

SOFTWARE

Open Access

# |tPRiors|: a tool for prior elicitation and obtaining posterior distributions of true disease prevalence



Konstantinos Pateras\*  and Polychronis Kostoulas

## Abstract

**Background:** Tests have false positive or false negative results, which, if not properly accounted for, may provide misleading apparent prevalence estimates based on the observed rate of positive tests and not the true disease prevalence estimates. Methods to estimate the true prevalence of disease, adjusting for the sensitivity and the specificity of the diagnostic tests are available and can be applied, though, such procedures can be cumbersome to researchers with or without a solid statistical background. This manuscript introduces a web-based application that integrates statistical methods for Bayesian inference of true disease prevalence based on prior elicitation for the accuracy of the diagnostic tests. This tool allows practitioners to simultaneously analyse and visualize results while using interactive sliders and output prior/posterior plots.

**Methods - implementation:** Three methods for prevalence prior elicitation and four core families of Bayesian methods have been combined and incorporated in this web tool. |tPRiors| user interface has been developed with R and Shiny and may be freely accessed on-line.

**Results:** |tPRiors| allows researchers to use preloaded data or upload their own datasets and perform analysis on either single or multiple population groups clusters, allowing, if needed, for excess zero prevalence. The final report is exported in raw parts either as .rdata or .png files and can be further analysed. We utilize a real multiple-population and a toy single-population dataset to demonstrate the robustness and capabilities of |tPRiors|.

**Conclusions:** We expect |tPRiors| to be helpful for researchers interested in true disease prevalence estimation and who are keen on accounting for prior information. |tPRiors| acts both as a statistical tool and a simplified step-by-step statistical framework that facilitates the use of complex Bayesian methods. The application of |tPRiors| is expected to aid standardization of practices in the field of Bayesian modelling on subject and multiple group-based true prevalence estimation.

**Keywords:** Prevalence estimation, Prior elicitation, Bayesian, Pooled samples, Statistical modelling, JAGS, Shiny

\*Correspondence: [kostas.pateras@gmail.com](mailto:kostas.pateras@gmail.com)

Laboratory of Epidemiology & Artificial Intelligence, Faculty of Public and One Health, University of Thessaly, Karditsa, Greece



© The Author(s). 2022 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Apparent prevalence represents the observed rate of positive test results. Tests can give false positive and false negative results, thus apparent prevalence estimates are biased estimates of the true disease prevalence, while the extend of this bias depends on the overall misclassification rate of testing. Methods are available and can be applied to estimate the true prevalence of disease and, thus, adjust for the imperfect sensitivity and specificity of the diagnostic tests, either within a frequentist or a Bayesian framework [1, 2]. The primary advantage of the latter framework lies in its ability to incorporate in the estimation process existing and relevant knowledge for the parameters of interest in the form of priors [3]. Nonetheless, the application of such approaches can be cumbersome for researchers with or without a statistical background, which results in limited application of existing Bayesian methods for calculation of true prevalence by non-experts. A statistical program that encompasses most of these prevalence models would boost the application of such methods and aid practitioners to perform basic and advanced Bayesian prevalence estimation.

In this manuscript we present Bayesian models for the estimation of apparent and true disease prevalence for a single and multiple populations (“Implementation section”). Then, we introduce a new free web-based application, *|tPRiors|*, that facilitates the use of Bayesian prevalence methods by non-experts (“The *|tPRiors|* web-based application” section), which has been developed with R and Shiny [4]. Finally, in “Results” section, we illustrate the use of *|tPRiors|* using a single toy population example [5], and a real multiple population example that is based on a systematic review and meta-analysis on the prevalence of Dementia in Europe [6]. The manuscript ends with a Discussion.

## Implementation

### Methods for prevalence estimation

#### Apparent prevalence estimation

Let us assume that in a single population of size  $N$  (i.e. nation, group of patients, animal herd or any cluster)  $y$  subjects test positive out of the  $n$  randomly sampled population subjects. Then  $y$  follows, approximately, a binomial distribution:

$$y|\pi_a \sim \text{Binomial}(n, \pi_a) \tag{1}$$

$$\pi_a \sim \text{Beta}(\alpha_{\pi_a}, \beta_{\pi_a})$$

where  $\pi_a$  is the apparent disease prevalence and  $\alpha_{\pi_a}, \beta_{\pi_a}$  are the parameters of the Beta prior distribution for  $\pi_a$ . The posterior distribution of the above model can be analytically calculated from  $\text{Beta}(\alpha_{\pi_a} + y, \beta_{\pi_a} - y + n)$  or through the use of Bayesian statistical programs (WinBugs, Jags, Stan) [7–9]. In the absence of historical information, parameters  $\alpha_{\pi_a}, \beta_{\pi_a}$  can be set close or equal to 1,

which is similar to assuming a Uniform(0,1) prior on  $\pi_a$ . Relevant existing knowledge for  $\pi_a$  can be incorporated by specifying values for  $\alpha_{\pi_a}, \beta_{\pi_a}$  that leads to a prior Beta distribution that captures such knowledge. *|tPRiors|* can be used to specify these values (see “The *|tPRiors|* web-based application” section).

#### True prevalence estimation

The relationship between apparent ( $\pi_a$ ) and true disease prevalence ( $\pi_t$ ) is known and given by  $\pi_a = \pi_t S_e + (1 - \pi_t)(1 - S_p)$  [10], where  $S_e, S_p$  denote the sensitivity and specificity of the diagnostic test. Hence, replacing  $\pi_a$  with the above in Eq. 1,

$$y|\pi_t, S_e, S_p \sim \text{Binomial}(n, \pi_t \cdot S_e + (1 - \pi_t) \cdot (1 - S_p))$$

$$\pi_t \sim \text{Beta}(\alpha_{\pi_t}, \beta_{\pi_t}) \tag{2}$$

$$S_e \sim \text{Beta}(\alpha_{S_e}, \beta_{S_e}), S_p \sim \text{Beta}(\alpha_{S_p}, \beta_{S_p})$$

where  $\alpha_{\pi_t}, \beta_{\pi_t}, \alpha_{S_e}, \beta_{S_e}, \alpha_{S_p}, \beta_{S_p}$  are the parameters of the Beta prior distributions for  $\pi_t, S_e$  and  $S_p$ , respectively [2].

#### Multiple populations

Let us suppose that  $n_i, i = 1, 2, \dots, k$  subjects are sampled randomly from multiple clusters ( $k$ ). When multiple clusters are available, commonly a two-stage cluster-design is implemented. In the first stage  $k$  clusters are selected randomly within a region and then  $n_i$  subjects are sampled randomly from each cluster ( $i$ ). In the end,  $y_i$  positive tests will be collected from each cluster. As in the single cluster case above, the number of tests, are assumed to follow a binomial distribution [1],

$$y_i|\pi_{ti}, S_e, S_p \sim \text{Binomial}(n_i, \pi_{ti} S_e + (1 - \pi_{ti})(1 - S_p)) \tag{3}$$

$$\pi_{ti}|\mu, \psi \sim \text{Beta}(\mu\psi, \psi(1 - \mu))$$

$$S_e \sim \text{Beta}(\alpha_{S_e}, \beta_{S_e}), S_p \sim \text{Beta}(\alpha_{S_p}, \beta_{S_p})$$

$$\mu \sim \text{Beta}(\alpha_{\mu}, \beta_{\mu}), \psi \sim \text{Gamma}(\alpha_{\psi}, \beta_{\psi}) \tag{4}$$

$S_e$  and  $S_p$  are modelled through beta prior distributions and are assumed invariant between clusters. The sampled group prevalences are considered exchangeable as  $\pi_{ti}|\mu, \psi \sim \text{Beta}(\mu\psi, \psi(1 - \mu))$ , where  $\mu$  denotes the average population prevalence and  $\psi$  is associated with the heterogeneity (between-group variance) of cluster prevalences around their mean.  $\mu, \psi$  are modelled through a Beta and Gamma prior distributions respectively. Large values of  $\psi$  correspond to lower between-study variability (heterogeneity). On all hyper-parameters (i.e.  $\alpha_{\mu}, \beta_{\mu}, \alpha_{\psi}, \beta_{\psi}$ ) a hyper-prior distribution can be assumed, also, all hyper parameters can be elicited based on expert opinion. A logical strategy, though, would be to reduce the number of unknown parameters, given that the data collection procedure relies solely on a basic binomial count [2].

### Allowing for zero true prevalence

All models above assume a flexible beta prior distribution on  $\pi_t$ . This implies that  $\pi_t \neq 0$ , an assumption that may not be realistic on low prevalence populations or on populations with specific free from disease clusters. To account for the probability of zero disease prevalence, a mixture distribution can be applied such that  $\pi_t \sim \text{Beta}(\alpha_{\pi_t}, \beta_{\pi_t})$  for a single population or  $\pi_{ti}|\mu, \psi \sim \text{Beta}(\mu\psi, \psi(1-\mu))$  for multiple populations with probability  $w$ , and  $\pi = 0$  with probability  $1-w$ , where  $w$  is the probability that the group is actually infected [2].

$$\pi^* = \begin{cases} \pi_t & \text{with probability } w \\ 0 & \text{with probability } 1-w \end{cases}$$

In a similar sense, a beta prior distribution can be placed on  $w$ , with  $\alpha_w, \beta_w = 1$  denoting prior ignorance. It should be noted that omitting to place a mixture distribution on  $\pi_a$  may result in biases and possibly narrower credible intervals. Following the multiple population modelling,  $w$  can be further modelled similarly to  $\pi_{ti}$  to allow for zero prevalence in the whole region [11].

The multinomial model allows for estimation of (a) the group-level prevalence, (b) the prevalence distribution for groups, (c) the mean of the prevalence distribution, (d) the probability that the whole (sub) population is free of disease (at a pre-specified level) and finally (e) the predicted probability that a group is free of disease (again at a pre-specified level).

### The |tPRiors| web-based application

|tPRiors| is an open-access web-based application with a graphical, user-friendly interface which aids the estimation of true disease prevalence from single and multiple population data using the models presented in "Implementation" section. Non-expert users on Bayesian analysis and relevant software, such as R/JAGS, can easily elicit priors, perform Bayesian statistical analysis, visualize results and download the corresponding reports.

One of the mechanisms of the |tPRiors| application is the "PriorGen" R package (Kostoulas, 2018), which powers the prior elicitation by translating prior beliefs into usable prior information in the form of hyper-parameters for the Beta (and Gamma) distributions. This package can be utilized for the generation of priors on the true prevalence of disease, the sensitivity and the specificity of diagnostic tests and the probability of zero true prevalence.

|tPRiors| is powered by Shiny, thus, it has been deployed on shinyapps.io that allows for easy user access without the need for installation of R or JAGS. The web application can be accessed from any device that has internet connection and a web browser, even though more optimal use is warranted via a computer or a laptop. |tPRiors| is currently freely accessible to download and use in R/Rstudio via <https://github.com/kpatera/>

or it can be directly accessed via <https://publicandonehealth.shinyapps.io/tPRiors/>. The application can be accessed through different operating systems and internet browsers, though, we recommend the use of an updated Google Chrome or Firefox browser.

### |tPRiors| step by step walkthrough

The initial "Start" page and the other menus of |tPRiors| are presented in Fig. 1. |tPRiors| consists of six individual pages/tabs: 1. Start, 2. Set-up, 3. Prior(s), 4. Input & Output, 5. Report and 6. Acks. Three of these pages, "Set up", "Prior(s)" and "Input & Output", contain the core mechanisms. On each page the user has to press the "Fix/Set" button, after choosing the preferred options and before clicking on the next |tPRiors| page. Frequently asked questions (FAQ) can be found in the 'Acks' page. A description for each page follows.

#### Start

The first page provides a general overview of |tPRiors| and a brief description on how the user should proceed.

#### Set up

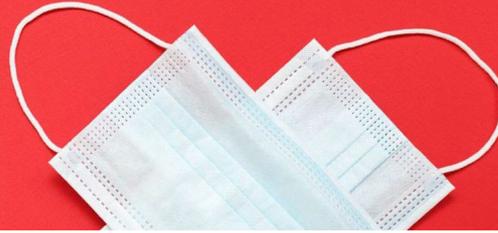
In this page the users select the modelling framework that best captures the structure of their data/population. The users first specify whether they will analyse data from one or multiple populations/clusters. Then, they define whether they wish to estimate true or apparent prevalence. Subsequently, they need to decide whether they wish to allow for the possibility of zero prevalence for some of the populations/clusters. Finally, they indicate whether there is a priori information for the prevalence (apparent or true) that they wish to use and then they decide which measure of central tendency (mean, median or mode) or dispersion (percentiles) they will employ for this prior specification on prevalence. The multiple population models option currently supports one prior specification measure (the mean).

#### Prior(s)

In this page the users should specify the available information that will be used for generation of the prior distribution for apparent or true prevalence. If, in the previous page, they have declared that they are estimating true prevalence then two additional sub-pages become available for the sensitivity and specificity of the diagnostic test. Clicking on the 'Help' button will provide an example on how to elicit a prior based on the available questions. We urge users of |tPRiors| to first answer all questions, input values on all available parameters, and only then press the 'Set priors' button, in this way small subjectivity would be inserted in the elicitation process. Prior specification is a guided-through process, though, further guidance can be found when the 'Help' button is clicked or in the FAQ section of the 'Acks' page. Fur-

## Bayesian true prevalence inference via elicited priors

# | tPRiors |



Below the user can find a brief description of the shiny application functions and options.

- (a) In tab (Set up) the user following questions can fix the parameters of the analysis (Choose model, priors, special characteristics)
- (b) In tab (Prior(s)) the user can elicitate the prior distribution(s) with the aid of sliders and visual confirmation
- (c) In tab (Input & Output) the user inputs the observed data and Jags sampling characteristics. A basic inference plot is presented. For multiple population both step 1 and step 2 should be selected before you proceed.
- (d) In tab (Report) the program returns a dynamic output that changes based on (a), (b) and (c).
- (e) In tab (Acks) acknowledgments and Frequently Asked Questions (FAQ) can be found.

Settings may still be changed even after a tab has been fixed by the user. Though, we advise users to perform a 'Reset' of |tPRiors| when they want to change a previously fixed setting. We also advice users to perform a 'Refresh' of the web page tPRiors when they perform a different analysis

A template file of a multiple population example can be found here ([https://1drv.ms/x/s!Ag8vFzeTRzBthzImmQV\\_nEDdj4Uy?e=BjwmQL](https://1drv.ms/x/s!Ag8vFzeTRzBthzImmQV_nEDdj4Uy?e=BjwmQL)). It is essential that at least two columns exist in the spreadsheet one named 'positive' and one named 'n'

The development of tPRiors was funded by H2020 project unCoVer: Unravelling Data for Rapid Evidence-Based Response. More details can be found in the manuscript, K Pateras and P Kostoulas, tPRiors: A tool for prior elicitation and obtaining posterior distributions of true disease prevalence

**Fig. 1** Initial screen and menus of the |tPRiors| web-based application. This screen-shot showcases the application and it is not meant for direct reading. For a dynamic web-page one should visit the link <https://publicandonehealth.shinyapps.io/tPRiors/>

thermore, additional information can be found at the PriorGen manual [12].

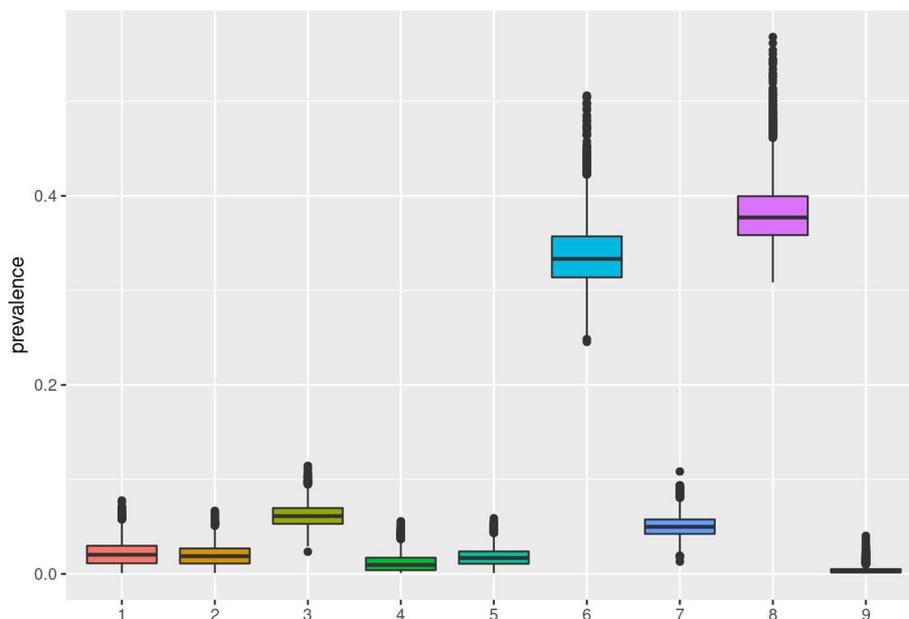
### I & O (Input & Output)

Based on their selection in the "Prior(s)" page, in this page users should input and model either single or multiple population data. For single population data, users should specify via a number of sliders, the sample size, number of events and Markov Chain Monte Carlo (MCMC) - related options. For multiple-population data, users should upload a dataset (.csv/.xls). Each row of the data table should correspond to a different site/study and the dataset should contain at least 2 columns. One column named "*positive*" that contains the observed number of events per site/population and one column named "*n*" that contains the sample sizes per site/population. When all parameters are set the user should then press the 'Fix model' button and only then click on the "Report" page. Caution! Both Step 1. Data and Step 2. Output tabs have to be clicked to fully run the multiple-population models.

### Report

After navigating through the above pages, users find a dynamic report in the fifth page. This report contains all previous selected options, information on the selected prior distributions as well as posterior prevalence inferences and MCMC diagnostics to check convergence. All figures can be directly downloaded. In the case of multiple populations, box plots are provided for each study in a dynamic graph (Fig. 2), together with information that aims to aid researchers in planning an upcoming prevalence study.

Near the end of the report a series of diagnostic plots are presented. More specifically, barplots and density plots of the main parameters are available at the top of the section, alongside traceplots and running mean plots, followed by complete versus partial chain plots and autocorrelation plots of the main parameters. When the MCMC analysis runs into issues users will identify non-common patterns in these graphs. Mostly diagnostic plots point towards convergence issues. The simple density plots can show abnormalities in the posterior densities, if any density



**Fig. 2** Box plots with estimates for each one of the nine studies presented in Table 2 - *|tPRiors|* output

look different than expected the user should check the data and prior inputs. The multiple chain or partial chain density plots should contain densities of different chains that mostly coincide, otherwise posterior Bayesian convergence is impacted by each MCMC run and the results may not be trusted. MCMC traceplots should depict multiple chains that often cross in order to validate convergence while, the ergodic mean plots and the correlation plots should move very quickly towards the mean and zero respectively, otherwise the user should increase the thinning interval and/or the number of MCMC samples. We refer interested readers and practitioners to the CODA and GSS R packages and we suggest that, when in doubt, they should always consult with a statistician.

Finally, *|tPRiors|* provides further access to the analysis by making available three files at the bottom of the "Report" page and analytical details on how to re-analyse them. These files are available under the 1) Input, 2) Model and 3) Output tabs. Interested users have the option to utilize Rstudio or R to work with the specific rData file of interest. By utilizing these files, i.e. study-level plots can be re-produced to identify and diagnose further methodological issues or change the appearance and labels of the figures. Detailed steps on how to reproduce diagnostic plots can be found at the end of the "report" page, right before the download data tabs. Briefly, the MCMC samples can be downloaded from the "Report" page via the 'Output' tab. Within the downloaded object, multiple MCMC chains are stored for each parameter of interest. Following the detailed steps on the "Report"

page, practitioners can repetition (mentioned right above) attempt further analyses.

#### Acks

In the final page/tab an overview of the dependencies on R packages can be found. *|tPRiors|* web application was built using a number of R packages such as: shiny, shinydashboard, rmarkdown, R2jags, rootSolve, reshape, ggplot2, dplyr, DT and plotly among others. The different colours represent the four main categories of use, i.e. the orange colour denotes main packages, the green colour denotes shiny-based packages, blue and variations denote packages that aid plotting/reporting and the red colour denotes packages for data manipulation.

#### Results

The following section presents two, a multiple and a single population, examples of true prevalence estimation while taking into account the sensitivity and specificity of the applied (imperfect) diagnostic test.

#### Multiple population illustrative example

The multinomial approach estimates true prevalences that can directly be compared among the multiple study sites (i.e. studies, regions, cities, groups) as they are adjusted for imperfect diagnostic testing.

We present the systematic review on prevalence from Bacigalupo et al. [6]. The authors conducted a systematic review and meta-analysis on dementia and Alzheimer's disease. They selected studies that diagnosed patients

**Table 1** Studies included in the systematic review of Bacigalupo et al. 2018 alongside their apparent prevalence

ID	Study	Country	Year(s)	Positive ( $y_i$ )	N ( $n_i$ )	App. Prev ( $\pi_\alpha$ )	95% CIs
1	Ravaglia et al.	Italy	1999	60	961	0.062	(0.047,0.078)
2	Tognoni et al.	Italy	2000	100	1662	0.060	(0.049,0.072)
3	Gascon-Bayarri et al.	Spain	2002	165	1754	0.094	(0.080,0.108)
4	Fish et al.	UK	2003	88	1754	0.053	(0.042,0.064)
5	Bermejo-Pareja et al.	Spain	1994-5, 1997-8	306	5278	0.058	(0.052,0.064)
6	Mathillas et al.	Sweden	2000-2, 2005-7	287	895	0.321	(0.290,0.351)
7	Tola-Arribas et al.	Spain	2009-10	184	2170	0.085	(0.073,0.097)
8	Lucca et al.	Italy	2009-10	894	2501	0.357	(0.339,0.376)
9	Perquin et al.	Luxembourg	2008	53	1377	0.038	(0.028,0.049)

according to the DSM (Diagnostic and Statistical Manual of Mental Disorders) IV criteria but also studies that had a high quality score according to Alzheimer's Disease International study quality score criteria [13]. The meta-analysis included 9 studies conducted in Europe between 1993 and 2018 with patients aged >65 years. The total number of participants were 18263, out of which 2137 were diagnosed with dementia based on DSM IV criteria. The meta-analysed pooled prevalence rate estimate was reported as equal to 12.4% (95% CI 7.6%–17.2%). The raw data are summarized in Table 1.

A number of subjects have been tested ( $n_i$ ) per study/country, out of which ( $y_i$ ) subjects showed positive test results. The test procedure was based on DSM IV criteria for all participants. In comparison with previously used diagnostic procedures, the DSM IV criteria result in higher specificity. On average, based on historical data and previous studies, sensitivity may be set equal to 95% and specificity can be set equal to 88% [14–16], though if comorbidities or mild dementia are present, specificity may drop even further [17].

On *|tPRiors|*, the user selects to model 1) true prevalence, 2) multiple populations, 3) non-zero true prevalence, 4) informative priors<sup>1</sup> and then clicks fix. Consequently, in the “Prior(s)” page and given the above information, the researcher specifies priors on the disease prevalence and the test specificity and sensitivity. Via the use of simple sliders in the tab “Priors”, a low-informative prior is elicited for the true prevalence, as a  $\pi_t \sim \text{Beta}(0.14, 0.55)$ . The *|tPRiors|* shiny application input for this prevalence prior would be:  $\pi_t(0.2, 0.99, 0.3, 0.986)$ . This prior has a median equal to 1.6% and IQR (Interquartile range) equal to (0.06%, 39.03%) denoting our limited knowledge on the true prevalence for dementia in Europe. Similarly, for the specificity parameter  $S_p \sim \text{Beta}(25, 3.5)$  is elicited, while for the sensitivity parameter  $S_e \sim \text{Beta}(25.6, 1.4)$

is elicited based on the user inputs on the *|tPRiors|* “Prior(s)” page. The *|tPRiors|* shiny application input for the sensitivity and specificity priors would be:  $S_e(0.88, 0.80, 0.90) - S_p(0.95, 0.85, 0.97)$ . The sensitivity prior implies a prior median value of 0.88 (IQR: 0.85, 0.92) and the specificity prior implies a prior median value of 0.95 (IQR: 0.93, 0.98).

#### Analysis and report

Regarding our multiple-population motivating example, after running the analysis with the defined priors, settings and uploaded datasets, the following statistics are reported by the *|tPRiors|* application. The reported probability that the posterior true prevalence would be higher than 0.05 was equal to 1. The prevalence posterior mean equals to 18.5% which is larger than the original pooled estimate of the Bacigalupo et al. review and may imply less modelled heterogeneity and thus larger weights on Mathillas et al (6) and Lucca et al (9) studies. In comparison with the simple apparent prevalence estimates, the lower studies' prevalence shrunk towards zero, while study 6 and 9 showed an increase in prevalence when accounting for imperfect testing, a behaviour that is partially explained by the test's low specificity. In this example, we did not model the probability of zero prevalence as we assumed that the dementia prevalence would not be zero in the age groups studied.

#### Single population illustrative example

In this section we re-analyse a single population toy example to reproduce a previously published analysis from Spreybroek et al. [5]. In their example, for simplicity the authors have assumed a sample size of 500 with two assumed apparent prevalences, one equal to 24% that corresponds to a number of positives equal to 120 and one equal to 5% that corresponds to a number of positives equal to 25. Spreybroek et al. considered two Bayesian models. They have assumed either 1) a fixed value for the sensitivity ( $S_e = 0.80$ ) and specificity parameter  $S_p = 0.90$ ), or 2) a uniform distribution for the two parameters

<sup>1</sup>Do you have single or multiple populations/clusters? Multiple, Do you want to model the true or the apparent prevalence? True prevalence, Do you want to account for zero true prevalence? No, Would you specify informative priors? Yes

$S_e \sim Unif(0.7, 0.9)$  and  $S_p \sim Unif(0.85, 0.95)$ . We partially adopt their assumptions by setting Beta distributions on prior parameters with mean test sensitivity equal to 80% and mean test specificity equal to 90%, while avoiding the use of fixed values (Table 2). As prior probability for the true prevalence we elicit a)  $\pi_t \sim Beta(1, 1)$ , denoting Setup A and a b)  $\pi_t \sim Beta(1.3, 1.3)$ , denoting Setup B. Both setups express our belief that, before observing the data all values are likely or almost equally likely possible for the true prevalence. In setup A we assume that the mean value of the true prevalence equals 0.5 and it is uniformly distributed around this value. The  $\pi_t \sim Beta(1.3, 1.3)$ , prior distribution on the true prevalence of Setup B has a mean value of 0.5 (95% CrI 0.28,0.72). The *|tPRiors|* application shiny input for setup A would be:  $[\pi_t(0.5, 0.05, 0.95) - S_e(0.8, 0.75, 0.97) - S_p(0.9, 0.85, 0.97)]$ . Setup B places slightly smaller weight in prior true prevalence values close to 0/1, a plausible assumption as we expect that the prevalence would neither be very small nor very large. The *|tPRiors|* application shiny input for setup B would be:  $[\pi_t(0.5, 0.08, 0.95) - S_e(0.8, 0.75, 0.97) - S_p(0.9, 0.85, 0.97)]$ . For the sensitivity and specificity parameters in all scenarios we assume a mean value of 0.8 and 0.9 with a 97% probability that the mean values are higher than 0.75 and 0.85 respectively. Therefore, both Setup A and Setup B place the same prior on the sensitivity (0.8, CrI: 0.78, 0.82) and the same prior on the specificity (0.9, CrI: 0.88, 0.92). For both setup A and setup B, the user is able to impose a non-zero prevalence prior, if the prior probability of this prior is low, posterior inference would set additional density weight on zero. Let us assume that when non-zero true prevalence is selected, the *|tPRiors|* application shiny input for the non-zero true prevalence alternative could be:  $[w(0.2, 0.1, 0.95)]$ , meaning that based on the available literature the mean value for non-zero

prevalence probability is expected to be equal to 0.2 and we can be 95% sure that it is higher than 0.1.

**Analysis and report**

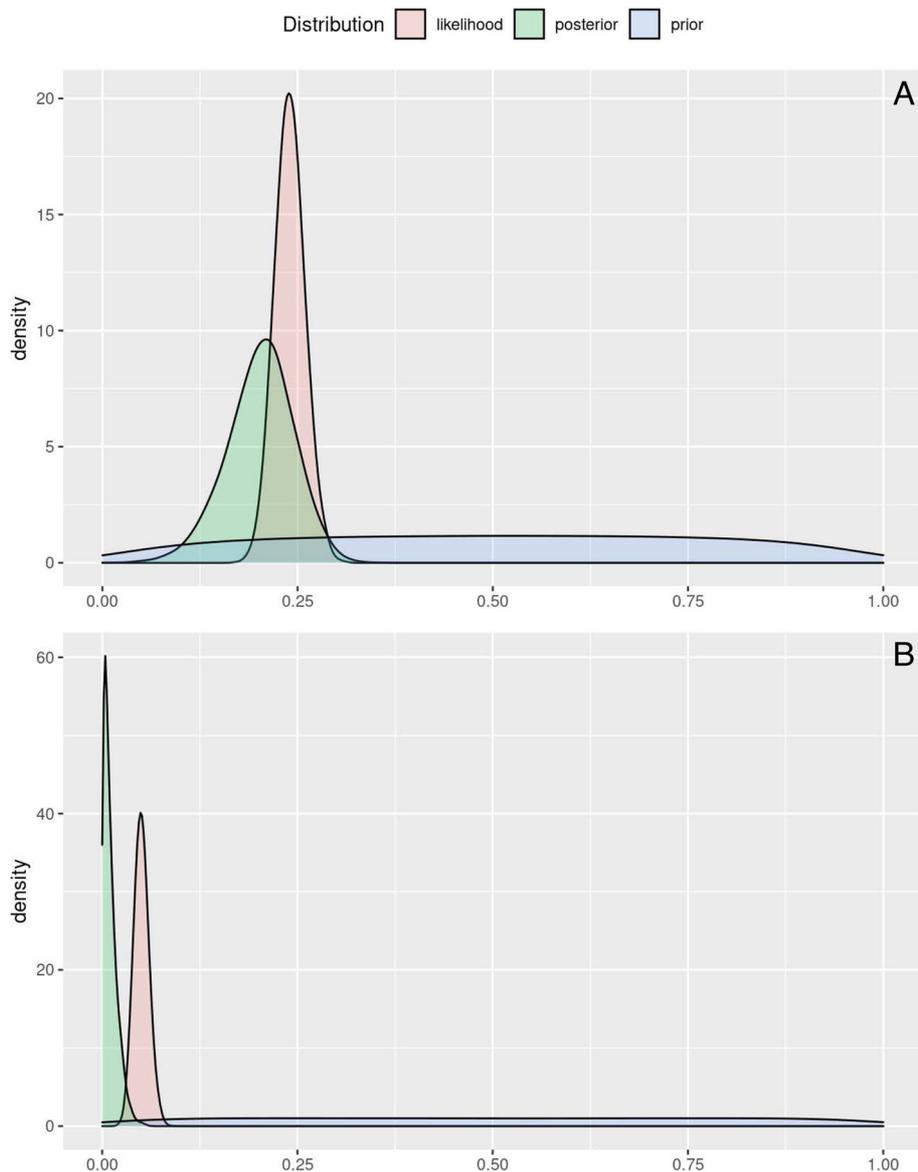
As depicted in Fig. 3(A-B), the posterior distribution of true prevalence points toward a lower prevalence than the apparent prevalence. Table 2 presents the toy example results of the original study [5] and the results produced by the *|tPRiors|* web-based application. The posterior mean of the original manuscript and the web-based application are very similar. Some discrepancies are observed on the width of the 95% credible intervals. More specifically, the *|tPRiors|* Setup A results in intervals (e.g. 24%: CrI (0.13, 0.27)) that lie between the two intervals produced by the original manuscript model 1 (24%: CrI (0.15, 0.25)) and model 2 (24%: CrI (0.11, 0.29)) Table 2]. This result is sensible as our prior assumptions are more relaxed than model 1 but more concentrated than the original model 2 prior distributions in Speybroeck et al. [5]. Finally, regarding the zero-prevalence prior alternative, Setup A, when assuming that the apparent prevalence equals to 5%, the analysis results in posterior mean true prevalence equal to 0.010 (95% CrI: 0.001, 0.032), while when assuming that the apparent prevalence equals to 24%, the analysis results in posterior mean true prevalence equal to 0.1988 (95% CrI: 0.1120, 0.2745). These values should be interpreted with caution, as the resulting posterior distribution is expected to have two peaks as expected and as it can be seen in the *|tPRiors|* 'Report' page.

**Discussion**

In this manuscript, we introduced *|tPRiors|*, a newly implemented web-based application for the calculation of prevalence based on prior information. During recent years such web-based applications have gained considerable popularity, as they are able to streamline and replicate otherwise complex analyses. Several similar applications are available in different areas of medical research and epidemiology as online tools or as statistical packages such as: i) IWA, an application that implemented full hierarchical model for prevalence estimation [11], ii) IPDmada an application that performs individual patient data meta-analysis [18], iii) covid19-tracker, an application that visualizes data from 2019-2020 covid pandemic [19] or iv) shinyCircos for visualizing genomic data [20]. To our knowledge, *|tPRiors|* is the first application that incorporates elicited prior knowledge, transform the latter into prior distributions and calculated posterior inferences for a variety of single and multiple population prevalence models. *|tPRiors|* makes the development and implementation of prevalence Bayesian inferences less cumbersome for researchers with or without strong statistical knowledge. Besides the basic Bayesian inference provided in *tPRiors*, an important function of such a web-based application is the interactive setting of the prior distributions

**Table 2** True prevalence results (mean and quantile of the 95% credible interval (Q) of the toy example with 24% and 5% apparent prevalence, 500 sample size, 120/25 positive tests with consistent assumptions on sensitivity (80%) and specificity (90%). Speybroek et al. 2012 Model 1 assumes fixed values for  $S_e$  and  $S_p$ , while Model 2 assumes  $S_e \sim Unif(0.7, 0.9)$  and  $S_p \sim Unif(0.85, 0.95)$ . For the *|tPRiors|* example,  $S_e \sim Beta(195.76, 48.94)$  and  $S_p \sim Beta(135.86, 15.1)$ , while for Setup A:  $\pi_t \sim Beta(1, 1)$  and Setup B:  $\pi_t \sim Beta(1.3, 1.3)$

Measure	Speybroek et al. 2012				tPRiors			
	Model 1		Model 2		Setup A.		Setup B.	
	24%	5%	24%	5%	24%	5%	24%	5%
Mean	0.2015	0.0048	0.2008	0.0096	0.2004	0.0103	0.2026	0.0127
2.5% Q	0.1493	0.0001	0.1133	0.0003	0.1266	0.0003	0.1203	0.0008
97.5% Q	0.2577	0.0173	0.2927	0.0336	0.2705	0.0327	0.2764	0.0367



**Fig. 3** Prior, posterior probability densities alongside the likelihood for Setup B with i. apparent prevalence: 24% or ii. apparent prevalence: 5% of the single population toy example - *tPRiors* output

on the relevant parameters, which can aid researchers towards prior elicitation in a natural and visual way.

Currently, four core models are incorporated in the web-application and “infinite” variations of these basic models can be constructed given input from the user, though, more models may be applied in practice. *tPRiors* has been developed as a framework that could encompass more models in the near future, either through a further systematic search or through alternative approaches researched from within our group, in an updated version of this web-application. No comparisons between approaches were performed, hence, interested readers are prompted to read the manuscripts that introduced

or critically discussed these models [1, 2, 5]. To aid researchers, we provide templates/preloaded multiple population datasets, as multiple population data should be uploaded in a specific format. No specific analyses for missing outcome data is currently incorporated, Bayesian inference, however, lends itself to accounting for missing outcome data via proper assumptions. Such considerations were out of scope but can be readily incorporated in the current models. Finally, *tPRiors* implements modelling with and without assuming a mixture distributions that allow for zero true infection prevalences. We should note that, especially in populations with low or zero prevalence, even though simpler models may be more intuitive,

if mixture models for prevalences are not applied, posterior inferences might be artificially inflated. In practice, we should note that the multiple population models perform more robustly when more than eight clusters are considered.

## Conclusions

The application of *tPRiors* is expected to aid standardization of practices in the field of Bayesian modelling on subject and multiple group-based true prevalence estimation. We further believe that *tPRiors* will help towards the increase of popularity of the aforementioned methods among practitioners and researchers in the public health sector. Although, *tPRiors* aims to aid researchers to perform Bayesian analysis with ease, we recommend at least one statistician to be part of the research team when such analyses are formally conducted.

## Highlights

- 1. *tPRiors* is a new web-based tool for conducting Bayesian inference of cross-sectional sampling via the use of single or multiple population diagnostic tests results.
- 2. *tPRiors* is a free tool, available for all researchers/practitioners and it does not require advanced knowledge or installation of any statistical software (i.e. R, SAS, SPSS).
- 3. *tPRiors* is designed to guide the user to the final analysis via the use of simple questions, leading to a printable final report.
- 4. Development of *tPRiors* will increase the quality of calculation and reporting of (true) prevalence studies.

## Availability and requirements

**Project name:** *tPRiors*

**Project home page:** <https://publicandonehealth.shinyapps.io/tPRiors/> and <https://github.com/kpatera/tPRiors>, publicly released [21]

**Operating system(s):** Platform independent

**Programming language:** R / Shiny

**Other requirements:** Firefox >v96 or another internet browser

**License:** GNU GPL v3.0

**Any restrictions to use by non-academics:** License needed

## Abbreviations

DSM: Diagnostic and statistical manual of mental disorders; IQR: Interquartile range; CrI: Credible interval; FAQ: Frequently asked questions; MCMC: Markov chain Monte Carlo

## Acknowledgments

The Shiny web application participated in the 3rd annual Shiny Contest 2021. The authors would like to thank two anonymous reviewers for considerably improving the manuscript/web application and Eleftherios Meletis for performing a thorough test of the *tPRiors* shiny application.

## Authors' contributions

KP and PK contributed equally. PK contributed to the concept. KP designed and developed the Shiny web application. KP and PK generated data and tested the application. KP drafted the manuscript. PK critically, under the reviewed the manuscript. Both authors further revised the design of the *tPRiors* application. The authors read and approved the final manuscript.

## Funding

The authors would like to acknowledge financial support from the European Commission, H2020 Health Grant no 101016216, unCoVer - Unravelling data for rapid evidence-based response to COVID-19. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

## Availability of data and materials

All datasets exploited in this manuscript can be found online in the corresponding *tPRiors* application specific single or multiple population pages [21].

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 25 October 2021 Accepted: 28 February 2022

Published online: 03 April 2022

## References

1. Branscum AJ, Gardner IA, Johnson WO. Bayesian modeling of animal- and herd-level prevalences. *Prev Vet Med.* 2004;66:101–12.
2. Hanson T, Johnson WO, Gardner IA. Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *J Agric Biol Environ Stat.* 2003;8:223–39.
3. Flor M, Weis M, Selhorst T, Müller-Graf C, Greiner M. Comparison of Bayesian and frequentist methods for prevalence estimation under misclassification. *BMC Public Health.* 2020;20:1135.
4. Chang W, Cheng J, Allaire J, Xie Y, McPherson J, et al. Shiny: web application framework for R. <https://shiny.rstudio.com/>. Accessed 28 Oct 2021.
5. Speybroeck N, Devleeschauwer D, Joseph L, Berkvens D. Misclassification errors in prevalence estimation: Bayesian handling with care. *Int J Public Health.* 2012;58:791–95.
6. Bacigalupo I, Mayer F, Lacorte E, Pucchio A, Marzolini F, Canevelli M, Fiandra T, Vanacore N. A Systematic Review and Meta-Analysis on the Prevalence of Dementia in Europe: Estimates from the Highest-Quality Studies Adopting the DSM IV Diagnostic Criteria. *J Alzheimers Dis.* 2018;66:1471–81.
7. Plummer M et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing. (Vol. 124, No. 125.10). 2003. p. 1–10.
8. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2020;10:325–37.
9. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. Stan: a probabilistic programming language. 2017;76:1–32.
10. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol.* 1978;107:71–76.
11. Kostoulas P, Giovannini A, Alba A, Comin A, Meletis E, Iannetti S, Napp S, Lindberg A, Solomakos N. Harmonisation Of Transmissible disease Interpretation in the EU. EFSA supporting publ. 2019;16(7):1678E.
12. Kostoulas P. PriorGen: Generates Prior Distributions for Proportions. 2018. <https://cran.r-project.org/web/packages/PriorGen/index.html>. Accessed 28 Oct 2021.

13. Prince M, Jackson J, Ferri CP, Sousa R, Albanese E, Ribeiro WS. Honyashiki MWorld Alzheimer Report 2009. *Alzheimers Dis Int*. [https://alz.org/national/documents/report\\_full\\_2009worldalzheimerreport.pdf](https://alz.org/national/documents/report_full_2009worldalzheimerreport.pdf). Accessed 28 Oct 2012.
14. Lopez OL, McDade E, Riverol M, Becker JT. Evolution of the diagnostic criteria for degenerative and cognitive disorders. *Curr Opin Neurol*. 2011;24:532–41.
15. Tsai JC, Chen CW, Chu H, Yang HL, Chung MH, Liao YM, Chou KR. Comparing the Sensitivity Specificity, and Predictive Values of the Montreal Cognitive Assessment and Mini-Mental State Examination When Screening People for Mild Cognitive Impairment and Dementia in Chinese Population. *Archiv fur Psychologie*. 2016;30:486–91.
16. Trapp W, Röder S, Heid A, Billman P, Daiber S, Hajak G. Sensitivity and specificity of the Bamberg Dementia Screening Test's (BDST) full and short versions: brief screening instruments for geriatric patients that are suitable for infectious environments. *BMC Med*. 2021;19(1):1–3.
17. Mackin RS, Ayalon L, Feliciano L, Areán PA. The Sensitivity and Specificity of Cognitive Screening Instruments to Detect Cognitive Impairment in Older Adults With Severe Psychiatric Illness. *J Geriatr Psychiatry Neurol*. 2010;23:94–99.
18. Wang J, Keusters WR, Wen L, Leeflang MMG. IPDmada. *Res Synth Methods*. 2020;12:45–54.
19. Tebé C, Valls J, Satorra P, Tobías A. COVID19-world: a shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic. *BMC Med Res Methodol*. 2020;20(1):1–7.
20. Yu Y, Ouyang Y, Yao W. shinyCircos: an R/Shiny application for interactive creation of Circos plot. *Bioinformatics*. 2017;34:1229–31.
21. Pateras K, Polychronis K. kpatera/tPRiors: tPRiors: A tool for prior elicitation and obtaining posterior distributions of true disease prevalence: Zenodo; 2022. <https://doi.org/10.5281/zenodo.5925650>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

