Edmund H. Wilkes*

veRification: an R Shiny application for laboratory method verification and validation

https://doi.org/10.1515/cclm-2023-0158
Received February 15, 2023; accepted March 23, 2023;
published online April 13, 2023

Abstract

Objectives: According to international standards, clinical laboratories are required to verify the performance of assays prior to their implementation in routine practice. This typically involves the assessment of the assay's imprecision and trueness vs. appropriate targets. The analysis of these data is typically performed using frequentist statistical methods and often requires the use of closed source, proprietary software. The motivation for this paper was therefore to develop an open-source, freely available software capable of performing Bayesian analysis of verification data.
Methods: The veRification application presented here was developed with the freely available R statistical computing environment, using the Shiny application framework. The codebase is fully open-source and is available as an R package on GitHub.
Results: The developed application allows the user to analyze imprecision, trueness against external quality assurance, trueness against reference material, method comparison, and diagnostic performance data within a fully Bayesian framework (with frequentist methods also being available for some analyses).
Conclusions: Bayesian methods can have a steep learning curve and thus the work presented here aims to make Bayesian analyses of clinical laboratory data more accessible. Moreover, the development of the application and seeks to encourage the dissemination of open-source software within the community and provides a framework through which Shiny applications can be developed, shared, and iterated upon.

Keywords: Bayesian; statistics; verification.

Introduction

Verification of new assays forms a central part of clinical laboratory practice in order to formally confirm, through provision of objective evidence, that specified requirements for an assay's performance have been fulfilled. Such analyses form a key part of clinical laboratory accreditation and a variety of guidelines exist that provide direction as to the experimental evidence that is required [1–3]. In general, these define experiments that are designed to assess an assay's imprecision (stochastic error) and trueness (systematic error), with the latter being able to be tested through a variety of means (e.g. against a reference material or an external quality assurance scheme). These documents also provide recommendations for ways in which the data produced by these experiments should be analyzed, with the vast majority detailing methods within the frequentist paradigm of statistical analysis (e.g. t-tests, nested analysis of variance, and Passing–Bablok regression) [1–3]. These methods can easily provide the probability of having collected the verification data (or more extreme values), assuming a given null hypothesis to be true ($P(data|hypothesis)$), yet are incapable of answering the more intuitive question: what is the probability that my assay is meeting a given requirement given the data I have collected ($P(hypothesis|data)$) [4]? Such questions are arguably much more intuitive and of clear practical significance in the verification of new assays in clinical laboratories.

The Bayesian statistical paradigm leverages Bayes theorem to incorporate prior information into an analysis and directly calculate $P(hypothesis|data)$ [5]. When used in a principled and transparent way, this paradigm can counter the type M (magnitude) and S (sign) errors, and misinterpretations common to frequentist methods, which have contributed to the replication crisis in the medical sciences [6–14]. Unfortunately, despite their increased tractability due to advances in computing and accessibility through open-source software (e.g. Python and/or R packages [14, 15]), Bayesian methods can have a relatively steep learning curve and, to the author's knowledge, no accessible tools exist for their specific use for laboratory method verification. Moreover, regardless of statistical paradigm, the vast majority of tools available for the analysis of laboratory method verification data are proprietary software with significant cost

***Corresponding author: Edmund H. Wilkes**, Department of Clinical Biochemistry, North West London Pathology, Imperial College Healthcare NHS Trust, Charing Cross Hospital, Fulham Palace Road, Hammersmith, W6 8RF, London, UK, Phone: +44(0)20 331 11400,
E-mail: edmund.wilkes@nhs.net

implications (e.g. Microsoft Excel, Analyse-it, and Finbiosoft's Validation Manager, amongst others). The aims of the work presented here were therefore two-fold: (i) to develop an open-source, accessible, and flexible application for the analysis of laboratory method verification data; and (ii) to expand on previous work [16] to improve the accessibility of the application of Bayesian statistical methods to laboratory method verification data in order to harness their aforementioned benefits. This manuscript outlines the development of an application that fulfils these aims and provides instructions for its use. The application detailed here was built using the freely available Shiny R package [17] and can therefore be used for free from a local instance of RStudio [18] or base R [19] running on a Linux, Windows, or macOS-based operating system. Alternatively, Shiny applications such as these can be hosted on a public or private web server and subsequently accessed through a web browser. The codebase is fully open-source in order to allow users to both understand how the analyses are conducted, and to encourage the continued development of its capabilities by the community. The structure of the application was heavily influenced and inspired by the Association of Clinical Biochemistry spreadsheets authored by Anders Kallner and colleagues [1].

Materials and methods

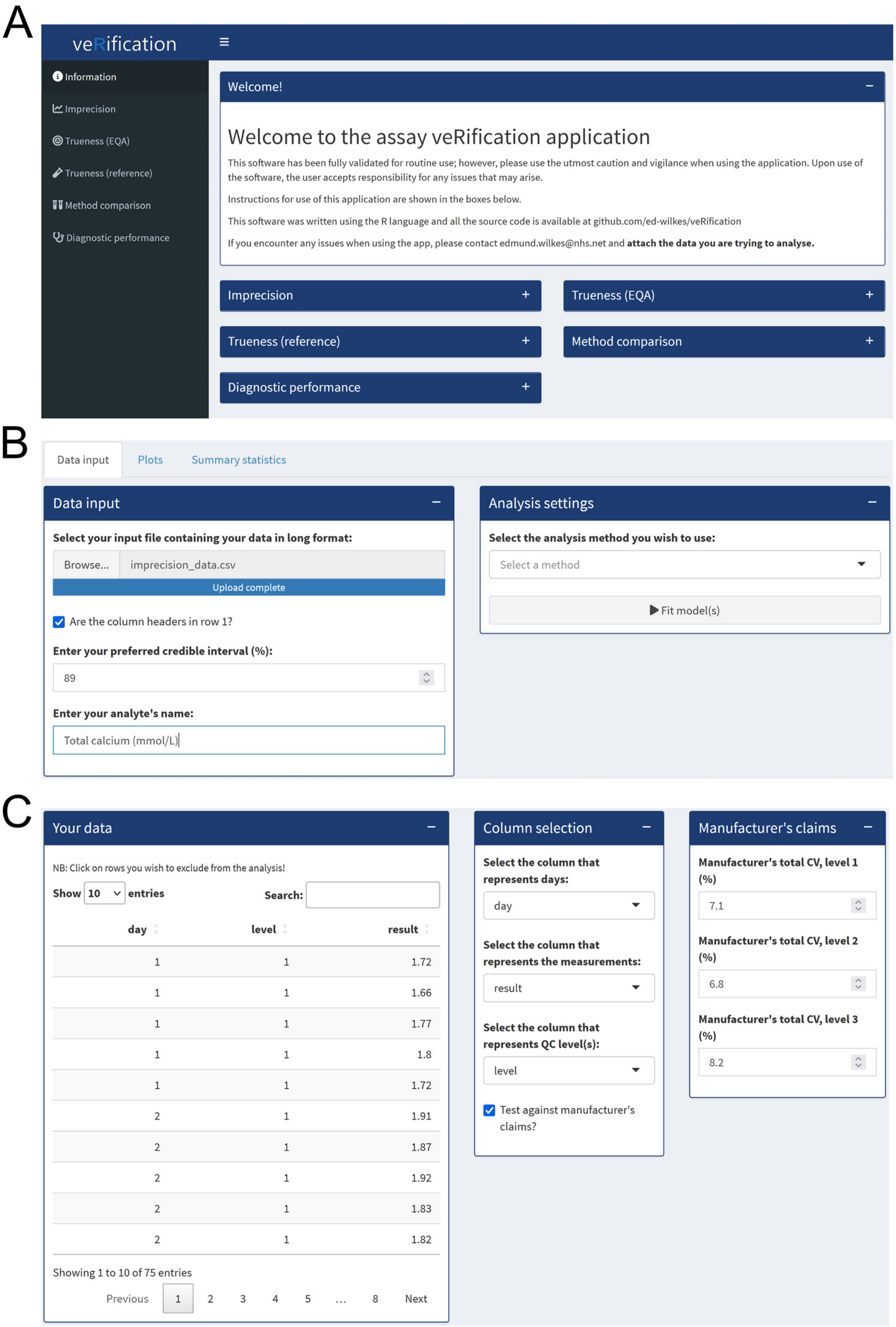
The open-source application shown here was developed within RStudio [18] and the R statistical computing environment [19] which are available to download for free at https://posit.co/downloads/ and https://cran.r-project.org/, respectively. The application was built through the use of the Shiny package (https://shiny.rstudio.com/) and makes use of a number of other R packages [14, 15, 20–27]. The full codebase for the application can be found at https://github.com/ed-wilkes/veRification and is written in a modular way within the *golem* framework [28]. This means that application can be installed as an R package and run from a local instance of Posit or R on a local machine running a Linux, Windows, or macOS-based operating system, or hosted on a web server and accessed through a web browser. It is worth noting that the performance will depend on the hardware on which you are running R and RStudio. In addition, the application's modularity enables interested users to fork the repository and easily develop and add new modules as they wish.

Results

Assessing assay imprecision

The assessment of an assay's imprecision is perhaps one of the first steps of a method's verification. In the UK, the Association for Clinical Biochemistry and Laboratory Medicine (ACB) recommends the measurement of at least two levels of internal quality control (IQC) material 5 times a day (spread throughout the day), across 5 different days [1]. Such an experiment allows the calculation of within- and between-day variability (referred to as repeatability and intermediate imprecision, respectively), alongside the total variation across the time course and the expected value. Analysis of these data is commonly performed with a nested ANOVA (analysis of variance), but can equivalently be modeled with a Bayesian linear varying-effects model [16] (Eqs. 4.1– 4.5) in order to make direct probability statements regarding the model's parameters and propagate our uncertainty regarding their values into our inferences. Bayesian (and frequentist) varying-effects models can be fitted within the "Imprecision" tab. The Bayesian analyses uses the default, weakly informative prior distributions recommended by the *rstanarm* team [15], scaled to the input data (Eqs. 4.4 and 4.5, where m_y and s_y represent the mean and standard deviation of the input data, respectively) and the frequentist models – estimated by means of restricted maximum likelihood (REML) – are fitted through the use of the *VCA* package [20]. This module allows the simultaneous analysis of multiple levels of IQC material, automatically fitting a separate model to each level.

First, the user is prompted to input their data in .csv or .xls(x) format. The inputted data are then displayed to the user and the user prompted to interactively select the columns within the data that represent the days of measurement, the IQC levels, and the measurements themselves. The user can also select whether to test the estimated total laboratory CV against a given claim (e.g. from a manufacturer's kit insert or another data source) (Figure 1A–C). Once these settings are chosen, the data are plotted with an interactive visualization and the modeling results are presented to the user when complete (Figure 2A). A number of basic checks of the Bayesian model's validity are performed and the results of these (pass or fail) are also shown. These checks determine if the Markov chains (MCMC) have converged (parameter $\widehat{\mathbb{R}}$ values ≤1.1 and that the effective sample sizes are adequate (≥10% of the total number of posterior samples) [29–31]. The MCMC traces for each of the model's parameters are shown within the "Bayesian model diagnostics" tab, alongside their full posterior distributions (Figure 2B). The user is strongly encouraged to check these for consistency [29, 30] and this remains relevant for each of the Bayesian models fitted within the application. If the user-defined credible intervals – which default to 89% – of the posterior distribution of the total laboratory CV for a given level of QC is higher than the user-provided claimed CV value (i.e. a one-tailed test), then this is highlighted to the user. An equivalent inference is performed if the frequentist method is chosen. It cannot be



Welcome to the assay veRification application

This software has been fully validated for routine use; however, please use the utmost caution and vigilance when using the application. Upon use of the software, the user accepts responsibility for any issues that may arise.

Instructions for use of this application are shown in the boxes below.

This software was written using the R language and all the source code is available at github.com/ed-wilkes/veRification

If you encounter any issues when using the app, please contact edmund.wilkes@nhs.net and **attach the data you are trying to analyse.**

Select your input file containing your data in long format: imprecision_data.csv

☑ Are the column headers in row 1?

Enter your preferred credible interval (%): 89

Enter your analyte's name: Total calcium (mmol/L)

Select the analysis method you wish to use:

NB: Click on rows you wish to exclude from the analysis!

Show 10 entries

Search:

day	level	result
1	1	1.72
1	1	1.66
1	1	1.77
1	1	1.8
1	1	1.72
2	1	1.91
2	1	1.87
2	1	1.92
2	1	1.83
2	1	1.82

Showing 1 to 10 of 75 entries

Select the column that represents days: day

Select the column that represents the measurements: result

Select the column that represents QC level(s): level

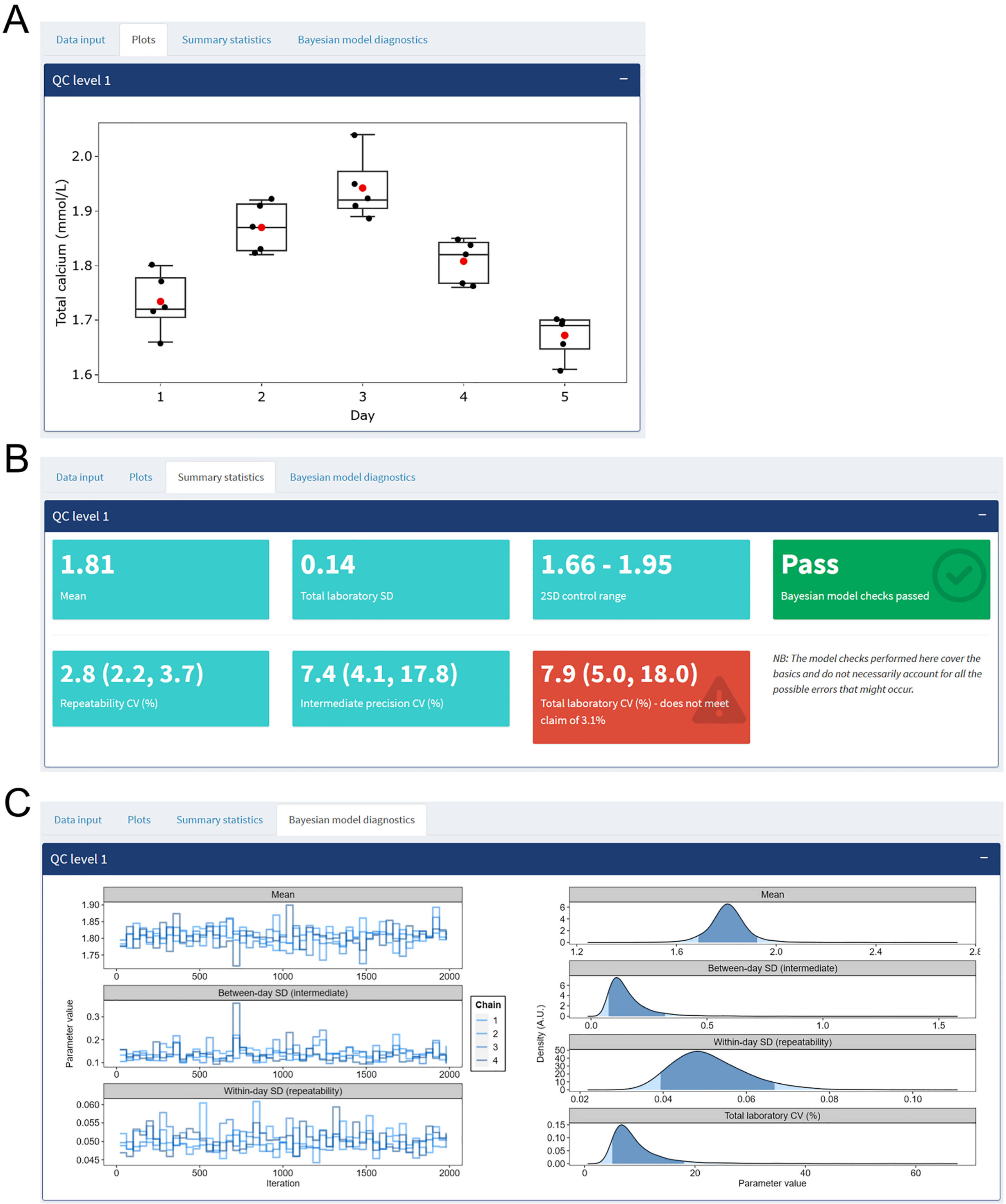
☑ Test against manufacturer's claims?

Manufacturer's total CV, level 1 (%) 7.1

Manufacturer's total CV, level 2 (%) 6.8

Manufacturer's total CV, level 3 (%) 8.2

Figure 1: The application homepage and imprecision module input. (A) The welcome screen to the application. The different modules can be accessed through the side bar. Guidance for each of the modules is provided on the welcome screen under dropdown boxes shown. (B) The data input screen for the "Imprecision" module. (C) Once data are uploaded, they can be viewed interactively. The user is prompted to select which columns within the data represent days, measured values, and QC levels. The user can then determine whether the data should be tested against imprecision values provided by a manufacturer (or other source) and input these as relevant.



1.81
Mean

0.14
Total laboratory SD

1.66 - 1.95
2SD control range

Pass
Bayesian model checks passed

2.8 (2.2, 3.7)
Repeatability CV (%)

7.4 (4.1, 17.8)
Intermediate precision CV (%)

7.9 (5.0, 18.0)
Total laboratory CV (%) - does not meet claim of 3.1%

NB: The model checks performed here cover the basics and do not necessarily account for all the possible errors that might occur.

Figure 2: Imprecision module output. (A) An interactive visualization of one QC level of the input data under the “Plots” tab. Boxplots are displayed for each day of data, showing the median, 25th and 75th percentiles, and tails representing 1.5 times the interquartile range. Red points represent the mean of each day. Individual data points are shown in black – a small amount of random jitter is added for increased clarity. (B) The summary of the posterior distributions of each of the imprecision model’s parameters under the “Summary statistics” tab. (C) Visualizations of the Markov chains and full posterior distributions of the model’s parameters. The dark, shaded areas of the posterior distributions represent the user-defined credible interval chosen under the “Data input” tab.

overstated that the choice of interval is entirely arbitrary and the user is strongly encouraged to view and inspect the entire posterior distribution in the “Bayesian model diagnostics” tab. Note that the credible intervals for the remaining parameters are not displayed for clarity, as it is assumed that the coefficients of variation are the parameters of most interest. If required, the user can view the full posterior distributions in order to assess the uncertainty in their values (Figure 2C).

$$y_i \sim N(\text{mean} = \mu_i, \text{sd} = \sigma) \tag{4.1}$$

$$\mu_i = \alpha_{day,\,i} \tag{4.2}$$

$$\alpha_{day} \sim N(\overline{\alpha}, \sigma_{day}) \tag{4.3}$$

$$\overline{\alpha} \sim N(m_y, 2.5 \cdot s_y) \tag{4.4}$$

$$\sigma \sim \text{exponential}(\text{rate} = 1/s_y) \tag{4.5}$$

Assessing trueness against external quality assurance materials

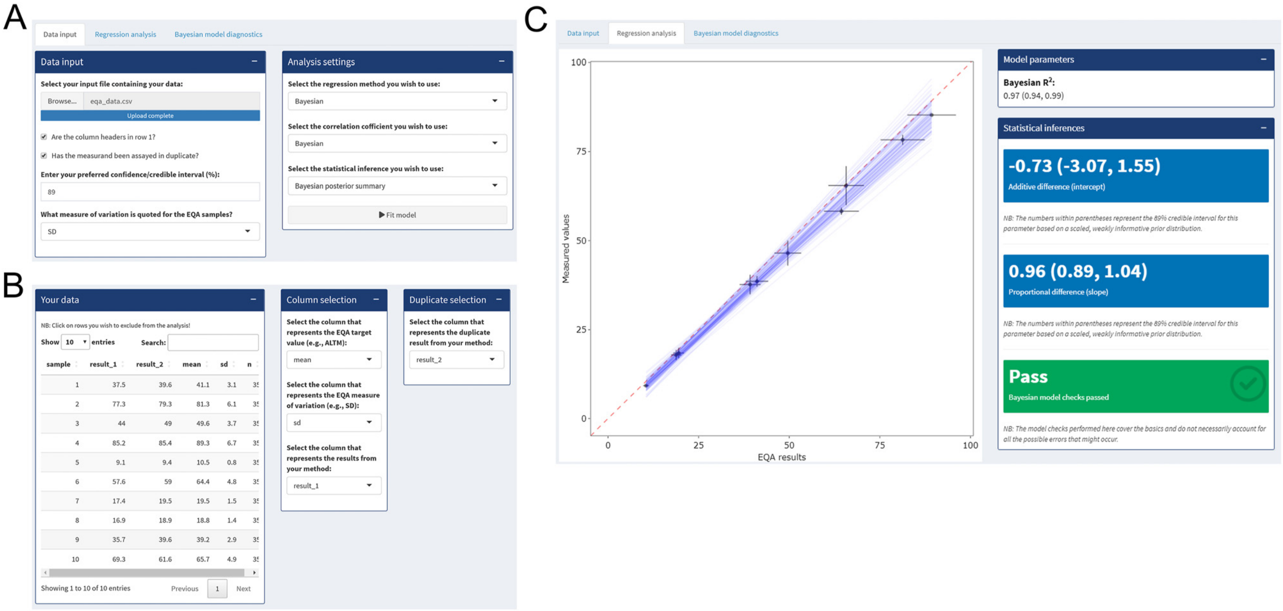
The trueness of an assay, defined as the closeness of agreement between the arithmetic mean of a large number of test results and the true or accepted reference value [32], is fundamental to its clinical utility. This is pragmatically tested through comparison of the values obtained using the assay in question to those reported through an external quality assurance (EQA) scheme. Ideally, this comparison encompasses a large range of clinically-relevant concentrations of the analyte and is preferably performed in duplicate to account for within-assay variability [1]. The application allows the user to analyze these data in a number of different ways within the “Trueness (EQA)” tab using both frequentist (ordinary least-squares, Deming, or Passing–Bablok regression analyses) or Bayesian methods, depending on the user’s preference (Figure 3A and B). The Bayesian linear regression models take the form shown in Eqs. 4.6– 4.10 (where s_x represents the standard deviation of all the EQA results) if no duplicate measurements are present and a full measurement error model (Eqs 4.11– 4.17) if they are available. In this measurement error model, the standard deviation of each measured data point is estimated from the duplicate measurements. Once fitted, the data are visualized and the modeling results presented to the user (Figure 3C). The user is again strongly encouraged to view the posterior distribution and MCMC traces of each parameter in their entirety under the “Bayesian model diagnostics” tab.

$$y_i \sim N(\mu_i, \sigma) \tag{4.6}$$

$$\mu_i = \beta_0 + \beta_1 \cdot x_i \tag{4.7}$$

$$\beta_0 \sim N(0, 2.5 \cdot s_y) \tag{4.8}$$

$$\beta_1 \sim \text{LogNormal}(\text{meanlog} = 0, \text{sdlog} = 1) \tag{4.9}$$



Data input

Select your input file containing your data: eqa_data.csv

☑ Are the column headers in row 1?

☑ Has the measurand been assayed in duplicate?

Enter your preferred confidence/credible interval (%): 89

What measure of variation is quoted for the EQA samples? SD

Analysis settings

Select the regression method you wish to use: Bayesian

Select the correlation cofficient you wish to use: Bayesian

Select the statistical inference you wish to use: Bayesian posterior summary

Your data

NB: Click on rows you wish to exclude from the analysis!

Show 10 entries Search:

sample	result_1	result_2	mean	sd	n
1	37.5	39.6	41.1	3.1	35
2	77.3	79.3	81.3	6.1	35
3	44	49	49.6	3.7	35
4	85.2	85.4	89.3	6.7	35
5	9.1	9.4	10.5	0.8	35
6	57.6	59	64.4	4.8	35
7	17.4	19.5	19.5	1.5	35
8	16.9	18.9	18.8	1.4	35
9	35.7	39.6	39.2	2.9	35
10	69.3	61.6	65.7	4.9	35

Showing 1 to 10 of 10 entries Previous 1 Next

Column selection

Select the column that represents the EQA target value (e.g., ALTM): mean

Select the column that represents the EQA measure of variation (e.g., SD): sd

Select the column that represents the results from your method: result_1

Duplicate selection

Select the column that represents the duplicate result from your method: result_2

Model parameters

Bayesian R^2: 0.97 (0.94, 0.99)

Statistical inferences

-0.73 (-3.07, 1.55) Additive difference (intercept)

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a scaled, weakly informative prior distribution.

0.96 (0.89, 1.04) Proportional difference (slope)

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a scaled, weakly informative prior distribution.

Pass Bayesian model checks passed

NB: The model checks performed here cover the basics and do not necessarily account for all the possible errors that might occur.

Figure 3: Trueness assessed against EQA material. (A) The data input options for the EQA module. (B) Interactive tabulation of the uploaded data. The user is prompted to dynamically select the columns that represent the EQA means, uncertainty values, and the measured values. (C) Visualization of the input data (black points and associated error bars) and a randomized sample from the posterior distributions of the fitted Bayesian model (blue lines). The posterior distributions of the model’s parameters are summarized and presented to the user, alongside the results of the model checks.

$$\sigma \sim \text{exponential}(1/s_y) \tag{4.10}$$

$$y_{obs,i} = N\left(y_{true,i}, \sigma_{y,i}\right) \tag{4.11}$$

$$y_{true,i} = N(\mu_i, \sigma) \tag{4.12}$$

$$\mu_i = \beta_0 + \beta_1 \cdot x_{true,i} \tag{4.13}$$

$$x_{obs,i} = N(x_{true,i}, \sigma_{x,i}) \tag{4.14}$$

$$\beta_0 \sim N(0, 2.5 \cdot s_y) \tag{4.15}$$

$$\beta_1 \sim \text{LogNormal}(0,1) \tag{4.16}$$

$$\sigma \sim \text{exponential}(1/s_y) \tag{4.17}$$

Assessing trueness against reference materials

In addition to the assessments vs. EQA material, the trueness of an assay can also be assessed by comparison to a reference material whose analyte concentration has been assigned through the use of a reference method (or appropriate substitute). Such analyses are typically performed through measurement of the reference material on 3–5 occasions in duplicate [1]. These data are then often analyzed through the use to assess the significance of the difference between the measured and assigned values in a Neyman–Pearson, frequentist framework. For reasons discussed elsewhere [5, 16], Bayesian methods provide a formal and more intuitive way through which the probability of a difference between the reference and test methods can be calculated. This is especially true in this case, where a relatively strong prior distribution for the results is already known (the assigned value ± an assigned uncertainty). Within the “Trueness (reference)” tab, the application prompts the user to upload their data, enter the assigned value and associated uncertainty, and choose the relevant columns of their uploaded data (Figure 4A and B). A model is fitted to the data using the assigned value and uncertainty as the prior distribution (Eqs. 4.18– 4.21, where m_p and s_p represent the reference material’s assigned mean and SD, respectively, and ν represents the degrees of freedom of the Student’s t distribution). The Student’s t distribution is used here in order to allow for more extreme observations [30]. Thus, the model directly assesses the question: what is the probability that the investigated assay is providing different results to those expected for the reference material, given the data we have collected? If duplicate measurements are included in the data set, then a varying-effects model is fitted to the data in order to take into account the within-assay variation (Eqs. 4.22–4.24, where $\alpha_{rep,i}$ represents the mean for the ith pair of measurements and σ_{rep} represents the between-replicate variation), using the same likelihood function and residual prior (Eqs. 4.18 and 4.21, respectively). The results of the modeling are visualized for the user and a Bayes factor is presented as a measure of evidence against the assigned mean of the reference material (Figure 4C) [27]. Once again, the full posterior distributions and MCMC traces are shown in the “Bayesian modeling diagnostics” tab.

$$y_i \sim N(\mu_i, \sigma) \tag{4.18}$$

$$\mu_i = \beta_0 \tag{4.19}$$

$$\beta_0 \sim T(\nu = 30, m_p, s_p) \tag{4.20}$$

$$\sigma \sim \text{exponential}(1/s_y) \tag{4.21}$$

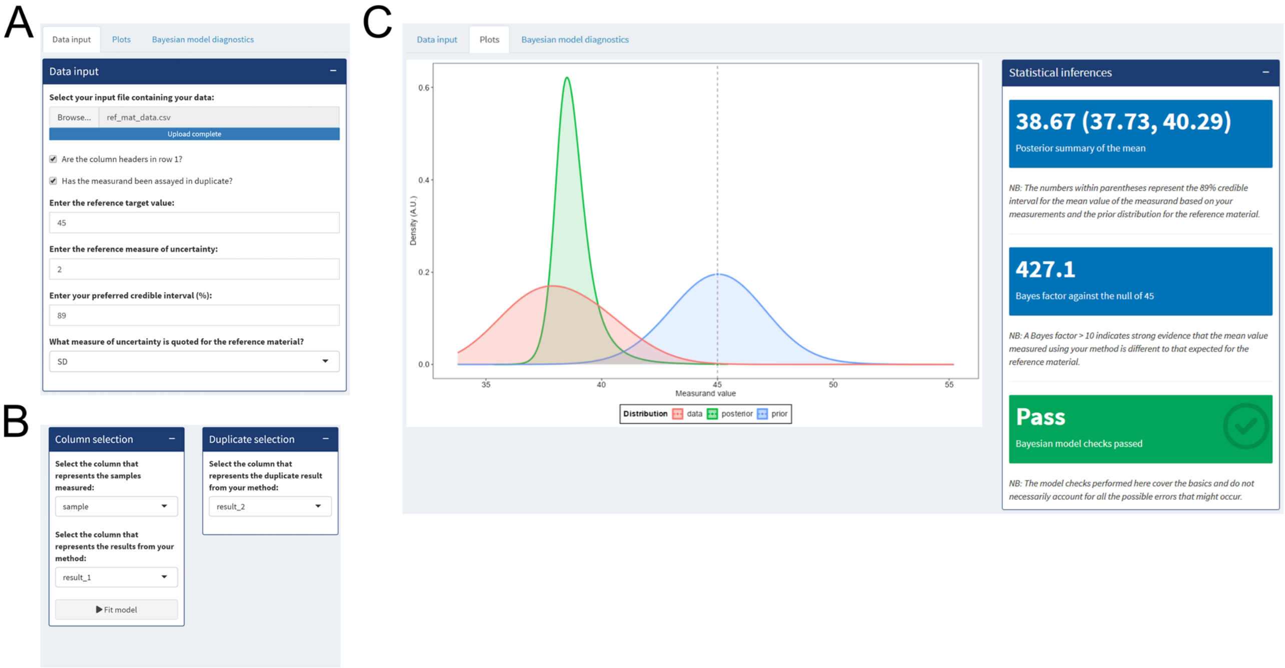
$$\mu_i = \alpha_{rep,i} \tag{4.22}$$

$$\alpha_{rep} \sim N(\overline{\alpha}, \sigma_{rep}) \tag{4.23}$$

$$\overline{\alpha} \sim T(\nu = 30, m_p, s_p) \tag{4.24}$$

Assessing trueness against a reference assay

Comparisons of two methods for the measurement of a given measurand are commonly performed within clinical laboratories. The current UK ACB guidelines [1] suggest that at least 20 samples consisting of patient material should be assayed with each method in a timely manner (preferably in duplicate). The application allows the user to analyze these data in a number of different ways using both frequentist (ordinary least-squares, Deming, or Passing–Bablok regression analyses) and Bayesian methods, depending on the user’s preference, within the “Method comparison” tab (Figure 5A and B). As with the EQA data analysis, the Bayesian linear regression models here take the form shown in Eqs. 4.6– 4.10 (where s_x represents the standard deviation of the reference method’s results) if no duplicate measurements are present, and a full measurement error model (Eqs. 4.11– 4.17) if they are available. The standard deviation of each measured data point is again estimated from the duplicate measurements present in the data. Once fitted, the data are visualized and the modeling results presented to the user (Figure 5C). This section of the application also performs a traditional Bland–Altman analysis [33], allowing the user to determine whether they wish to explore absolute (additive) or relative (proportional) differences between the two methods’ results (Figure 5D). In this application, relative differences are presented as $\log_2$ fold-ratios of the results



Data input

Select your input file containing your data:

Browse... ref_mat_data.csv

Upload complete

☑ Are the column headers in row 1?

☑ Has the measurand been assayed in duplicate?

Enter the reference target value: 45

Enter the reference measure of uncertainty: 2

Enter your preferred credible interval (%): 89

What measure of uncertainty is quoted for the reference material? SD

Column selection

Select the column that represents the samples measured: sample

Select the column that represents the results from your method: result_1

▶ Fit model

Duplicate selection

Select the column that represents the duplicate result from your method: result_2

Statistical inferences

38.67 (37.73, 40.29)

Posterior summary of the mean

NB: The numbers within parentheses represent the 89% credible interval for the mean value of the measurand based on your measurements and the prior distribution for the reference material.

427.1

Bayes factor against the null of 45

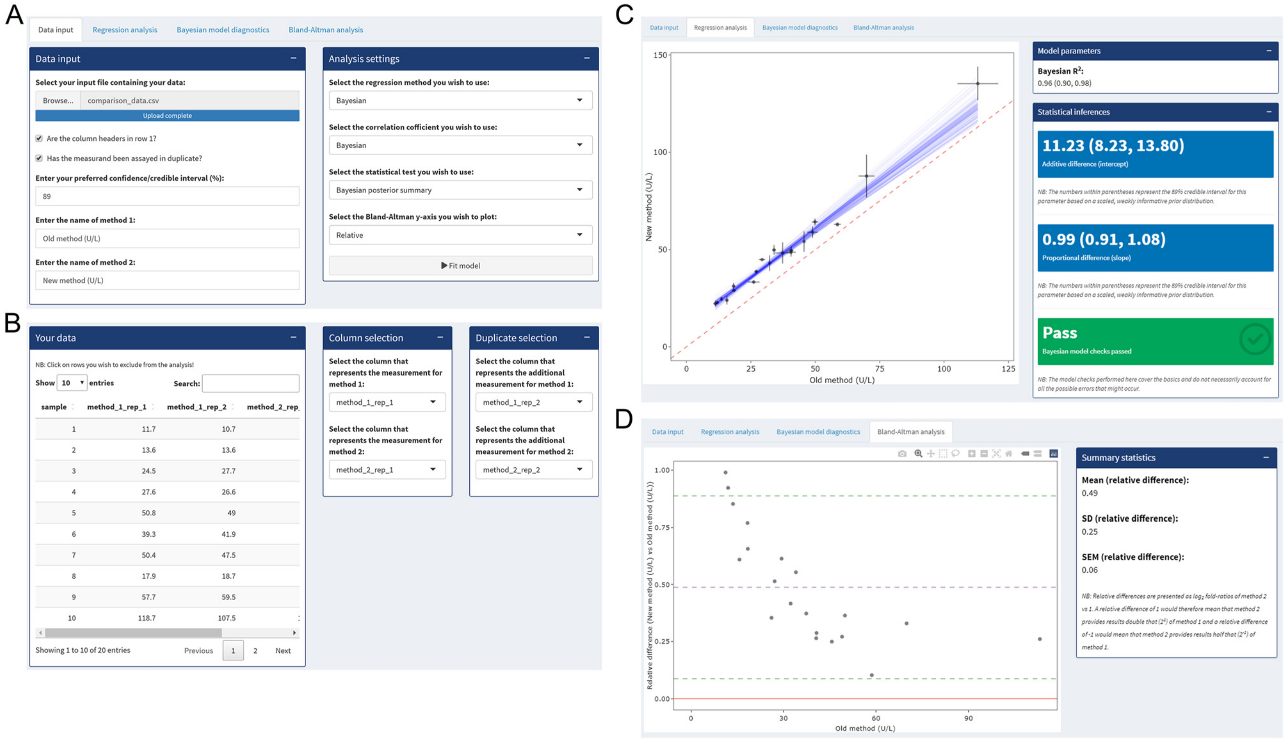
NB: A Bayes factor > 10 indicates strong evidence that the mean value measured using your method is different to that expected for the reference material.

Pass ✓

Bayesian model checks passed

NB: The model checks performed here cover the basics and do not necessarily account for all the possible errors that might occur.

Figure 4: Trueness assessed against reference material. (A) The data input options for the trueness against reference materials module. (B) The user is prompted to dynamically select the columns in the data that represent the measurements and their duplicates, if present. (C) Visualization of the prior, data, and posterior distributions. The posterior distribution for the mean parameter is summarized and the Bayes factor vs the prior distribution is shown alongside the result of basic model checks.



Data input

Select your input file containing your data:

Browse... comparison_data.csv

Upload complete

☑ Are the column headers in row 1?

☑ Has the measurand been assayed in duplicate?

Enter your preferred confidence/credible interval (%): 89

Enter the name of method 1: Old method (U/L)

Enter the name of method 2: New method (U/L)

Analysis settings

Select the regression method you wish to use: Bayesian

Select the correlation cofficient you wish to use: Bayesian

Select the statistical test you wish to use: Bayesian posterior summary

Select the Bland-Altman y-axis you wish to plot: Relative

▶ Fit model

Your data

NB: Click on rows you wish to exclude from the analysis!

Show 10 entries Search:

sample	method_1_rep_1	method_1_rep_2	method_2_rep
1	11.7	10.7	
2	13.6	13.6	
3	24.5	27.7	
4	27.6	26.6	
5	50.8	49	
6	39.3	41.9	
7	50.4	47.5	
8	17.9	18.7	
9	57.7	59.5	
10	118.7	107.5	

Showing 1 to 10 of 20 entries Previous 1 2 Next

Column selection

Select the column that represents the measurement for method 1: method_1_rep_1

Select the column that represents the measurement for method 2: method_2_rep_1

Duplicate selection

Select the column that represents the additional measurement for method 1: method_1_rep_2

Select the column that represents the additional measurement for method 2: method_2_rep_2

Model parameters

Bayesian R^2: 0.96 (0.90, 0.98)

Statistical inferences

11.23 (8.23, 13.80)

Additive difference (intercept)

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a scaled, weakly informative prior distribution.

0.99 (0.91, 1.08)

Proportional difference (slope)

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a scaled, weakly informative prior distribution.

Pass ✓

Bayesian model checks passed

NB: The model checks performed here cover the basics and do not necessarily account for all the possible errors that might occur.

Summary statistics

Mean (relative difference): 0.49

SD (relative difference): 0.25

SEM (relative difference): 0.06

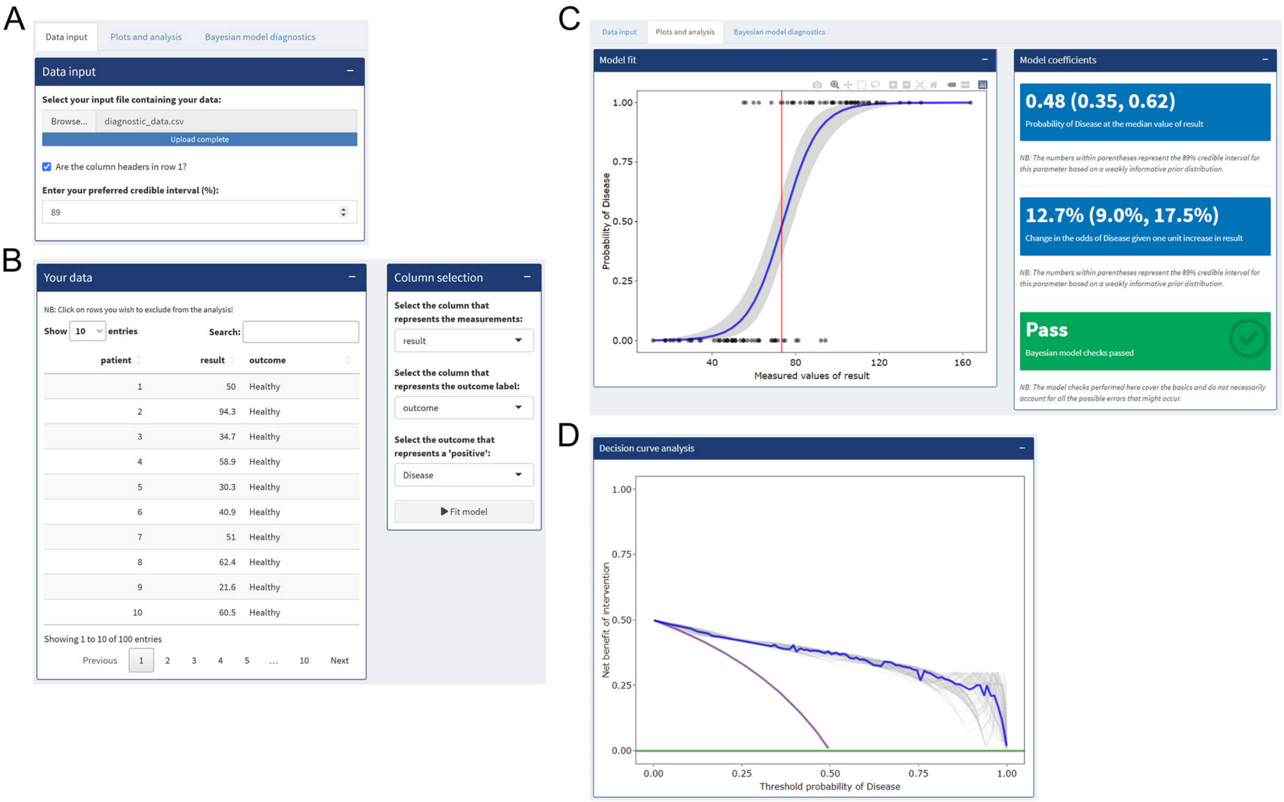
NB: Relative differences are presented as $\log_2$ fold-ratios of method 2 vs 1. A relative difference of 1 would therefore mean that method 2 provides results double that (2^1) of method 1 and a relative difference of -1 would mean that method 2 provides results half that (2^{-1}) of method 1.

Figure 5: Method comparison module. (A) The data input options for the “Method comparison” module. Several regression methods are available and the user can select which type of Bland–Altman analysis they wish to perform. (B) Visualization of the input data and dynamic selection of the columns that represent the measurements and their duplicates, if present. (C) Visualization of the input data and a randomized sample from the posterior distributions of the fitted Bayesian model (blue lines). The posterior distributions of the model’s parameters are summarized and presented to the user, alongside the results of the model checks. (D) Bland–Altman analysis. The mean, SD, and SEM of the differences are shown.

from the two methods – analogous to MA plots in high-throughput transcriptomics/genomics analyses [34] – as these more readily represent the direction and magnitude of the differences on a multiplicative scale and are conveniently symmetrical around zero.

Assessing the diagnostic performance of a test

The final tool within the application is designed to assess the performance of a given test to correctly predict a binary categorical outcome using a continuous variable (e.g. predicting “healthy” vs “disease” through the measurement of a biomarker). This is often achieved through the use of receiver operating characteristic (ROC) analysis, which typically involves the derivation of a hard analyte threshold based on minimizing a loss function that balances a test’s sensitivity and specificity. This type of analysis is problematic in medicine, however, for several reasons. Firstly, as clinical decision makers, we are most often analyzing scenarios in which there is significant stochasticity in a given clinical outcome due to measurement error, biological variation, and sampling variability. As such, there is often a significant overlap between the two groups being compared and thus probability estimates are most appropriate to best quantify the tendency towards one group or the other – i.e. directly inferring and interpreting the forward probability, $P(disease|data)$ [35, 36]. Secondly, ROC – and indeed the alternative precision-recall curve – analyses do not directly inform the analyst as to the clinical utility of a new test, but instead focus on the prediction of a dichotomous classification through minimization of a loss function that is assumed to be equal across individuals [35–38]. Lastly, ROC analyses are insensitive to the outcome’s distribution and thus are inappropriate for scenarios in which one category is much less frequent than the other. The “Diagnostic performance” module in this application instead first fits a Bayesian logistic regression model (Eqs. 4.25–4.30) to convert the predictor variable’s measurements to probabilities of the given



Select your input file containing your data: diagnostic_data.csv

☑ Are the column headers in row 1?

Enter your preferred credible interval (%): 89

NB: Click on rows you wish to exclude from the analysis!

Show 10 entries Search:

patient	result	outcome
1	50	Healthy
2	94.3	Healthy
3	34.7	Healthy
4	58.9	Healthy
5	30.3	Healthy
6	40.9	Healthy
7	51	Healthy
8	62.4	Healthy
9	21.6	Healthy
10	60.5	Healthy

Showing 1 to 10 of 100 entries

Previous 1 2 3 4 5 … 10 Next

Select the column that represents the measurements: result

Select the column that represents the outcome label: outcome

Select the outcome that represents a 'positive': Disease

0.48 (0.35, 0.62)
Probability of Disease at the median value of result

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a weakly informative prior distribution.

12.7% (9.0%, 17.5%)
Change in the odds of Disease given one unit increase in result

NB: The numbers within parentheses represent the 89% credible interval for this parameter based on a weakly informative prior distribution.

Pass
Bayesian model checks passed

NB: The model checks performed here cover the basics and do not necessarily account for all the possible errors that might occur.

Figure 6: Diagnostic performance module. (A) Data input for the “Diagnostic performance” module. (B) Visualization of the input data and dynamic selection of the measurements, outcome measure, and what is considered a “positive” result. (C, left panel) Posterior draws of the expected value of the posterior predictive distribution from the model (blue line) are plotted against the input data (black points). The chosen width of credible interval is shown as a gray ribbon. (C, right panel) A summary of the posterior distributions of the model’s parameters with the chosen credible interval. (D) The results of the decision curve analysis. The results of intervening in all and no patients are shown in purple and green, respectively. The decision curve analysis performed using expected value of the posterior predictive distribution of the model is shown in blue, with 100 posterior draws shown in gray in order to visualize the uncertainty in the analysis.

outcome. The module then uses the output from this model (and the associated uncertainty in its parameters) to perform a decision curve analysis [37, 38] to best assess the clinical utility of the predictor variable.

$$y_i \sim \text{Bernoulli}(p) \tag{4.25}$$

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot x_i \tag{4.26}$$

$$\beta_0, \beta_1 \sim N(0, 1.5) \tag{4.27}$$

As with the previous modules, the user is prompted to enter their data set in.csv or.xls(x) format and select the columns in their data that represent the analyte measurements, binary outcome measure, and which category of the outcome measure is considered a "positive" (Figure 6A and B). The results of the analysis are then generated within the "Plots and analysis" tab and include: (i) posterior draws of the expected value of the posterior predictive distribution vs the input data; (ii) a summary of the posterior distributions of the model's parameters; and (iii) the results of a decision curve analysis (including 100 posterior draws of expected value of the posterior predictive distribution in gray in order to display uncertainty) (Figure 6C and D).

Discussion and conclusions

The verification of assays in clinical laboratories is of the utmost importance to ensure that given performance characteristics are achieved in a given laboratory's hands prior to clinical use. Several documents provide guidance for the experiments required to achieve this [1–3]; however, the analysis of the produced data typically requires either the use of closed-source, accessible proprietary software, or open-source, freely available software that can have a steep learning curve and usually requires programming experience [14, 15, 24]. Moreover, the vast majority of these software – particularly those that are closed-source and/or proprietary – impose the frequentist paradigm of statistical analysis on the user. This is problematic due to the pitfalls associated with these methods discussed extensively elsewhere [4–13, 16]. As such, there is a motivation to develop an open-source, free, and accessible application that allows end users to analyze verification data using the Bayesian – in addition to the frequentist, if so desired – statistical paradigm. Here, an application is presented that fulfils these requirements and can either be installed as an R package or, as with other Shiny applications, hosted on a web server if required. The full source-code is available on GitHub and, due to its development within the *golem* framework [28], users are easily able to adapt the code to their needs by editing or adding modules as they see fit.

Research funding: None declared.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Not applicable.
Ethical approval: Not applicable.

References

1. Khatami Z, Hill R, Sturgeon C, Kearney E, Breadon P, Kallner A. Measurement verification in the clinical laboratory: a guide to assessing analytical performance during the acceptance testing of methods (quantitative examination procedures) and/or analysers. Available from: https://www.acb.org.uk/asset/34B3F3F5%2DAF91%2D4B44%2DAF184C565EDC162B/ [Accessed 19 Jan 2023].
2. Theodorsson E. Validation and verification of measurement methods in clinical chemistry. Bioanalytical 2012;4:305–20.
3. Pum J. A practical guide to validation and verification of analytical methods in the clinical laboratory. Adv Clin Chem 2019;90:215–81.
4. Colling LJ, Szűcz D. Statistical inference and the replication crisis. Rev Philos Psychol 2021;12:121–47.
5. van de Schoot R, Depaoli S, King R, Kramer B, Martens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Primers 2021; 1. https://doi.org/10.1038/s43586-020-00001-2.
6. Gelman A, Hennig C. Beyond subjective and objective in statistics. J R Stat Soc Ser A Stat Soc 2017;180:967–1033.
7. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p < 0.05". Am Stat 2019;73:1–19.
8. McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. Am Stat 2019;73:235–45.
9. van Zwet EW, Cator EA. The significance filter, the winner's curse and the need to shrink. Stat Neerl 2021;75:1–16.
10. Gelman A, Tuerlinckx F. Type S error rates for classical and Bayesian single and multiple comparison procedures. Comput Stat 2000;15: 373–90.
11. Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci 2014;9:641–51.
12. Gelman A. The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Pers Soc Psychol Bull 2018;44:16–23.
13. Szűcs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci 2017;11: 943.
14. Bürkner PC. brms: an R package for Bayesian multilevel models using stan. J Stat Softw 2017;80:1–28.
15. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan; 2022. R package version 2.21.3.
16. Wilkes EH. A practical guide to Bayesian statistics in laboratory medicine. Clin Chem 2022;68:893–905.
17. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, et al. shiny: web application framework for R; 2022. R package version 1.7.3.

18. Posit Team. RStudio: integrated development environment for R. Boston, USA: Posit Software PBC; 2022.
19. R Core Team. R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022.
20. Schuetzenmeister A, Dufey F. VCA: variance component analysis; 2022. R package version 1.4.5.
21. Chang W, Borges Ribeiro B. shinydashboard: create dashboards with 'Shiny'; 2022. R package version 0.7.2.
22. Sali A, Attali D. shinycssloaders: add loading animations to a 'shiny' output while it's recalculating; 2022. R package version 1.0.0.
23. Merlino A, Howard P. shinyFeedback: display user feedback in Shiny apps; 2022. R package version 0.4.0.
24. Manuilova E, Schuetzenmeister A. mcr: method comparison regression; 2022. R package version 1.3.0.
25. Wickham H, Averick M, Bryan J, Chang W, McGowan L, Francois R, et al. Welcome to the tidyverse. J Open Source Softw 2019;43:1686.
26. Sievert C. Interactive web-based data visualization with R, plotly, and shiny. New York: Chapman and Hall/CRC; 2020.
27. Majowski D, Mattan SB, Lüdecke D. bayestestR: describing effects and their uncertainty, existence and significance within the Bayesian framework. J Open Source Softw 2019;40:1541.
28. Fay C, Guyader V, Rochette S, Girvard C. golem: a framework for robust shiny applications; 2022. R package version 0.3.5.
29. Gelman A, Hill J, Vehtari A. Regression and other stories (analytical methods for social research). Cambridge: Cambridge University Press; 2020.
30. McElreath R. Statistical rethinking: a Bayesian course with examples in R and Stan. Florida, FL: CRC Press; 2020.
31. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner PC. Rank-normalisation, folding, and localisation: an improved for assessing convergence of MCMC. Bayesian Anal 2021;16:667–718.
32. International Organization for Standardization. Medical laboratories: requirements for quality and competence (ISO Standard No. 15189:2022; 2022. Available from: https://www.iso.org/standard/76677.html.
33. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983;32:307–17.
34. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Stat Sin 2002;12:111–39.
35. Harrel F. Statistical thinking – classification vs prediction. Available from: https://www.fharrell.com/post/classification/ [Accessed 19 Jan 2023].
36. Harrell F. Statistic thinking – clinicians' misunderstanding of probabilities makes them like backwards probabilities such as sensitivity, specificity, and type I error. Available from: https://www.fharrell.com/post/backwards-probs/ [Accessed 19 Jan 2023].
37. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.
38. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res 2019;3s. https://doi.org/10.1186/s41512-019-0064-7.