncows){
      y_cows[i,j] ~ dbern(pSer_cows[i,j])
      pSer_cows[i,j]=Fp_cows*r_cows[i,j]+(1-r_cows[i,j])*Fn_cows
      r_cows[i,j]=1-mean(lnSer_cows[i]) }
    logit(lnSer_cows[i])=bc0+bc1*Ncows_HH[i]+bc2*Ncaprids_HH[i]
    ystar_cows ~ dbin(lnSer_cows[i],Ncows_HH[i])

#caprids
    for(j in 1:Ncaprids){
      y_caprids[i,j] ~ dbern(pSer_caprids[i,j])
      pSer_caprids[i,j]=Fp_cows*r_caprids[i,j]+(1-r_caprids[i,j])*Fn_caprids
      r_caprids[i,j]=1-mean(lnSer_caprids[i]) }
    logit(lnSer_caprids[i])=bs0+bs1*Ncaprids_HH[i]+bs2*Ncows_HH[i]
    ystar_caprids ~ dbin(lnSer_caprids[i]*Ncaprids_HH[i])

#Humans
    for(j in 1:Nhumans){
      y_humans[i,j] ~ dbern(pSer_humans[i,j])
      pSer_humans[i,j]=Fp_humans*r_humans[i,j]+(1-r_humans[i,j])*Fn_humans
      r_humans[i,j]=1-mean(pr_humans[i]) }
    logit(lnSer_humans[i])=bh0+bh1*ystar_cows[i]+bh2*ystar_caprids[i]
```

#Genetics

#Cattle

ym_cows[i] ~ dbin(pType_m_cows[i], Ncows_typed[i])
logit(pType_m_cows[i])=bcm0+bcm1*Ncows_HH[i]+bcm2*Ncaprids_HH[i]
ystar_m_cows ~ dbin(pType_m_cows[i],ystar_cows[i])

ya_cows[i] ~ dbin(pType_a_cows[i], Ncows_typed[i])
logit(pType_a_cows[i])=bca0+bca1*Ncows_HH[i]+bca2*Ncaprids_HH[i]
ystar_a_cows ~ dbin(pType_a_cows[i],ystar_cows[i])

#caprids

ym_caprids[i] ~ dbin(pType_m_caprids[i], Ncaprids_typed[i])
logit(pType_m_caprids[i])=bsm0+bsm1*Ncaprids_HH[i]+bsm2*Ncows_HH[i]
ystar_m_caprids ~ dbin(pType_m_caprids[i],ystar_caprids[i])

ya_caprids[i] ~ dbin(pType_a_caprids[i], Ncaprids_typed[i])
logit(pType_a_caprids[i])=bsa0+bsa1*Ncaprids_HH[i]+bsa2*Ncows_HH[i]
ystar_a_caprids ~ dbin(pType_a_caprids[i],ystar_caprids[i])

#Humans

for(j in 1:Nhumans){
y_humans[i,j] ~ dbern(pType_humans[i,j])
pType_humans[i,j]=FpType_humans*rType_humans[i,j]+
(1-rType_humans[i,j])*FnType_humans
r_Type_humans[i,j]=1-mean(pType_humans[i]) }
logit(pType_humans[i])=bhType0+bha1*ystar_a_cows[i]+bha2*ystar_a_caprids[i]+
bhm1*ystar_m_cows[i]+bhm2*ystar_m_caprids[i]

}#end household loop

For prior specification please see table S2.

}#end model

S2. Simulations

The simulation of the data for this study was performed based on the modelling framework described above. Twelve distinct datasets were simulated, one for each of the combinations of the three epidemiological scenarios and the four population structures described in the main manuscript.

To simulate human infection status a binomial process was used to generate the number of *Brucella* positive humans at each household (yType_humans(i)) as a function of the total number of animals infected with each *Brucella* species at each household (e.g. ym_caprids[i], ya_caprids[i], ym_cows[i], and/or ya_cows[i]). The probability of a human in the i^{th} household being infected with *Brucella* (pType_humans(i)) was simulated as the binomial logistic proportion of the individuals that were tested (Nhumans[i]). Details of the R code used for these simulations are given below (See box *R code for simulations*).

The simulated datasets include binary indicators of infection status (infected or not) for all humans and animals in each population. For humans the (*Brucella* infection status was simulated, analogous to serological diagnostic test data. For animals the *B. melitensis* and *B. abortus* infection status was simulated for each individual. This animal infection status corresponds to genetic typing data. The serostatus of all animals positive for *B. melitensis* and/or *B. abortus* was defined as positive. All other animals were defined as negative.

The first part of this process that simulates *Brucella* positive humans is analogous to Equation 2 above except that the ystar values in Equation 2 are estimates of the actual Y values used in the simulations (e.g. ystar_m_caprids[i] is an estimate of the ym_caprids[i] value generated in the simulations). The number of infected humans at each household was simulated at the household level as described above and positive status was then randomly allocated to the simulated number of positive humans for each household to generate the individual based dataset.

To generate the Y values used above (for the number of animals infected with each *Brucella* species in each household), the infection status of animals (cattle and caprids) was

simulated independently for *B. melitensis* and *B. abortus*. Bernoulli processes were used to describe the probability that an individual animal was positive for a given *Brucella* species. The probability of infection in each case (e.g. pType_m_caprids[i,j], was defined through a logit transformed linear function of covariates that describe the *Brucella* species transmission processes in animals as a function of the number of cows and caprids present at the household (i). These processes are analogous to the model Equations 7 to 10 above.

The parameter values used in the simulations were specified for each of the distinct epidemiological scenarios as given in Table S3. Parameter values were selected and set by examining the relationships between key values (e.g. cattle seroprevalence and cattle population size) in the Tanzanian data set included in this study and with reference to other previously published studies. Values of baseline prevalence parameters were selected to ensure plausible final seroprevalence values for the simulated human, cattle and caprid populations(e.g. ranging between 1 and 16% for humans and 1 and 10% for cattle and caprids).

R code for simulations

```
  for (i in 1:Nhh){

                              #Human infections
   pType_humans[i]<- invlogit (bhType0 + bha1 * ya_cows[i] + bha2 * ya_caprids[i]
   + bhm1 * ym_cows[i] + bhm2 * ym_caprids[i])
   yType_humans[i]<- rbinom(1, Nhumans[i], pType_humans[i])
       }

                              #Animal infections
#caprid melitensis
for (j in 1:Ncaprids){
   pType_m_caprids[i,j]<- invlogit (bsm0 + bsm1 * Ncaprids_HH[i] + bsm2 * Ncows_HH)[i]
   ym_caprids[i,j]<- rbinom (1, 1,pType_m_caprids[i,j])
       }

#caprid abortus
for (j in 1:Ncaprids){
   pType_a_caprids[i,j]<- invlogit (bsa0 + bsa1 * Ncaprids_HH[i] + bsa2 * Ncows_HH)[i]
   ya_caprids[i,j]<- rbinom (1, 1, pType_a_caprids[i,j])
       }

#Cattle melitensis
for (j in 1:Ncows){
   pType_m_cows[i,j]<- invlogit (bcm0 + bcm1 * Ncows_HH[i] + bcm2 * Ncaprids_HH)[i]
   ym_cows[i,j]<- rbinom (1, 1, pType_m_caprids[i,j])
       }

#Cattle abortus
for (j in 1:Ncows){
   pType_a_cows[i,j]<- invlogit (bca0 + bca1 * Ncows_HH[i] + bca2 * Ncaprids_HH[i])
   ya_cows[i,j]<- rbinom (1, 1, pType_a_cows[i,j])
       }
```

Table S3: Parameter values used for the simulation of alternative epidemiological scenarios. The parameter notation is shown as presented in the model descriptions (*Par.*) with corresponding notation in the R code (*Par. R*).

Variable	**Par.**	**Par. R**	**Scenario 1**	**Scenario 2**	**Scenario 3**
Human genetics					
Baseline prevalence	-	bpType_human	0.01	0.02	0.01
Intercept	β_{0h}	bhType0	log(bpType_human/1-bpType_human)		
N cattle with abortus	β_{1h}	bha1	0.35	0.0000001	0.0000001
N caprids with abortus	β_{2h}	bha2	0.0000001	0.0000001	0.0000001
N cattle with melitensis	β_{3h}	bhm1	0.0000001	0.0000001	0.35
N caprids with melitensis	β_{4h}	bhm2	0.7	0.7	0.7
Abortus in Cattle					
Baseline prevalence	-	bpType_a_cows	0.02	0.0000001	0.02
Intercept	β_{0ca}	bca0	log(bpType_a_cows/(1-bpType_a_cows))		
N cattle in herd	$\beta_{1c,a}$	bca1	0.08	0.0000001	0.0000001
N caprids in flock	$\beta_{2c,a}$	bca2	0.0000001	0.0000001	0.0000001
Abortus in caprids					
Baseline prevalence	-	bpType_a_caprids	0.0000001	0.0000001	0.001
Intercept	β_{0sa}	bsa0	log(bpType_a_caprids/(1-bpType_a_caprids))		
N caprids in flock	$\beta_{1s,a}$	bsa1	0.0000001	0.0000001	0.0000001
N cattle in herd	$\beta_{2s,a}$	bsa2	0.0000001	0.0000001	0.0000001
Melitensis in Cattle					
Baseline prevalence	-	bpType_m_cows	0.0000001	0.02	0.01
Intercept	β_{0cm}	bcm0	log(bpType_m_cows/(1-bpType_m_cows))		
N cattle in herd	$\beta_{1c,m}$	bcm1	0.0000001	0.0000001	0.0000001
N caprids in flock	$\beta_{2c,m}$	bcm2	0.0000001	0.002	0.002
Melitensis in caprids					
Baseline prevalence	-	bpr_m_caprids	0.02	0.02	0.02
Intercept	β_{0sm}	bsm0	log(bpType_m_caprids/(1-bpType_m_caprids))		
N caprids in flock	$\beta_{1s,m}$	bsm1	0.008	0.008	0.008
N cattle in herd	$\beta_{2s,m}$	bsm2	0.0000001	0.0000001	0.0000001