ShinyPrior: A tool for estimating probability distributions from published evidence

Nicole M White[1] and Robin Blythe[1]

[1]Australian Centre for Health Services Innovation and Centre for Healthcare Transformation, School of Public Health & Social Work, Queensland University of Technology, Australia

February 10, 2023

1 Introduction

Health economic modelling uses quantitative methods to estimate the costs and consequences associated with changes in healthcare delivery [1]. As a framework for decision-making, modelling aims to synthesise the best available evidence about one or more interventions, compared with current policy or practice. Ideally, evidence needed to populate a health economic model is obtained from a single source, for example, data collected as part of a clinical trial [2]. However, access to these data is not always feasible. When primary data are not available, published estimates are commonly used to define model inputs, either alone or in combination with other sources [3]. Examples of model inputs that published estimates may inform include probabilities of a patient experiencing a clinical event (e.g, disease relapse), the estimated effectiveness of the intervention(s) being evaluated, and resource costs.

Whilst easily accessible, incorporating published data into health economic models can be challenging. First, data sourced from published studies are typically available only as summary statistics, for example, from tables or quoted in the main text of an article. Common choices for summary statistics include means with standard deviations or standard errors, medians with quantiles, and confidence intervals. Differences in reported summary statistics between studies, or even within the same study, make it difficult to standardise estimates for use as model inputs. Second, published estimates at best approximate the true value of model inputs within the target population being evaluated [4, 5]. Sensitivity associated with chosen estimates on modelling outcomes should therefore be assessed. The Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist recommends that uncertainty be characterised by assigning probability distributions to key model inputs [6]. Defined probability distributions can then be used to assess the downstream effects of uncertainty on model outcomes [7]. Depending on the form of data available, different methods are needed to infer the parameters of appropriate probability distributions with reasonable accuracy.

ShinyPrior is a web-based application for estimating probability distributions from summary statistics. The application implements appropriate methods to convert different summary statistics into distributions commonly assumed for continuous variables in health economic models. The application was developed using the `shiny` [8] and `shinydashboard` [9] packages available in R [10], and is available at `https://aushsi.shinyapps.io/ShinyPrior`.

The purpose of this article is to outline the main steps involved when using ShinyPrior. Each step is represented by a separate menu displayed on the left-hand side of the application:

- **Define distribution inputs** (Section 2): The user specifies the distribution family and form of evidence available to estimate distribution parameters
- **Customisation** (Section 3): Provides options to summarise one or more estimated distributions, to produce publication-ready figures and tables.

2 Define distribution inputs

2.1 Distribution family

Table 1 provides further details about supported distributions. Assumed parameterisations align with the R help documentation. Users wishing to use application outputs in other software are advised to check software-specific parameterisations to ensure consistency across platforms. Guidance on using application outputs to simulate random variates in R, STATA, Microsoft Excel and TreeAge is provided in the Appendix.

Distribution family	**Parameterisation**	**Forms of evidence**
Normal, $\mathcal{N}(\mu, \sigma)$	Mean: $-\infty < \mu < \infty$	Mean with uncertainty
	Standard deviation: $\sigma > 0$	Percentiles
Gamma, $\mathcal{G}(a, b)$	Shape: $a > 0$	Mean with uncertainty
	Scale: $b > 0$	Percentiles ‡
log-Normal, $\mathcal{LN}(\mu, \sigma)$	Mean: $-\infty < \mu < \infty$	Mean with uncertainty
	Standard deviation: $\sigma > 0$	Percentiles
Weibull, $\mathcal{W}(a, b)$	Shape: $a > 0$	Mean with uncertainty‡
	Scale: $b > 0$	Percentiles
Beta, $\mathcal{B}(a, b)$	Shape: $a > 0$	Mean with uncertainty
	Scale: $b > 0$	Percentiles‡
		Number of events, sample size
Uniform, $\mathcal{U}(l, u)$	Minimum: l	Mean with Uncertainty
	Maximum: u	Percentiles
	$-\infty < l < u < \infty$	Minimum, Maximum

Table 1: Available distributions and forms of evidence supported. ‡ denotes parameter estimation by numerical optimisation; otherwise, closed-form solutions are used.

2.2 Description

ShinyPrior was designed to summarise multiple distributions within the same session. This functionality allows users to compare different distributions, and to create figures and tables for use in research outputs. Figure and table options are described in Section 3. To support this functionality, users must supply a text label in the *Description* box for each unique distribution. Supplied descriptions are not subject to a character limit, however, we recommend a short label of up to 20 characters. Entered labels are used to index all application outputs, and are displayed in both the Visualisation and Summary table panes. An error message will appear if a user selects *Estimate distribution* before supplying a description. If an existing description is supplied for a new distribution, the previous result matching the same description will be overwritten.

2.3 Form of evidence

Distribution parameters are estimated from the form of evidence selected by the user. Forms of evidence reflect commonly used summary statistics for describing unbounded and bounded continuous variables. All distributions include the options "Mean with uncertainty", and "Percentiles". Additional available options will update based on the distribution selected from the *Distribution family* dropdown menu. Further guidance on each option is provided below:

- Mean with uncertainty: Uncertainty is defined on the standard deviation scale and must be greater than 0. Users should therefore ensure that any appropriate transformations are applied to summary statistics before using this option; e.g., sample variance, standard error.
- Percentiles: Examples of summary statistics based on percentiles include confidence intervals and interquartile ranges. This option requires the user to specify a lower and upper value, corresponding to the lower and upper limits of the percentile interval. Percentile coverage

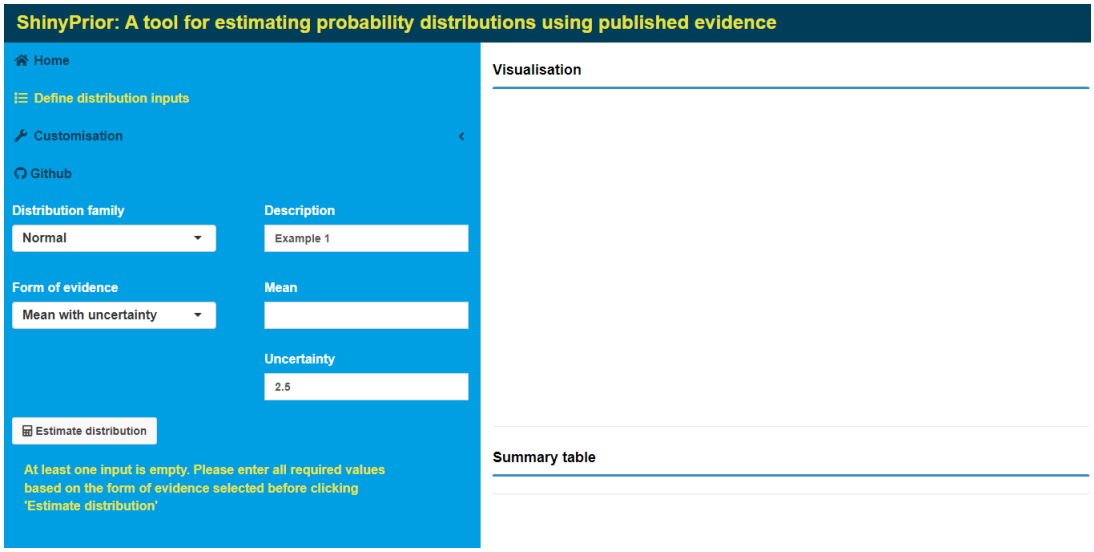

Figure 1: Example error message generated after supplying incorrect inputs. In this case, the Mean value is missing

must be provided as a percentage between 0 and 100%. The percentage coverage value is used to determine the distribution percentiles corresponding to the lower and upper values. For example, a 95% confidence level assumes the lower and upper values represent the 2.5^{th} and 97.5^{th} percentiles of the distribution, respectively. Similarly, an interquartile range assumes the lower and upper values represent the 25^{th} and 75^{th} percentiles.

- Number of events, sample size: For the Beta distribution only. The number of events must be greater than 0 and less than the defined sample size.
- Minimum and Maximum: For the Uniform distribution only.

An error message will appear if incompatible values are entered or if one or more inputs are missing. Errors messages include guidance on feasible values, as appropriate. An example error message is shown in Figure 1.

2.4 Estimate distribution

Supplied inputs are used to estimate distribution parameters based on closed-form solutions, when available, or by numerical optimisation. Numerical optimisation methods are pre-specified within the application based on the number of parameters without a closed-form solution. Numerical optimisation of a single unknown parameter is conducted using one-dimensional root finding using the `uniroot()` function, conditional on the closed-form estimate of the other parameter. In the case of two unknown parameters, a quasi-Newton routine is implemented using the general optimisation function `optim()`, from the `stats` package. Optimisation aims to minimise the sum of squares between expected and observed values.

Results for the current distribution are displayed in the Visualisation and Summary table windows (Figure 2). Previous results can be added via the Customisation menu.

3 Customisation

ShinyPrior offers several options for customising the appearance of density plots and summary table information. Results can be customised for a single distribution result, or for multiple results simultaneously. For all outputs, the distribution(s) of interest are specified in the *Select distribution(s)* box.

Outputs can be exported anytime into selected file formats for future use. The application saves all results for use in figures and tables until the user deletes them (see Section 3.3).

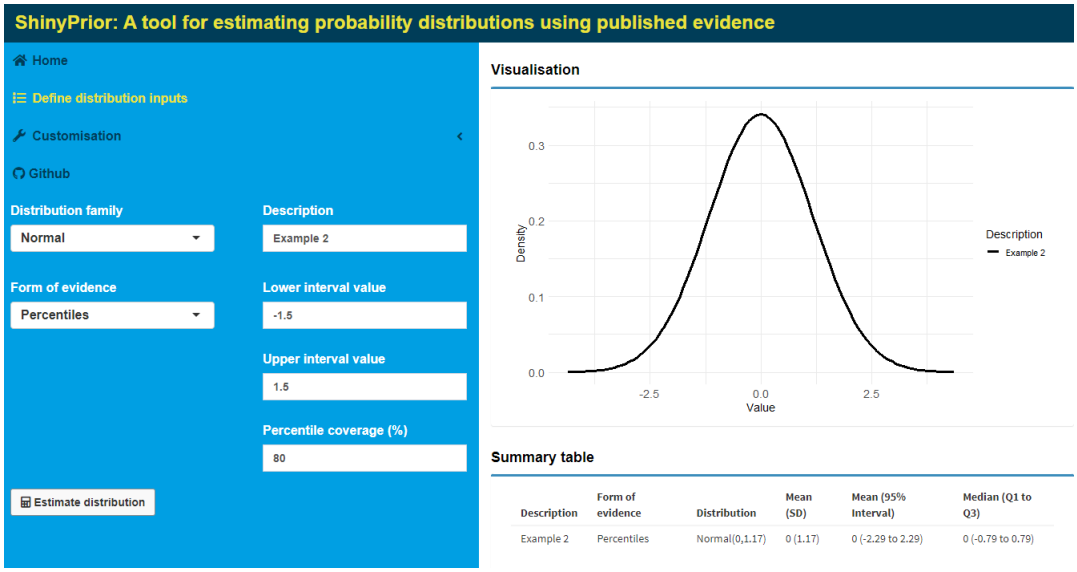

Figure 2: Example application display for a single distribution. Outputs are generated after specifying all inputs under *Define distribution inputs* and clicking *Estimate distribution*.

3.1 Visualisation

Density plots are created using the `ggplot2` package [11]. Options to customise plot output are colour palette (*Choose colour scheme*), plot theme (*Select plot theme*), axis/legend labels (*x-axis label*, *y-axis label*, *Legend title*) and legend display (*Display legend*). See Table 2 for a description of plot elements.

Plot element	**Format**	**Options available**
Colour scheme	Dropdown selection	Greyscale ‡, Accent, Dark2, Paired, Set1, Set2, Set3, Colourblind-1, Colourblind-2
Plot theme	Dropdown selection	Minimal ‡, Light, Black/White, Classic, Gray
x-axis label	Free-text	–
y-axis label	Free-text	–
Display legend	Radio button	Yes ‡, No
Legend title	Free-text	–
Format	Dropdown selection	.png ‡, .tiff, .jpeg
Resolution	Dropdown selection	300 dpi ‡, 600 dpi
Height, cm	Numeric	–
Width, cm	Numeric	–
Figure name	Free-text	–

Table 2: Figure options. ‡ denotes the default option for a plot element.

Nine pre-defined palettes are available for use [12, 13]. All colour palettes include up to eight colours. The greyscale scheme includes a maximum of five colours. Hexidecimal values by colour scheme are given in Table 3.

When multiple distributions are displayed, legend colours and label ordering will match the order of results as entered in the *Select distributions(s)* box. An example is shown in Figure 3

The following ggplot2 themes are supported: `theme_minimal()` (default), `theme_light()`, `theme_bw()`, `theme_classic()`, and `theme_gray()`.

Axis labels and the legend title can be updated in the corresponding free-text boxes. The figure legend can be included or excluded by selecting the appropriate option under *Display legend*. If

Colour scheme	Hexidecimal values
Greyscale	#000000, #737373, #BDBDBD, #D9D9D9, #F0F0F0
Accent	#7FC97F, #BEAED4, #FDC086, #FFFF99, #386CB0, #F0027F, #BF5B17, #666666
Dark2	#1B9E77, #D95F02, #7570B3, #E7298A, #66A61E, #E6AB02, #A6761D, #666666
Paired	#A6CEE3, #1F78B4, #B2DF8A, #33A02C, #FB9A99, #E31A1C, #FDBF6F, #FF7F00
Set1	#E41A1C, #377EB8, #4DAF4A, #984EA3, #FF7F00, #FFFF33, #A65628, #F781BF
Set2	#66C2A5, #FC8D62, #8DA0CB, #E78AC3, #A6D854, #FFD92F, #E5C494, #B3B3B3
Set3	#8DD3C7, #FFFFB3, #BEBADA, #FB8072, #80B1D3, #FDB462, #B3DE69, #FCCDE5
Colourblind-1	#000000, #E69F00, #56B4E9, #009E73, #F0E442, #0072B2, #D55E00, #CC79A7
Colourblind-2	#999999, #E69F00, #56B4E9, #009E73, #F0E442, #0072B2, #D55E00, #CC79A7

Table 3: Hexidecimal values by colour scheme.

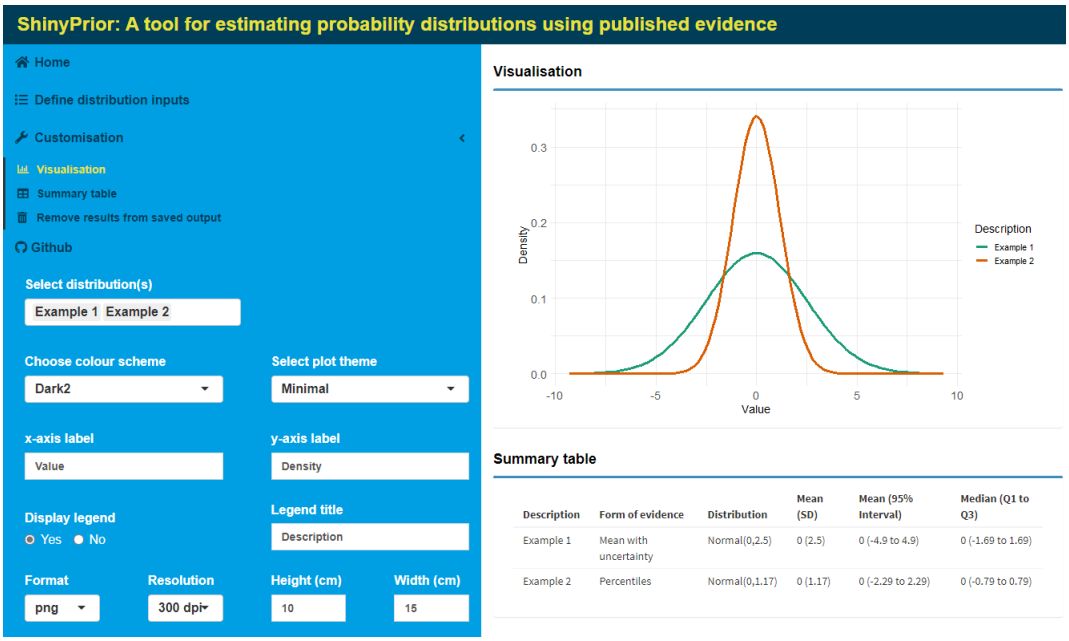

Figure 3: Example figure for two estimated distributions, using the Dark2 colour palette

"Yes" is selected, the legend will appear on the right-hand side of the figure.

Users can further specify the saved figure size (*Height (cm)*, *Width (cm)*), figure resolution (*Resolution*), and file format (*Format*). Plots can be saved at 300 or 600 dots per inch (dpi), as a .png, .tiff or .jpeg file. Custom filenames are allowed and can be specified in the *Figure name* box.

3.2 Summary table

Distribution summary statistics are presented in tabular form, generated using the `flextable` package [14]. Customisation options include the inclusion/exclusion of columns, and row ordering by *Description* or *Distribution family*.

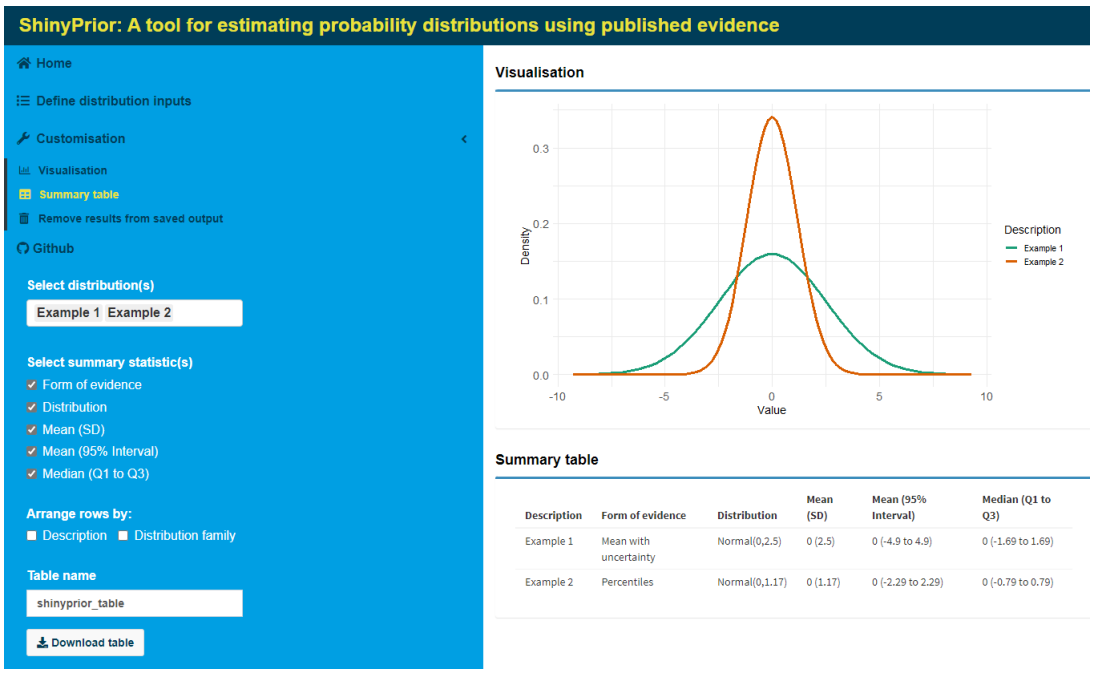

Figure 4: Summary table options for the example shown in Figure 3

All table columns will appear in the Summary table window by default (Figure 4). Available columns are: Description, Form of evidence, Distribution, Mean (SD), Mean (95% Interval), and Median (Q1 to Q3). The Description column cannot be removed within the application. Otherwise, any number of columns may be included or excluded from the final table via the checkboxes under the Summary table sub-menu. Similar to figures, a custom filename may be entered in the *Table name* box. In the current version of ShinyPrior, tables can only be exported as Word documents (.docx).

3.3 Remove results from output

Users can permanently delete results for one or more distributions at any time. Results to be deleted are specified in the *Select distribution(s)* box. Once entered, confirm the selection by clicking *Remove selected distribution(s).*

4 Acknowledgements

We thank Hannah Carter and Adrian Barnett for their contributions that improved earlier versions of the application, and David Borg for feedback on the accompanying vignette.

References

[1] Michael F Drummond, Mark J Sculpher, Karl Claxton, Greg L Stoddart, and George W Torrance. *Methods for the economic evaluation of health care programmes*. Oxford university press, 2015.

[2] Amy K O'Sullivan, David Thompson, and Michael F Drummond. Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations? *Value in Health*, 8(1):67–79, 2005.

[3] Pedro Saramago, Andrea Manca, and Alex J Sutton. Deriving input parameters for cost-effectiveness modeling: taxonomy of data types and approaches to their statistical synthesis. *Value in Health*, 15(5):639–649, 2012.

[4] Andrew H Briggs, Milton C Weinstein, Elisabeth AL Fenwick, Jonathan Karnon, Mark J Sculpher, A David Paltiel, ISPOR-SMDM Modeling Good Research Practices Task Force, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value in Health*, 15(6):835–842, 2012.

[5] Andrew H Briggs. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics*, 17: 479–500, 2000.

[6] Don Husereau, Michael Drummond, Federico Augustovski, Esther de Bekker-Grob, Andrew H Briggs, Chris Carswell, Lisa Caulley, Nathorn Chaiyakunapruk, Dan Greenberg, Elizabeth Loder, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 explanation and elaboration: a report of the ISPOR CHEERS II good practices task force. *Value in health*, 25(1):10–31, 2022.

[7] Karl Claxton, Mark Sculpher, Chris McCabe, Andrew Briggs, Ron Akehurst, Martin Buxton, John Brazier, and Tony O'Hagan. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health economics*, 14(4):339–347, 2005.

[8] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2020. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.5.0.

[9] Winston Chang and Barbara Borges Ribeiro. *shinydashboard: Create Dashboards with 'Shiny'*, 2021. URL `https://CRAN.R-project.org/package=shinydashboard`. R package version 0.7.2.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[11] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

[12] Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2022. URL `https://CRAN.R-project.org/package=RColorBrewer`. R package version 1.1-3.

[13] Winston Chang. *R Graphics Cookbook*, chapter Using Colors in Plots. O'Reilly Media, 2013.

[14] David Gohel and Panagiotis Skintzos. *flextable: Functions for Tabular Reporting*, 2022. URL `https://CRAN.R-project.org/package=flextable`. R package version 0.8.2.

Appendix

Distribution	Software	Parameters	Random variate generation
Normal	R	`mean`$= \mu$, `sd` $= \sigma$	`rnorm(n = 1,mean,sd)`
$\mathcal{N}(\mu, \sigma)$	STATA	`m`$= \mu$, `s` $= \sigma$	`rnormal(m,s)`
	Excel	`mean`$= \mu$ `standard_dev` $= \sigma$	`NORM.INV(RAND(),mean,standard_dev)`
	TreeAge	`Mean` $= \mu$, `std dev` $= \sigma$	–
Gamma	R	`shape` $= a$, `scale` $= b$	`rgamma(n = 1,shape,scale)`
$\mathcal{G}(a, b)$	STATA	`a` $= a$, `b` $= b$	`rgamma(a,b)`
	Excel	`alpha` $=$ a,`beta` $=$ b	`GAMMA.INV(alpha,beta)`
	TreeAge	`Alpha` $= a$, `Lambda` $= b$	Alternative parameters
log-Normal	R	`meanlog` $= \mu$, `sdlog` $= \sigma$	`rlnorm(n = 1,meanlog,sdlog)`
$\mathcal{LN}(\mu, \sigma)$	STATA	`m`$= \mu$, `s` $= \sigma$	`exp(rnormal(m,s))`
	Excel	`mean` $= \mu$ `standard_dev` $= \sigma$	`LOGNORM.INV(RAND(),mean,standard_dev)`
	TreeAge	`mean of logs` $= \mu$ `std. dev of logs` $= \sigma$	Alternative parameters
Weibull	R	`shape` $=$ a, `scale` $=$ b	`rweibull(n = 1,a,b)`
$\mathcal{W}(a, b)$	STATA	`a` $= a$, `b` $= b$	`rweibull(a,b)`
	Excel	`a` $= a$, `b` $= b$	`a*(-LN(1-RAND()))^(1/b)`
	TreeAge	`Scale` $= a$, `Shape` $= b$	Alternative parameters
Beta	R	`shape1` $= a$, `shape2` $= b$	`rbeta(n = 1,shape1,shape2)`
$\mathcal{B}(a, b)$	STATA	`a` $= a$, `b` $= b$	`rbeta(a,b)`
	Excel	`alpha` $= a$, `beta` $= b$	`BETA.INV(RAND(),alpha,beta)`
	TreeAge	`Alpha` $= a$, `Beta` $= b$	Alternative parameters
Uniform	R	`min` $= l$, `max` $= u$	`runif(n = 1,min,max)`
$\mathcal{U}(l, u)$	STATA	`a` $= a$, `b` $= b$	`runiform(a,b)`
	Excel	`bottom` $= l$, `top` $= u$	`RANDBETWEEN(bottom,top)`
	TreeAge	`Min` $= l$, `Max` $= u$	–

Table 4: ShinyPrior parameterisations mapped to R, STATA, Microsoft Excel and TreeAge for supported distributions. Code for random variate generation is provided for R, STATA and Excel. For TreeAge, the parameterisation is provided when multiple options are available.