Reproducibility of Statistical Tests Based on Randomised Response Data

Fatimah M. Alghamdi[1], Frank P. A. Coolen[2] and Tahani Coolen-Maturi[2*]

[1]Department of Mathematical Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11564, Saudi Arabia.
[2]Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, United Kingdom.

*Corresponding author. E-mail: tahani.maturi@durham.ac.uk;
Contributing authors: fmalghamdi@pnu.edu.sa;
frank.coolen@durham.ac.uk;

Abstract

Reproducibility of experimental conclusions is an important topic in various fields, including social studies. The lack of reproducibility in research results not only limits scientific progress but also wastes time, resources, and undermines society's confidence in scientific findings. This paper focuses on the statistical reproducibility of hypothesis test outcomes based on data collected using randomised response techniques (RRT). Nonparametric predictive inference (NPI) is used to quantify reproducibility, which is well-suited to treat reproducibility as a prediction problem. NPI relies on few model assumptions and provides lower and upper bounds for reproducibility probabilities. This paper concludes that less variability in the reported responses of RRT methods leads to higher reproducibility of statistical hypothesis tests based on RRT data with the same degree of privacy.

Keywords: Reproducibility Probability, Nonparametric Predictive Inference, Randomised Response Data, Greenberg Method, Forced Method.

1 Introduction

In statistics, reproducibility refers to the ability to reproduce a study's conclusions if the study is repeated in the same way. Science depends substantially on reproducibility to ensure that its findings are valid. Goodman [1] emphasised the importance of the statistical reproducibility challenge for investigations. He argued that p-values have been inaccurately used and misunderstood in research, providing the results a misleading aspect of confidence and ability for generalisation. He pointed out that p-values do not indicate effect size or reproducibility probability, which are crucial to research. Goodman [1] advocated a more detailed and open approach to statistical inference, one that involves effect sizes, confidence intervals, reproducibility probability, and other measures. Senn [2] agrees with Goodman that the p-value and reproducibility probability are separate measurements. He did not agree with Goodman's claim that the p-value overstates the strength of the evidence against the null hypothesis; however, Senn [2] argued that there is a connection between the p-values and reproducibility probability.

Coolen and BinHimd [3] presented NPI for the reproducibility of some basic tests. Wilcoxon's signed rank test and the two sample rank sum test were used to introduce nonparametric predictive inference (NPI) for reproducibility probability (RP) [3]. NPI for Bernoulli quantities [4] and for real-valued data [5] were both used for these inferences. They produced NPI lower and upper reproducibility probabilities, $\underline{RP}$ and $\overline{RP}$, instead of precise values. The NPI-bootstrap approach, as developed and demonstrated by BinHimd [32] for the Kolmogorov-Smirnov test, can be used to provide NPI for more complex test situations.

In order to increase the validity of scientific findings, Billheimer [6] emphasises the significance of predictive inference and scientific reproducibility. Additionally, Billheimer argues that predictive inference provides a suitable method for inference on reproducibility by taking the distribution of future data into account. Next, he models the predictive distribution for the next observation, X_{n+1}, given the original observation X_n, and uses the de Finetti representation theorem [7]. His viewpoint is that parametric modelling is a useful approximation of the prior distribution of either parameters or possible observations, with parameter choices only affecting the distribution of future observables. Additionally, findings or actions based on predictions should be evaluated in the context of the research problem. He also refers to the importance of the predictive inference method which encourages statisticians to characterise interesting findings using observable quantities and predict the probability of them in future studies.

This paper reports the first study of reproducibility of statistical inferences based on data collected using RRT methods. The RRT methods used can be considered to be classical methods, which seemed to be a good starting point to explore aspects of reproducibility. In recent years, many important contributions have been published on RRT methodology, leading to quite a wide variety of RRT methods. While these mostly still build on the classic

ideas, they consider, for example, more explicitly the respondent privacy and efficiency [40, 41]. It will be of great importance to study reproducibility of inferences based on data using such more modern RRT methods in the future.

Thus, in this paper, we study the reproducibility of statistical tests based on data collected from randomised response techniques (RRT). These RRTs can be used in social studies to elicit a true response to sensitive questions, which can be an effective method to determine the proportion of sensitive characteristics. We define the reproducibility probability (RP) of a test as the probability that the test result, whether the null hypothesis is rejected or not, will be the same if the test is repeated using an experiment done in the same way as the original experiment.

This paper is organised as follows. Section 2 introduces the RRT methods for the study in this paper. Nonparametric Predictive Inference (NPI) is demonstrated in Section 3. Section 4 explains NPI for RP (NPI-RP) for one-sided tests. Section 5 introduces a measure of reproducibility probability (MRP) and presents a comparison of the reproducibility of hypothesis tests using data collected by RRT. Section 6 presents a discussion of related topics for further research.

2 Randomised response techniques (RRT)

Randomised response techniques (RRT) are used to avoid possible embarrassment when respondents are asked sensitive questions. A spinner, a deck of cards, or a coin can be used as a randomisation device, and the responses are hidden from the interviewer. These methods help individuals to maintain their privacy. There are two basic RRT method approaches: qualitative randomised response techniques using 'Yes' or 'No' responses and quantitative randomised response methods using real numbers. In this paper, we only consider qualitative RRT methods.

2.1 Qualitative randomised response techniques

In this section, we introduce qualitative RRT for surveys in which sensitive questions are answered using qualitative binary response variables, typically 'Yes' or 'No'. Warner [8] presented the first RRT method, which we refer to as the Warner Method (WM). Suppose that we want to estimate the proportion π of a population who have a sensitive characteristic A using the WM method. In this method, there are two questions, Q_1 and Q_2, to determine if the respondent is in the target group A (they have the sensitive characteristic) or if they do not have the sensitive characteristic so they belong to $\bar{A}$, as follows:

$$Q_1 : \text{Are you a member of group } A?$$
$$Q_2 : \text{Are you a member of group } \bar{A}?$$

Assume that a sample of size n is selected from a target population, and there is a randomisation device which helps respondents to choose the question. Suppose that with probability γ, the respondent is asked question Q_1, which is sensitive, and with probability $1-\gamma$, the respondent is asked question Q_2, which is also sensitive, where γ is known to the interviewer. As a result, the number of people who get question Q_1 is Binomially distributed with sample size n and parameter γ. Each response is either Yes ($\dot{Y}$) or No ($\dot{N}$). The probability of a 'Yes' answer is:

$$P_W^* = \gamma\pi + (1-\gamma)(1-\pi) \tag{1}$$

Warner [8] suggested that the probability of a sensitive question in the randomisation device should be greater than 0.5. The reason for this choice is that if $\gamma = 0.5$, then the probability of respondent i saying 'Yes' will not depend on π in Equation (1), so, the response would provide no information about π. If $\gamma = 1$, we just return to the non-RRT method and use the direct question. If we choose $0.5 < \gamma < 1$ or $0 < \gamma < 0.5$, the respondent provides a useful response, and the respondent does not reveal to which group they belong [8].

Assume that Y is the Binomial random quantity of the number of 'Yes' responses to the chosen question where $Y \sim Bin(n, P_W^*)$ and $Y \in \{0, 1, \ldots, n\}$. Then, the expected value of Y is $\mathrm{E}(Y) = nP_W^*$, and the estimator $\hat{\pi}(Y)$ of the proportion π of people who have the sensitive characteristic is

$$\hat{\pi}(Y) = \frac{n(\gamma-1)+Y}{(2\gamma-1)n} \quad \text{where} \quad 0 \le \gamma \le 1, \gamma \ne \frac{1}{2} \tag{2}$$

The expectation of the estimator $\hat{\pi}(Y)$ [8] is

$$\mathrm{E}(\hat{\pi}(Y)) = \mathrm{E}\left[\frac{n(\gamma-1)+Y}{(2\gamma-1)n}\right] = \pi \tag{3}$$

So $\hat{\pi}(Y)$ is an unbiased estimator of π. The variance of the estimator $\hat{\pi}(Y)$ is [9]:

$$\mathrm{Var}(\hat{\pi}(Y)) = \frac{(\pi-\pi^2)}{n} + \frac{\gamma(1-\gamma)}{n(2\gamma-1)^2} \quad \text{where} \quad 0 \le \gamma \le 1,\ \gamma \ne \frac{1}{2} \tag{4}$$

The first term in Equation (4) is the binomial variance related to the sensitive question. The second term is the extra variance due to the uncertainty caused by using a randomisation device.

The Greenberg technique [9] and the Forced Method [10] are other RRT approaches for binary responses which are used in this paper. The Greenberg Method (GB) [9] is a variation of the WM method [8] in which respondents are also randomly asked one of two questions using the randomisation device. Assume that we have a sample of size n, and random quantity Y is the number

of 'Yes' answers to the chosen question. Let A represent the sensitive characteristic of interest, whereas B denotes a neutral characteristic that is unrelated to A. The unrelated question aims to encourage respondents to answer the selected question truthfully. Let π_A and π_B represent the proportions of individuals belonging to groups A and B, respectively. If both proportions π_A and π_B are unknown, it is necessary to choose two independent samples from the population, assuming a basic random sampling with replacement method and two separate methods of randomisation are used for the two samples. If π_B is only known, only one sample and decks of cards are needed. Each card contains either a sensitive or an unrelated question, which occurs with probability γ and $1-\gamma$, respectively. Each question can result in one of two possible answers: a Yes ($\dot{Y}$) or a No ($\dot{N}$). The two questions could be:

$$\begin{aligned} Q_1 &: \text{ Are you a member of group } A? \\ Q_2 &: \text{ Are you a member of group } B? \end{aligned}$$

Then, the probability of the event that a person answers 'Yes' to the selected question of the GB Method is

$$P_G^* = \gamma\pi_A + (1-\gamma)\pi_B \tag{5}$$

Note that, as for WM, in applying GB, the interviewer is unaware of the question being asked. It is preferable to choose π_B close to zero [11]. The estimator $\hat{\pi}_A(Y)$ of proportion of people who have the sensitive characteristic is

$$\hat{\pi}_A(Y) = \frac{\frac{Y}{n} - \pi_B(1-\gamma)}{\gamma} \tag{6}$$

Using Bayes' rule, the conditional probabilities that the respondent belongs to groups A or B are calculated as follows:

$$P_G(A|\ \dot{Y}) = \frac{\pi_A P_G(\dot{Y}|\ A)}{P_G^*} \tag{7}$$

where $P_G^*(\dot{Y}|\ A) = \pi_B + (1-\pi_B)\gamma$, and $P_G^*(\dot{Y}|\ B) = \pi_B(1-\gamma)$. The expected value of the estimator $\hat{\pi}_A(Y)$ is

$$\begin{aligned} \mathrm{E}(\hat{\pi}_A(Y)) &= \mathrm{E}\left(\frac{\frac{Y}{n} - (1-\gamma)\pi_B}{\gamma}\right) = \frac{P_G^* - (1-\gamma)\pi_B}{\gamma} \\ &= \frac{\gamma\pi_A + (1-\gamma)\pi_B - (1-\gamma)\pi_B}{\gamma} = \pi_A \end{aligned} \tag{8}$$

So, $\hat{\pi}_A(\mathrm{Y})$ is an unbiased estimator of the population proportion π_A. The variance of $\hat{\pi}_A(Y)$ is [12]:

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_A(Y)) &= \mathrm{Var}\left(\frac{P_G^* - (1-\gamma)\pi_B}{\gamma}\right) \\ &= \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-\gamma)^2\pi_B(1-\pi_B) + \gamma(1-\gamma)(\pi_A + \pi_B - 2\pi_A\pi_B)}{n\gamma^2}\end{aligned} \tag{9}$$

where $0 < \gamma \leq 1$ and $\gamma \neq \frac{1}{2}$, using that $\mathrm{Var}\left(\frac{(1-\gamma)\pi_B}{\gamma}\right) = 0$ because γ and π_B are constants.

The Forced Method (FM) [10] is another RRT method, where the randomisation device forces the respondent to answer 'Yes' to the selected question with probability γ_1, or 'No' with probability γ_2, or to answer the sensitive question with probability γ, where $\gamma = 1 - \gamma_1 - \gamma_2$ and $0 < \gamma_1 < 1$, $0 < \gamma_2 < 1$ and $\gamma_1 + \gamma_2 < 1$ [10]. Each response can result in one of two possible outcomes: a Yes ($\dot{Y}$) or a No ($\dot{N}$).

Assume a sample of size n, and random quantity Y is the number of people who answer 'Yes' to the sensitive question they are asked. The probability of a respondent answering 'Yes' is

$$P_F^* = \gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2) \tag{10}$$

where π_A is again the proportion of people who have the sensitive characteristic A. The estimator of π_A is

$$\hat{\pi}_A(Y) = \frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2} \tag{11}$$

Using Bayes' rule, the conditional probabilities of the event that the respondent belongs to groups A given the response 'Yes' or 'No' are

$$P_F(A|\ \dot{Y}) = \frac{\pi_A P_F(\dot{Y}|\ A)}{P_F^*} \tag{12}$$

$$P_F(\bar{A}|\ \dot{Y}) = \frac{\pi_{\bar{A}} P_F(\dot{Y}|\ \bar{A})}{P_F^*} \tag{13}$$

where $P_F^*(\dot{Y}|\ A) = 1 - \gamma_2$, and $P_F^*(\dot{Y}|\ \bar{A}) = \gamma_1$. The expected value of $\hat{\pi}_A(Y)$ is

$$\mathrm{E}(\hat{\pi}_A(Y)) = \mathrm{E}\left(\frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) = \frac{P_F^* - \gamma_1}{1 - \gamma_1 - \gamma_2}$$

$$= \frac{\gamma_1 + \pi_A(1-\gamma_1-\gamma_2) - \gamma_1}{1-\gamma_1-\gamma_2} = \pi_A \tag{14}$$

So $\hat{\pi}_A(\mathrm{Y})$ is an unbiased estimator of the population proportion π_A. The variance of the estimator $\hat{\pi}_A(Y)$ is [10]:

$$\begin{aligned}\mathrm{Var}(\hat{\pi}_A(Y)) &= \mathrm{Var}\left(\frac{\frac{Y}{n}-\gamma_1}{1-\gamma_1-\gamma_2}\right) = \mathrm{Var}\left(\frac{P_F^*}{n(1-\gamma_1-\gamma_2)^2}\right) \\ &= \frac{\pi_A(\pi_A-1)}{n} + \frac{\pi_A(\gamma_2-\gamma_1)}{n(1-\gamma_1-\gamma_2)} + \frac{\gamma_1(1-\gamma_1)}{n(1-\gamma_1-\gamma_2)^2}\end{aligned} \tag{15}$$

Other RRT methods have been proposed, each with specific procedures and assumptions, such as scenarios with studying reproducibility of statistical inference based on data collected by these methods is an interesting topic for future research, e.g. multiple randomisation devices [13–16]. Such methods are not considered in this paper.

2.2 RRT efficiency comparison and privacy degree

When applying RRT methods, the efficiency and degree of privacy need to be considered. The efficiency of randomised response methods refers to the ability of these methods to accurately estimate the proportion of individuals in a population who have a sensitive characteristic. An efficient randomised response method produces estimates that are close to the true proportion of people who have the sensitive characteristic. There are several measures of the efficiency of RRT methods [9, 17]. Since the basic attention in this research is the relationship between RP and the variation in reported RRT responses, we do not take efficiency measures into account. However, RRT methods are considered more efficient when reported responses have less variability.

Another fundamental challenge in RRT is how to provide accurate estimates of the population proportion of people with sensitive characteristics while maintaining respondents' privacy. Several privacy measures have been proposed for qualitative and quantitative randomised response methods, with different implications for optimal study design. When there is a high degree of privacy, respondents are more likely to participate in surveys and to answer truthfully. If respondents are satisfied with their privacy, they reduce bias resulting from false responses. Furthermore, protecting the privacy of respondents is an essential ethical consideration.

Privacy measures typically involve conditional probabilities for the event that the respondents have the sensitive characteristic A given the response 'Yes' or 'No' [18, 19]. Clearly, the higher the conditional probability of belonging to A, given a 'Yes' response, $P(A|\,\dot{Y})$, the more embarrassing it may be to provide that response even when the actual question being asked is unknown to the interviewer. One RRT method could be considered more useful than another if $\max(P(A|\,\dot{Y}), P(A|\,\dot{N}))$ of the first RRT method is smaller than for the second method [20].

Zhimin and Zaizai [12] presented a method to measure the privacy of RRT methods. To derive the privacy measure, remember that the conditional probabilities of the event that a respondent has the sensitive characteristic A given the response 'Yes' or 'No' are

$$P(A|\dot{Y}) = \frac{\pi_A P(\dot{Y}|A)}{P^*} \tag{16}$$

$$P(\bar{A}|\dot{Y}) = \frac{\pi_A P(\dot{Y}|\bar{A})}{P^*} \tag{17}$$

Then, the proposed privacy measure Δ is:

$$\Delta = \left|1 - \frac{1}{2}\left(\frac{P(\dot{Y}|\ A)}{P(\dot{Y}|\ \bar{A})} + \frac{P(\dot{N}|\ A)}{P(\dot{N}|\ \bar{A})}\right)\right| \tag{18}$$

where small values of Δ indicate a high privacy level because the conditional probabilities $P(A|\ \dot{Y})$ of the event that the respondents have the sensitive characteristic A given the response 'Yes' or 'No' are close to π_A, which means both $\frac{P(\dot{Y}|\ A)}{P(\dot{Y}|\ \bar{A})}$ and $\frac{P(\dot{N}|\ A)}{P(\dot{N}|\ \bar{A})}$ close to 1, that hence Δ is close to 0.

The privacy degrees Δ_{GB} of the Greenberg Method, and Δ_{FM} of the Forced Method as explained in Section 2.1 and using Equation (18) are

$$\Delta_{GB} = \left|\frac{\gamma(1 - 2\pi_B(1-\gamma))}{(2\pi_B(1-\gamma)(1-\pi_B(1-\gamma))}\right| \tag{19}$$

$$\Delta_{FM} = \left|\frac{\gamma_1(3 - 2\gamma_1) + \gamma_2(1 - 2\gamma_1) - 1}{2\gamma_1(1-\gamma_1)}\right| \tag{20}$$

We consider these measures together with reproducibility in Example 6.

3 Nonparametric Predictive Inference (NPI)

Nonparametric Predictive Inference (NPI) is a statistical method based on Hill's assumption $A_{(n)}$ [21], which provides direct conditional probabilities for a future observable random quantity based on observed values of related random quantities [4, 22]. To introduce the assumption $A_{(n)}$, suppose that there are $n+1$ real-valued random quantities, $Y_1, ..., Y_n, Y_{n+1}$. Assume that the ordered observed values of the random quantities $Y_1, ..., Y_n$ are denoted by $y_1 < y_2 < ... < y_n$, and define $y_0 = -\infty$ and $y_{n+1} = \infty$. The n observations split the real-line into $n+1$ intervals $I_i = (y_{i-1}, y_i)$, where $i = 1, ..., n+1$. The assumption $A_{(n)}$ [21] for one future observation Y_{n+1} is

$$P(Y_{n+1} \in I_i) = \frac{1}{n+1} \quad \text{for} \quad i = 1, ..., n+1 \tag{21}$$

$A_{(n)}$ is a post-data assumption related to exchangeability [7]. The lower and upper probabilities for the future observations $Y_{n+1} \in \mathfrak{A}$, for any $\mathfrak{A} \subset \mathbb{R}$, are [4, 22]:

$$\underline{P}(Y_{n+1} \in \mathfrak{A}) = \sum_{i=1}^{n+1} \mathbf{1}\{I_i \subseteq \mathfrak{A}\} P(Y_{n+1} \in I_i) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{I_i \subseteq \mathfrak{A}\} \tag{22}$$

$$\overline{P}(Y_{n+1} \in \mathfrak{A}) = \sum_{i=1}^{n+1} \mathbf{1}\{I_i \cap \mathfrak{A} \neq \emptyset\} P(Y_{n+1} \in I_i) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{I_i \cap \mathfrak{A} \neq \emptyset\} \tag{23}$$

where $\mathbf{1}\{E\}$ is the indicator function which is equal to 1 if event E is true and 0 otherwise.

In NPI, De Finetti's Fundamental Theorem of Probability [7] is used to determine optimal bounds for the probability of an event of interest involving Y_{n+1} [4], given the probabilities in Equation (21). This theory has strong consistency properties and provides reliable predictive results [4] in the theory of imprecise probability [23] and interval probability [24]. NPI has been introduced for several applications such as statistical process control [25, 26], the field of trading [27] and the area of finance [28, 29].

3.1 NPI for Bernoulli random quantities

In this paper, NPI for Bernoulli random quantities is used [5]. It is based on a latent variable representation of Bernoulli data. This presentation assumes underlying real-valued quantities and a threshold so that values on one side of the threshold are successes and values on the other side of the threshold are failures. The consecutive assumptions $A_{(n)}, ..., A_{(n+m-1)}$ are used for linking the m future observations to the n data observations.

Assume there is a sequence of $n + m$ exchangeable Bernoulli trials, each having the possible outcomes 'success' and 'failure', with data consisting of s successes in n trials. If Y_1^n denotes the random number of successes in trials 1 to n, then an adequate representation of the data for NPI is $Y_1^n = s$. Let Y_{n+1}^{n+m} denote the random number of successes in the future trials $n + 1$ to $n + m$. Let $R_t = \{r_1, r_2, ..., r_t\}$ with $1 \leq t \leq n + 1$ and integer values $0 \leq r_1 < r_2 < ... < r_t \leq m$. The NPI upper probability [5, 30] for the event $Y_{n+1}^{n+m} \in R_t$ given $Y_{n+1}^{n+m} = s$, for $s \in \{0, 1, ..., n\}$, is

$$\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = \binom{n+m}{n}^{-1} \sum_{j=1}^{t} \left[\binom{s - r_j}{s} - \binom{s - r_{j-1}}{s} \right] \binom{n - s + m - r_j}{n - s} \tag{24}$$

It is assumed that all $\binom{n+m}{n}$ orderings of the successes are equally likely. The corresponding NPI lower probability can be derived using the conjugacy property, that is $\underline{P}(A) = 1 - \overline{P}(A^c)$ for any event A and its complementary

event A^c, so

$$\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c | Y_1^n = s) \tag{25}$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

The NPI lower and upper probability for the events $Y_{n+1}^{n+m} \geq c$ and $0 \leq c \leq n$, are [5, 31]:

$$\underline{P}(Y_{n+1}^{n+m} \geq c \mid Y_1^n = s) = 1 - \binom{n+m}{n}^{-1} \times \left[\sum_{l=1}^{c-1} \binom{s+l-1}{s-1} \binom{n+m-s-l}{n-s} \right] \tag{26}$$

$$\overline{P}(Y_{n+1}^{n+m} \geq c \mid Y_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{s-c}{s} \binom{n+m-s-c}{n-s} + \sum_{l=c+1}^{n} \binom{s-l-1}{s-1} \binom{n+m-s-l}{n-s} \right] \tag{27}$$

where $s \in \{1, ..., n-1\}$. The minimum value of the NPI lower probability is 0.5, which happens when half of all orderings of the successes s based on the future test comes before the ordering of the successes r based on the original test due to the exchangeability assumption. This is shown in detail by Coolen and BinHimd [3]. The maximum value of the NPI upper probability is 1 for $Y_1^n = 0$ and $Y_1^n = n$, which occurs if all outcomes in the original test are failures or if all outcomes are successes, respectively [32].

If the observed data are all successes (so $s = n$) or all failure (so $s = 0$), then the NPI upper probabilities for this event $Y_{n+1}^{n+m} \geq c$ are:

$$\overline{P}(Y_{n+1}^{n+m} \geq c | Y_1^n = n) = 1 \tag{28}$$

$$\overline{P}(Y_{n+1}^{n+m} \geq c | Y_1^n = 0) = \binom{n+m}{n}^{-1} \binom{n+m-c}{n} \tag{29}$$

and the NPI lower probabilities for this event $Y_{n+1}^{n+m} \geq c$ are:

$$\underline{P}(Y_{n+1}^{n+m} \geq c | Y_1^n = n) = 1 - \binom{n+m}{n}^{-1} \binom{n+c-1}{n} \tag{30}$$

$$\underline{P}(Y_{n+1}^{n+m} \geq c | Y_1^n = 0) = 0 \tag{31}$$

3.2 NPI reproducibility

One important feature of practical research related to test results is the reproducibility of a given test. The concept and understanding of reproducibility have attracted more attention within the traditional frequentist statistical

framework, encouraging further research and academic interest in the past few years. The NPI method of frequentist statistics focuses explicitly on future observations while making few assumptions and using lower and upper probabilities to quantify uncertainty. This makes it possible to draw inferences about reproducibility probability (RP) given the explicitly predictive nature of NPI.

NPI reproducibility was first introduced by Coolen and BinHimd [5], denoted by NPI-RP, and defined as the probability that, if a test is repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is, whether the null hypothesis is rejected or not, will be the same. Coolen and BinHimd [5] considered a few basic nonparametric tests, namely the sign test, Wilcoxon's signed rank test, and the two sample rank sum test [33]. NPI for Bernoulli quantities [32] and for real-valued data [34] were used for these inferences. This led to NPI lower and upper reproducibility probabilities, denoted by $\underline{RP}$ and $\overline{RP}$, respectively, rather than precisely determined reproducibility probabilities.

The NPI-RP method has also been presented for two basic tests using order statistics [35]: a test for a specific population quantile value and a precedence test for comparing data from two populations. These latter tests are typically used for lifetime data experiments when one wishes to reach a conclusion before all observations are available. For these inferences, NPI for future order statistics is used to provide the lower and upper reproducibility probability for quantile and basic precedence tests [35].

More research has been published on NPI-RP, such as NPI for test reproducibility by sampling future data orderings [36]. In this work, Coolen and Marques investigated the NPI reproducibility of likelihood ratio tests using the test criterion in terms of the sample mean. This happens by taking into account all orderings of m future observations among the n data observations, all of which are equally likely based on an exchangeability assumption. However, because of the computing limitations of this method, exact lower and upper probability can only be computed for very small values of n. Then, the ordering sampling method is proposed to generate possible ordering of all samples for both exponential and normal distributions and it is examined how well it works to approximate the NPI lower and upper reproducibility probability.

Furthermore, another study examines reproducibility probability for likelihood ratio tests [37] between two Beta distributions. For simple hypotheses, the exact distribution is obtained using Gamma or Generalized Integer Gamma distributions. For more complex cases, near-exact or asymptotic approximations are developed using logarithm transformation and characteristic function. Numerical studies demonstrate the precision of the approximations, while simulations analyse test power and reproducibility probability.

More investigation introduces a statistical reproducibility for pairwise t-tests in pharmaceutical research using an NPI algorithm [38]. Simkus et al. [38] studied the statistical reproducibility of pairwise t-tests in pharmaceutical product development. They compared the reproducibility of t-tests and

Wilcoxon Mann-Whitney tests, and also considered the reproducibility of final decisions based on multiple related t-tests.

4 Reproducibility of one-sided hypothesis tests based on RRT data

Reproducibility of one-sided hypothesis tests based on data sampled using randomised response methods (NPI-RP-RRT) considers how likely it is that a future similar test of the null hypothesis will lead to the same conclusion as the original test. In this paper, we restrict attention to qualitative data collected using an RRT method. We consider the one-sided hypothesis test on the proportion π_A of people with a sensitive characteristic A:

$$H_0' : \pi_A = \pi_{A_0} \quad \text{versus} \quad H_1' : \pi_A > \pi_{A_0} \tag{32}$$

where $\pi_{A_0} \in [0, 1]$. Let $P_G^*(Y = \dot{Y} \mid H_0') = P_{G_0}^*$ be the probability of a 'Yes' answer to the selected question for the Greenberg method (GB) based on the proportion π_{A_0}, which is the proportion of people who have characteristic A. In this section, we use Equation (5) to link between π_{A_0} and $P_{G_0}^*$, and then investigate how the reproducibility probability is affected by π_{A_0} and $P_{G_0}^*$ under H_0. Therefore, the hypothesis test in P_G^*, corresponding to the hypothesis test using Equation (32), with level of significance $\alpha = 0.05$, is

$$H_0 : P_G^* = P_{G_0}^* \quad \text{and} \quad H_1 : P_G^* > P_{G_0}^* \tag{33}$$

This test can be performed based on the respondents' answers. A logical test rule is to reject the null hypothesis if $Y \geq c$, where c is determined, for chosen significance level α, as the minimal integer value for which:

$$P(Y \geq c \mid H_0) \leq \alpha \tag{34}$$

Let Y_1^n denote the random number of 'Yes' answers in the original sample and Y_{n+1}^{2n} denote the random number of 'Yes' answers in the future sample. The NPI upper and lower reproducibility probabilities for the event $Y_{n+1}^{2n} \geq c$, given $Y_1^n = y$, are

$$\overline{RP}(y) = \overline{P}(Y_{n+1}^{2n} \geq c \mid Y_1^n = y), \quad \underline{RP}(y) = \underline{P}(Y_{n+1}^{2n} \geq c \mid Y_1^n = y) \tag{35}$$

If the random number of 'Yes' answers in the original test 1 to n is less than c, so H_0' is rejected, then the upper and lower reproducibility probabilities for the event $Y_{n+1}^{2n} < c$ are:

$$\overline{RP}(y) = \underline{P}(Y_{n+1}^{2n} < c \mid Y_1^n = y), \quad \underline{RP}(y) = \overline{P}(Y_{n+1}^{2n} < c \mid Y_1^n = y) \tag{36}$$

Examples 1 and 2 illustrate this method.

Table 1 NPI-RP-GB at $\alpha = 0.05$, $c = 22$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9956	0.9980	22	0.5	0.6145
1	1.0000	1.0000	12	0.9909	0.9956	23	0.6145	0.7240
2	1.0000	1.0000	13	0.9824	0.9909	24	0.7240	0.8198
3	1.0000	1.0000	14	0.9680	0.9824	25	0.8198	0.8954
4	1.0000	1.0000	15	0.9449	0.9680	26	0.8954	0.9479
5	1.0000	1.0000	16	0.9101	0.9449	27	0.9479	0.9790
6	1.0000	1.0000	17	0.8605	0.9101	28	0.9790	0.9939
7	0.9999	1.0000	18	0.7941	0.8605	29	0.9939	0.9990
8	0.9997	0.9999	19	0.7102	0.7941	30	0.9990	1
9	0.9992	0.9997	20	0.6106	0.7102			
10	0.9980	0.9992	21	0.5	0.6106			

Table 2 NPI-RP-GB at $\alpha = 0.01$, $c = 23$.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9981	0.9992	22	0.5	0.6145
1	1.0000	1.0000	12	0.9959	0.9981	23	0.5	0.6195
2	1.0000	1.0000	13	0.9916	0.9959	24	0.6195	0.7340
3	1.0000	1.0000	14	0.9837	0.9916	25	0.7340	0.8333
4	1.0000	1.0000	15	0.9702	0.9837	26	0.8333	0.9097
5	1.0000	1.0000	16	0.9483	0.9702	27	0.9097	0.9601
6	1.0000	1.0000	17	0.9149	0.9483	28	0.9601	0.9872
7	1.0000	1.0000	18	0.8666	0.9149	29	0.9872	0.9977
8	0.9999	1.0000	19	0.8007	0.8666	30	0.9977	1
9	0.9997	0.9999	20	0.7163	0.8007			
10	0.9992	0.9997	21	0.6145	0.7163			

***Example** 1* This example explains NPI reproducibility for one-sided hypothesis tests based on data collected using the GB method (NPI-RP-GB). Suppose that we have a sample of size $n = 30$ and are interested in a sensitive characteristic A. The unknown proportion of people with the sensitive characteristic is $\pi_{A_0} = 0.7$, and $\pi_B = 0.3$ is the proportion of people who would respond 'Yes' to the unrelated question. In this example, we assume that a randomisation device is used with a probability $\gamma = 0.7$ that the sensitive question is asked. We want to test:

$$H_0' : \pi_A = 0.7 \quad \text{versus} \quad H_1' : \pi_A > 0.7 \tag{37}$$

with level of significance $\alpha = 0.05$. The hypothesis test on P_G^*, corresponding to the hypothesis test in Equation (37), is

$$H_0 : P_G^* = 0.58 \quad \text{versus} \quad H_1 : P_G^* > 0.58 \tag{38}$$

The corresponding threshold value for this one-sided test is $c = 22$ calculated using Equation (34). Therefore, H_0 is rejected at 0.05 level of significance if $Y_1^n \geq 23$. Then, the claim that the proportion of people who answer 'Yes' is 0.7 would be rejected at the 0.05 significance level.

The NPI lower and upper reproducibility probabilities for the event $Y_{n+1}^{2n} \geq c = 23$ under H_0 are presented in Table 1. The minimum value of the lower reproducibility probability is 0.5 as explained in Section 3.1. This happens for the values $y = 21$ and $y = 22$. Similarly, the NPI lower and upper reproducibility probabilities for the event

Table 3 NPI-RP-FM with $\alpha = 0.05$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.15$, $\gamma_2 = 0.10$, $c = 24$.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9993	0.9997	22	0.6195	0.7240
1	1.0000	1.0000	12	0.9983	0.9993	23	0.5	0.6195
2	1.0000	1.0000	13	0.9964	0.9983	24	0.5	0.6260
3	1.0000	1.0000	14	0.9925	0.9964	25	0.6260	0.7469
4	1.0000	1.0000	15	0.9854	0.9925	26	0.7469	0.8505
5	1.0000	1.0000	16	0.9731	0.9854	27	0.8505	0.9273
6	1.0000	1.0000	17	0.9527	0.9731	28	0.9273	0.9738
7	1.0000	1.0000	18	0.9210	0.9527	29	0.9738	0.9947
8	1.0000	1.0000	19	0.8742	0.9210	30	0.9947	1
9	0.9999	1.0000	20	0.8092	0.8742			
10	0.9997	0.9999	21	0.7240	0.8092			

Table 4 NPI-RP-FM with $\alpha = 0.01$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.15$, $\gamma_2 = 0.10$, $c = 25$.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9998	0.9999	22	0.7340	0.8198
1	1.0000	1.0000	12	0.9994	0.9998	23	0.6260	0.7340
2	1.0000	1.0000	13	0.9986	0.9994	24	0.5	0.6260
3	1.0000	1.0000	14	0.9969	0.9986	25	0.5	0.6347
4	1.0000	1.0000	15	0.9937	0.9969	26	0.6347	0.7642
5	1.0000	1.0000	16	0.9875	0.9937	27	0.7642	0.8729
6	1.0000	1.0000	17	0.9765	0.9875	28	0.8729	0.9486
7	1.0000	1.0000	18	0.9580	0.9765	29	0.9486	0.9881
8	1.0000	1.0000	19	0.9284	0.9580	30	0.9881	1
9	1.0000	1.0000	20	0.8837	0.9284			
10	0.9999	1.0000	21	0.8198	0.8837			

$Y_{n+1}^{2n} \geq c$ are presented in Table 2 for significance level 0.01, and the worst case for NPI lower reproducibility probability under the assumed model is 0.5 for the values $y = 22$ and $y = 23$.

If the original test leads to rejection of $H_0' : \pi_A = 0.7$ for the event $Y_1^n \geq c = 22$ at $\alpha = 0.05$, then the NPI reproducibility probability is the probability that the null hypothesis will also be rejected in the future test. Then, the NPI lower reproducibility and the NPI upper reproducibility probabilities for the event $Y_1^n > 22$ has the probability of $y > 23$: $\underline{RP}(y) = \overline{RP}(y-1)$ due to $\underline{P}(Y_{n+1}^{2n} \geq c | Y_1^n = y) = \overline{P}(Y_{n+1}^{2n} \geq c | Y_1^n = y-1)$. Conversely, if the reproducibility probability of Y_1^n which is less than the rejection threshold $c = 22$, the NPI lower and upper probabilities of the events $Y_1^n < 21$, which is $\underline{RP}(y) = \overline{RP}(y+1)$ for $Y_1^n < 21$ due to $\underline{P}(Y_{n+1}^{2n} < c | Y_1^n = y) = \overline{P}(Y_{n+1}^{2n} < c | Y_1^n = y+1)$.

In Tables 1 and 2, the NPI lower and upper reproducibility probabilities are presented and can be drawn as a line-segment between these values, based on data collected from the GB at significance level $\alpha = 0.05$ and $\alpha = 0.01$, with rejection threshold values 22 and 23 respectively. The larger value of the NPI lower and upper reproducibility probabilities suggest that a test gets the same outcome as the hypothesis test, with a probability close to 1.

***Example** 2* This example introduces the reproducibility probability for one-sided hypothesis tests with data collected using the Forced Method. Assume that a sample of size n is taken from a population with a possible sensitive characteristic A. Suppose that the H_0 value which we want to test is $\pi_{A_0} = 0.7$. The randomisation device leads to the sensitive question being asked with probability $\gamma = 0.75$, or the answer is forced to 'Yes' with probability $\gamma_1 = 0.10$ or forced to 'No' with probability $\gamma_2 = 0.15$. The significance level for the hypothesis test is $\alpha = 0.05$.

To start with, assume a sample with size $n = 30$, the null hypothesis that the proportion of people who have characteristic A is $H_0' : \pi_A = 0.7$, which is tested against $H_1' : \pi_A > 0.7$. So, the hypothesis test is

$$H_0' : \pi_A = 0.7 \quad \text{vs} \quad H_1' : \pi_A > 0.7 \tag{39}$$

Using Equation (10), this hypothesis test corresponds to the test:

$$H_0 : P_F^* = 0.625 \quad \text{vs} \quad H_1 : P_F^* > 0.625 \tag{40}$$

where the probability $P_{F_0}^*$ of a respondent saying 'Yes', using Equation (10), is

$$P_{F_0}^* = \gamma_1 + \pi_{A_0}(1 - \gamma_1 - \gamma_2) = 0.625 \tag{41}$$

The threshold value for this one-sided test is $c = 24$. Consequently, the null hypothesis H_0 is rejected when the observed value of Y_1^n is greater than or equal to 24; otherwise, the null hypothesis is not rejected. Similarly, the threshold value for this one-sided test is $c = 25$ at the significance level of 0.01. The NPI lower and upper probabilities for the event $Y_{n+1}^{2n} \geq c$, based on the FM data, are presented in Tables 3 and 4. The threshold values in Tables 3 and 4 are greater than the threshold values of reproducibility probability of statistical tests based on the GB data as presented in Tables 1 and 2.

As shown in Tables 3 and 4, the NPI lower and upper reproducibility probabilities based on FM data increase more than the NPI lower and upper reproducibility probabilities based on GB data as shown in Tables 1 and 2. In addition, the NPI lower reproducibility probabilities are closer to the NPI upper reproducibility probabilities based on FM data than the NPI lower reproducibility probabilities are closer to the NPI upper reproducibility probabilities based on GB data.

In general, the NPI lower and upper reproducibility probabilities based on FM data are greater than the NPI lower and upper reproducibility probabilities based on GB data, as shown in Tables 1 and 2 and Tables 3 and 4 respectively. Furthermore, the NPI lower reproducibility probabilities based on FM data are closer than the NPI upper reproducibility probabilities and the NPI lower reproducibility probabilities based on GB data.

5 A measure of reproducibility for statistical hypothesis tests

One objective of the study of reproducibility of hypothesis tests based on RRT methods is to compare RRT methods with regard to such reproducibility. This is non-trivial, particularly if the different RRT methods require different sample sizes to achieve a similar level of significance and power for a specific alternative hypothesis. In this section, we propose a new measure of reproducibility for such comparisons.

5.1 A measure of lower reproducibility for statistical hypothesis tests

The measure of the lower reproducibility probability under H_0 ($\text{MRP}_0^l(z)$) is the probability, under H_0, for the event that $\underline{RP}(Y) \geq z$, for $z \in [0,1]$. Therefore, with a sample of size n and probability P_0^* of a 'Yes' answer under H_0, MRP_0^l under H_0 for the one-sided test is

$$\begin{aligned} MRP_0^l(z) &= P(\underline{RP}(Y) \geq z | H_0) = P[\underline{RP}(Y) \geq z \ | Y \sim \text{Bin}(n, P_0^*)] \\ &= 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P_0^*)^y (1 - P_0^*)^{n-y} \end{aligned} \tag{42}$$

where $a(z)$ and $b(z)$ are any two y values for any two consecutive $\underline{RP}(Y)$ values. Due to the fact that $\text{MRP}_0^l(z)$ is based on the NPI lower reproducibility probability and that its lowest value is 0.5, so $MRP_0^l(z) = 1$ for $z \in [0, 0.5)$. The probability P_0^*, in this paper, depends on the RRT method used, so it is either $P_{G_0}^*$ or $P_{F_0}^*$, which are derived from Equations (5) or (10) in Section 2.1. To apply this measure, we specify all the values of $y = a(z)$ and $y = b(z)$ for any two consecutive $\underline{RP}(Y)$ values where $\underline{RP}(Y) \geq z$, and then calculate the summation of probabilities of all these $Y = y$ values except $[a(z), b(z)]$ as shown in Equation(42) such that $a(z)$ is the lowest integer values such that the condition $\underline{RP}(y) \geq z$ for $y < a(z)$ is selected. Similarly, the $b(z)$ is the largest integer value such that the condition $\underline{RP}(y) \geq z$ for each $y > b(z)$ is selected.

Similarly, we can investigate the measure of reproducibility under the alternative hypothesis $H_1 : P^* > P_0^*$, where the probability of people who say 'Yes' under H_1 is P_1^* which is not a single value. These values P_1^* can be selected to provide high power more than 0.90 for the statistical hypothesis test. So, the measure of lower reproducibility probability under H_1 is:

$$\begin{aligned} MRP_1^l(z) &= P(\underline{RP}(Y) \geq z | H_1) = P[\underline{RP}(Y) \geq z \ | Y \sim Bin(n, P_1^*)] \\ &= 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P_1^*)^y (1 - P_1^*)^{n-y} \end{aligned} \tag{43}$$

where P_1^* represents the probability of people who say 'Yes' to the selected question under the hypothesis H_1. The probability P_1^* depends on the RRT method used, so it is either $P_{G_1}^*$ or $P_{F_1}^*$, which are derived from Equations (5) and (10) in Section 2.1, which relate to H_1. Example 3 illustrates this measure using the Greenberg method as explained in Section 2.1.

It is noticed that the alternative hypothesis $H_1 : P^* > P_0^*$ where the values of P_1^* are greater than P_0^* and less than 1. Therefore, we linked between the power and the alternative hypothesis because power is defined as the probability of being able to reject the null hypothesis correctly in the case that the alternative

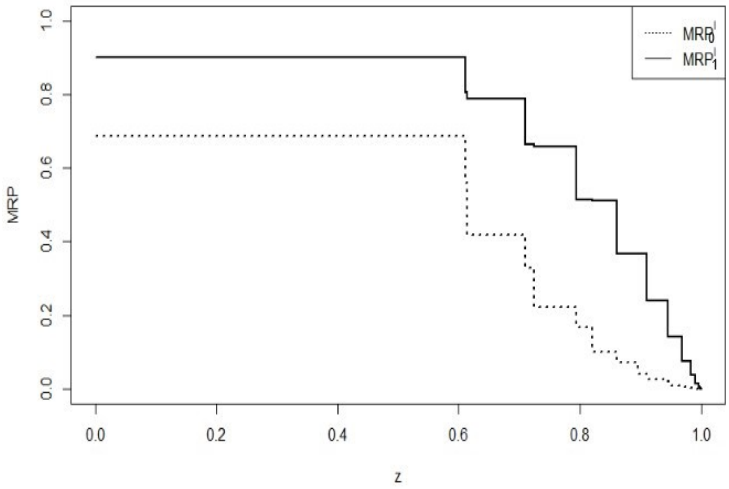

Fig. 1 MRP$_0^l(z)$ and MRP$_1^l(z)$ with GB data with $n = 30$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\pi_B = 0.3$, $\gamma = 0.7$, $\alpha = 0.05$, $P^*_{G_0} = 0.58$, $P^*_{G_1} = 0.72$

hypothesis is true, then power is equal to 1 minus the probability of Type II error β in the event that the alternative hypothesis is true. So, in order to increase power and the probability that the alternative hypothesis would come true, we select $\beta = 0.1$ and $P^*_1 = 0.9$.

Example *3* This example illustrates the measure of reproducibility probability (MRP$_0^l$) for one-sided hypothesis tests using data collected by the GB method [11]. We use the same parameters of the GB method of n, π_{A_0}, π_B and γ in Example 1. We want to test the null hypothesis $H'_0 : \pi_A = 0.7$ against the alternative hypothesis $H'_1 : \pi_A > 0.7$. The corresponding null hypothesis $H_0 : P^*_G = 0.58$ against the alternative hypothesis $H_1 : P^*_G > 0.58$ using Equation (5) of the GB method. Under the null hypothesis H'_0, assume that $\pi_{A_0} = 0.7$ and under alternative hypothesis H'_1, suppose that $\pi_{A_1} = 0.9$, so the proportion under H_0 and H_1 are:

$$P^*_{G_0} = \gamma\pi_{A_0} + (1-\gamma)\pi_B = 0.58 \tag{44}$$

$$P^*_{G_1} = \gamma\pi_{A_1} + (1-\gamma)\pi_B = 0.72 \tag{45}$$

The MRP$_0^l$ and MRP$_1^l$ are calculated as explained in Section 5.1. The results for MRP$_0^l$ and MRP$_1^l$ for different values of z are shown in Figure 1 and Tables 2 and 3, respectively. It has been observed that the MRP$_0^l(z)$ and MRP$_1^l(z)$ show a decreasing trend when the value of z increases. In the case that the values of z get closer to 1, they cause MRP$_0^l(z)$ and MRP$_1^l(z)$ to decrease because of the lower reproducibility probabilities get higher values due to the number of 'Yes' responses y is close to either 0 or the total number of responses n. In both cases, these responses provide substantial support for either the null or alternative hypothesis, and the NPI lower reproducibility probabilities indicate that if all responses are 'Yes' ('No'), the responses do not provide evidence against the possibility that the responses are 'No' ('Yes').

Table 5 $\mathrm{MRP}_0^l(z)$ with GB data with $n = 30$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\pi_B = 0.3$, $\gamma = 0.7$, $\alpha = 0.05$, $P^*_{G_0} = 0.58$, $P^*_{G_1} = 0.72$

z	$\mathrm{MRP}_0^l(z)$	z	$\mathrm{MRP}_0^l(z)$	z	$\mathrm{MRP}_0^l(z)$
0.5000	0.9020	0.8954	0.3667	0.9939	0.0151
0.6106	0.8067	0.9101	0.2400	0.9956	0.0056
0.6145	0.7898	0.9449	0.1420	0.9980	0.0018
0.7102	0.6644	0.9479	0.1419	0.9990	0.0018
0.7240	0.6575	0.9680	0.0755	0.9992	0.0005
0.7941	0.5137	0.9790	0.0754	0.9997	0.0001
0.8198	0.5115	0.9824	0.0358	0.9999	0.0000
0.8605	0.3673	0.9909	0.0151	1.0000	0.0000

Table 6 $\mathrm{MRP}_1^l(z)$ with GB data with $n = 30$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\pi_B = 0.3$, $\gamma = 0.7$, $\alpha = 0.05$, $P^*_{G_0} = 0.58$, $P^*_{G_1} = 0.72$

z	$\mathrm{MRP}_1^l(z)$	z	$\mathrm{MRP}_1^l(z)$	z	$\mathrm{MRP}_1^l(z)$
0.5000	0.6866	0.8605	0.0721	0.9824	0.0009
0.6106	0.5618	0.8954	0.0392	0.9909	0.0007
0.6145	0.4181	0.9101	0.0254	0.9939	0.0001
0.7102	0.3299	0.9449	0.0197	0.9956	0.0001
0.7240	0.2221	0.9479	0.0071	0.9980	0.0001
0.7941	0.1678	0.9680	0.0050	0.9990	0.0000
0.8198	0.1013	0.9790	0.0016		

The reproducibility probabilities included in $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ area close to 1 for all values of z within the range of 0 to 0.6. Both $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ have values on nearby to 0 when $z = 1$. The values of MRP_1^l are greater than the values of MRP_0^l for all values of z inside the interval $[0, 1]$ if the GB has a large threshold value or large P_1^* under H_1. Variations in the $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ are caused by variations in the method's parameters γ, π_{A_0}, π_B, and α. When these values are increased, there is a corresponding increase in $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ respectively.

The upper reproducibility probabilities can be used to derive $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_1^u(z)$ for the RRT methods, similar to the derivation of $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the lower reproducibility probabilities. Furthermore, the FM technique or other RRT methods can be used to make a comparison between them.

However, the RRT method parameters must be chosen carefully in order to calculate the minimum required sample size to obtain the required power of the hypothesis tests and with a specific significance level that can provide a high reproducibility probability of hypothesis tests based on RRT data, as applied in Example 4.

To determine the minimum sample size n_r required for this case for getting an approximate power (i.e. $1-\beta$) at a level of significance of (i.e., $\alpha = 0.05$), use Equation (46) [39].

$$\lceil n_r \rceil \geq \left[\frac{z_{1-\alpha}\sqrt{P_0^*(1-P_0^*)} + z_{1-\beta}\sqrt{P_1^*(1-P_1^*)}}{P_1^* - P_0^*} \right]^2 \tag{46}$$

where the approximate power is calculated using [39]:

$$1 - \beta \approx P\left(Z \geq \frac{n(P_0^* - P_1^*) + z_{1-\alpha}\sqrt{nP_0^*(1-P_0^*)}}{\sqrt{nP_1^*(1-P_1^*)}} \right) \tag{47}$$

The value of $z_{1-\alpha}$ and $z_{1-\beta}$ indicate to the $(1-\alpha)\times 100$ and $(1-\beta)\times 100$ percentiles of standard normal distribution respectively. If the hypothesis tests do not provide a required power of 0.90 with sample size n_r, Fleiss et.al [15] recommended adding $\frac{1}{|P_1^*-P_0^*|}$ as a continuity correction to $\lceil n_r \rceil$.

$$n = \lceil n_r \rceil + \frac{1}{|P_1^* - P_0^*|} \tag{48}$$

where $\lceil n_r \rceil$ is the minimal integer greater than or equal to n_r, and the probability P_0^* are $P_{G_0}^*$ or $P_{F_0}^*$ which are derived from Equations (5) or (10) in Section 2.1 under H_0.

5.2 The area under MRP (AUMRP)

In Section 5.1, we introduce MRP as a measurement of reproducibility probability of statistical tests based on data collected by the GB and the FM methods. In order to compare the reproducibility probability of statistical tests based on different RRT methods, we introduce an overall measure based on MRP, namely the area under MRP(z) under H_0 and under H_1 which are denoted by $\text{AUMRP}_0^l(z)$ and $\text{AUMRP}_1^l(z)$, respectively. Given $\text{MPR}_0^l(z)$ and $\text{MPR}_1^l(z)$, computed by Equations (42) and (43), the AUMRP_0 and AUMRP_1 are calculated as follows.

Let $AUMRP : [0,1] \rightarrow \mathbb{R}$ be a function defined on a closed interval $[0,1]$ of the real numbers, $\mathbb{R}$, and D as a partition of the interval $[0,1]$. Let z_i represent the real number that bounds each subinterval on the number line. Here, i ranges from 0 to n, and D is defined as follows: $D = \{[z_0, z_1], [z_1, z_2], \ldots [z_{n-1}, z_n]\}$ where $0 = z_0 < z_1 < z_2 < \cdots < z_n = 1$. Therefore, AUMRP_0^l and AUMRP_1^l over $[0,1]$ with partition D are

$$AUMRP_0^l = \sum_{i=1}^{n} MRP_0^l(z_i^*)\,\Delta z_i \tag{49}$$

$$AUMRP_1^l = \sum_{i=1}^{n} MRP_1^l(z_i^*)\,\Delta z_i \tag{50}$$

where $\Delta z_i = z_i - z_{i-1}$ where $z_i^* \in [z_{i-1}, z_i]$. Example 3 introduces MRP_0^l and MRP_1^l of the GB method.

Example 4 This example derives AUMPR_0^l and AUMRP_1^l of one-sided hypothesis tests based on the GB method [11] using the minimum required sample size as explained in Section 5.1 to get high reproducibility. We use the same combinations of the GB method of π_{A_0}, π_B and γ in Example 1.

Assume that H_0 and H_1 values which we want to test are $\pi_{A_0} = 0.7$ and $\pi_{A_1} = 0.9$, respectively, with significance level $\alpha = 0.05$ and power 0.90. Then, we derive AUMPR_0^l and AUMRP_1^l using Equations (49) and (50) and using the minimum required sample size n.

As shown in Table 7, the results give the required minimum sample sizes for different values of π_B. So, if $\pi_B = 0.10$, then the threshold value is 71, the AUMRP_0^l equals

Table 7 AUMRP_0^l, AUMRP_1^l of the GB method with $\gamma = 0.7$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $\beta = 0.1$

π_B	**0**	**0.1**	**0.25**	**0.3**	**0.45**	**0.6**
n	121	119	115	113	106	98
c	68	71	74	74	74	73
$P^*_{G_0}$	0.4900	0.5200	0.5650	0.5800	0.6250	0.6700
$P^*_{G_1}$	0.6300	0.6600	0.7050	0.7200	0.7650	0.8100
power	0.9262	0.9123	0.9090	0.9227	0.9316	0.9313
AUMRP_0^l	0.8070	0.8190	0.8225	0.8112	0.8029	0.8049
AUMRP_1^l	0.8235	0.8114	0.8087	0.8200	0.8281	0.8278

Table 8 The AUMRP_0^l, AUMRP_1^l of the FM method with $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\gamma_2 = 0.10$, $\alpha = 0.05$, $\beta = 0.1$

γ_1	**0.10**	**0.13**	**0.15**	**0.23**	**0.27**	**0.29**
n_r	76	80	84	100	109	115
c	57	60	64	77	85	90
$P^*_{F_0}$	0.6600	0.6690	0.6750	0.6990	0.7110	0.7170
$P^*_{F_1}$	0.8200	0.8230	0.8250	0.8330	0.8370	0.8390
power	0.9210	0.9367	0.9124	0.9358	0.9275	0.9315
AUMRP_0^l	0.8122	0.7966	0.8200	0.8010	0.8073	0.8047
AUMRP_1^l	0.8194	0.8335	0.8119	0.8319	0.8238	0.8274

0.8190 and AUMRP_1^l equals 0.8114 with power is 0.9123, whereas the AUMRP_0^l equals to 0.8225 and AUMRP_1^l equals to 0.8087 for $\pi_B = 0.25$ with threshold value is 74 and power is 0.9090. It is noted that for all values of $\pi_B \in [0, 0.6]$, AUMRP_0^l and AUMRP_1^l taking values between 0.80 and 0.81 and the AUMRP_1^l is always greater than the AUMRP_0^l except the case of $\pi_B = 0.1$ and 0.25. Similarly, we drive AUMPR_0^l and AUMRP_1^l of one-sided hypothesis tests based on the FM method using the same procedure as explained in Example 5.

***Example** 5* This example derives AUMPR_0^l and AUMRP_1^l of one-side hypothesis tests using the FM method. Assume that the probability of being asked the sensitive question is 0.75, the forced 'Yes' answer is $\gamma_1 = 0.10$ and the forced 'No' answer is $\gamma_2 = 0.15$, where the significance level is $\alpha = 0.05$, and power 0.90.

Let us consider that the H_0 value which we want to test is $\pi_{A_0} = 0.7$, while the alternative proportion of people with the sensitive characteristic is $\pi_{A_1} = 0.90$. For varying values of γ_2, the required minimum sample sizes and values of AUMRP_0^l and AUMRP_1^l are determined.

Table 8 presents the values of AUMRP_0^l and AUMRP_1^l for the FM method under the null hypothesis H_0 and the alternative hypothesis H_1, calculated using Equations (49) and (50), respectively. The values of AUMRP_0^l are within the range of 0.79 to 0.82 while AUMRP_1^l takes values within the range of 0.81 to 0.83. The AUMRP_0^l is always greater than the AUMRP_1^l except in the case of $\gamma_1 = \gamma_2 = 0.15$. Nevertheless, the patterns of AUMRP_0^l or AUMRP_1^l are not clear. The FM method requires a smaller sample size ($n_r = 76$) compared to the GB method ($n = 121$) in order to reach a power of 0.92 and obtain value 0.81 of the AUMRP_0^l and AUMRP_1^l.

Table 9 The $\text{Var}(\hat{\pi}_{A_0})_{GB}$ AUMRP_0^l, AUMRP_1^l of one-sided tests based on the GB method with $\pi_B = 0.4$, $\gamma = 0.5554$, $\pi_{A_1} = 0.9$, $P^*_{G_1} = 0.6777$, $\alpha = 0.05$, $\beta = 0.1$, $\Delta_{GB} = 1.224$

π_{A_0}	0.570	0.575	0.580	0.585	0.590	0.595	0.600	0.700
$\text{Var}(\hat{\pi}_{A_0})_{GB}$	0.8103	0.8104	0.8105	0.8104	0.8104	0.8103	0.8101	0.7961
n	72	74	76	78	80	83	85	181
$P^*_{G_0}$	0.4944	0.4972	0.5000	0.5027	0.5055	0.5083	0.5110	0.5666
Power	0.9075	0.9183	0.9278	0.9363	0.9126	0.9100	0.9200	0.9262
AUMRP_0^l	0.8207	0.8121	0.8033	0.7945	0.8154	0.8193	0.8098	0.8103
AUMRP_1^l	0.8097	0.8184	0.8267	0.8348	0.8133	0.8110	0.8192	0.8218

Now, it is worth to compare the reproducibility of statistical tests based on different RRT methods, taking into account the variance of the estimators and reproducibility of statistical hypothesis tests at the same degree of privacy. In order to increase the reproducibility probability, we choose the required minimum sample size when using the GB and FM methods while selecting different parameters for the RRT methods to get equivalent privacy and variance for the estimator $\hat{\pi}_{A_0}$. This is due to the study of the relationship between using required minimum sample sizes and reproducibility probability at the same degree of privacy. This choice of the parameters gives the same values of both variances of the estimator $\hat{\pi}_{A_0}$ and the same privacy degree of the GB and FM method to check the changes in reproducibility of statistical hypothesis tests as assumed in Example 6.

***Example** 6* Assume that we use the requird minimum sample size n of the parameters $\gamma = 0.5554$, $\pi_{A_1} = 0.9$ of the GB method, and $\gamma_1 = 0.20829$, $\gamma_2 = 0.10$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $\beta = 0.1$ as parameters of the FM method.

The aim of this example is to compare the GB and FM methods throughout various values of π_{A_0}, specifically focusing on their reproducibility. Both methods are assumed to have the same privacy degree of approximately 1.224 but differ in regards to the variance of the estimator $\hat{\pi}_{A_0}$. Tables 9 and 10 provide the relevant details for this comparison.

The variance for various values of $\hat{\pi}_{A_0}$ ranging from 0.79 to 0.81 is presented in Table 9 for the GB technique with a privacy degree of $\Delta_{GB} = 1.224$. Both AUMRP_0^l and AUMRP_1^l show no visible pattern, with a high power level of more than 0.90. The value of AUMRP_0^l shows a range of values from 0.79 to 0.82, whereas AUMRP_0^l displays a range of values from 0.80 to 0.83 for varying π_{A_0}.

The FM method with privacy degree, denoted as $\Delta_{FM} = 1.224$, has reduced variance for various values of π_{A_0} compared to the variance of the estimator of the GB, as seen in Table 10. The AUMRP_0^l and AUMRP_1^l show no apparent trend, and the power is more than 0.90. The value of AUMRP_0^l shows a range of values between 0.78 and 0.81, whereas AUMRP_0^l shows a range of values between 0.86 and 0.88 for varying π_{A_0}. The estimator of the FM method has a smaller variance compared to the estimator of the GB methods, although the AUMRP_0^l of the GB method shows higher reproducibility than the FM method. Conversely, the AUMRP_1^l of the FM method shows more reproducibility than the GB method when both are given the same privacy degree of 1.224.

As evidenced by the information presented in Tables 9 and 10, while the variances of the estimator $\hat{\pi}_{A_0}$ are low in order to improve reproducibility, it could be less

Table 10 The $\text{Var}(\hat{\pi}_{A_0})_{FM}$, AUMRP_0^l, AUMRP_1^l of one-sided tests based on FM method with $\gamma_2 = 0.10$, $\gamma_1 = 0.20829$, $\pi_{A_1} = 0.9$, $P_{F_1}^* = 0.8308$, $\alpha = 0.05$, $\beta = 0.1$, $\Delta_{FM} = 1.224$

π_{A_0}	0.570	0.575	0.580	0.585	0.590	0.595	0.600	0.700
$\text{Var}(\hat{\pi}_{A_0})_{FM}$	0.2554	0.2546	0.2539	0.2531	0.2523	0.2515	0.2507	0.2351
n	49	50	52	53	55	56	58	117
$P_{F_0}^*$	0.6026	0.6060	0.6095	0.6129	0.6164	0.6199	0.6233	0.6925
Power	0.9041	0.9020	0.9052	0.9026	0.9050	0.9019	0.9716	0.9672
AUMRP_0^l	0.7958	0.8035	0.7843	0.7913	0.8113	0.8173	0.7973	0.8102
AUMRP_1^l	0.8769	0.8690	0.8881	0.8808	0.8658	0.8581	0.8776	0.8690

possible to assume the parameters mentioned above in order of reducing the level of privacy. Hence, it is essential to consider several hypothetical values and different parameters in order to achieve an equivalent level of privacy with less variability in the actual responses and higher reproducibility.

6 Concluding remarks

This paper introduces an innovative method to assess the reproducibility probability of statistical hypothesis tests using data obtained by RRT methods, including the GB and FM methods. This approach uses the number of 'Yes' responses within a specific sample and the threshold to perform the tests. Next, use the Nonparametric Predictive Inference (NPI) method for Bernoulli variables in order to calculate the lower and upper reproducibility probability of one-sided hypothesis tests. The advantage of employing reproducibility of statistical tests is that they can be designed for any RRT method because this method depends on the number of orderings of yes responses, not on the binomial distribution.

For reproducibility of one-sided hypothesis tests, we introduced the measurement of lower and upper reproducibility probability MRP_0^l and MRP_1^l under H_0 using the threshold values. Then, we compared the GB and the FM methods by derivation of the required minimum sample size with respect to a higher power of more than 0.90 and $\alpha = 0.05$. After that, we calculated the area under MRP_0^l and MRP_1^l. In addition, derive the lower and upper threshold values to find the same area of the threshold value of MRP_0^l and MRP_1^l using different parameters of the RRT method. The finding is the GB method has more reproducibility than the FM methods for one-sided tests under H_0 especially if both methods have the same sample size, the threshold value or the probability of people who say 'Yes'. Conversely, the FM method has more reproducibility than the GB methods for one-sided tests under H_1.

For using the required minimum sample size, the same privacy degree, and with the same proportion of sensitive characteristics in the population π_{A_0}, the FM method takes smaller samples than the GB method requires. As a result, choosing the same parameters within significance level $\alpha = 0.05$ and power more than 0.90 needs to increase the sample size of the GB method than the FM method to obtain the AUMRP_0^l and AUMRP_1^l for one-sided tests with

the same privacy degree. In addition, high reproducibility of hypothesis tests based on a randomised response method provides a probability, denoted as P_0^* (P_1^*), which represents the probability of people that respond 'Yes' and close to the H_0 and H_1 values which we want to test are π_{A_0} (π_{A_1}), under the null hypothesis H_0 and H_1 respectively.

Furthermore, less variability in the reported responses of any RRT method leads to higher reproducibility with the same degree of privacy.

Acknowledgements

The research described in this article was conducted during Fatimah Alghamdi's PhD studies at the Department of Mathematical Sciences, Durham University. Financial funding for this research was provided by the Ministry of Education in Saudi Arabia, Princess Nourah bint Abdulrahman University, and the Saudi Arabian Cultural Bureau in London. We express our gratitude to Professor Sat Gupta for his valuable contributions and insightful discussions during this research project.

References

[1] S.N. Goodman, A comment on replication, p-values and evidence. *Statistics in Medicine*, **11**, 875–879 (1992).

[2] S. Senn, A comment on replication, p-values and evidence S.N. Goodman. *Statistics in Medicine*, **21**, 2437–2444 (2002).

[3] F.P.A. Coolen, S. BinHimd. Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, **8**, 591-618 (2014).

[4] T. Augustin, F.P.A. Coolen, Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251–272 (2004).

[5] F.P.A. Coolen, Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, **36**, 349–357 (1998).

[6] D. Billheimer, Predictive inference and scientific reproducibility. *The American Statistician*, **73**, 291-295 (2019).

[7] B. De Finetti, Theory of Probability. London: Wiley (1974).

[8] S.L. Warner, Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **62**, 63–69 (1965).

[9] B.G. Greenberg, A.L.A. Abul-Ela, W. R. Simmons, D. G. Horvitz, The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, **65**, 520–539 (1969).

[10] A. Chaudhuri, Randomized response and indirect questioning techniques in surveys. New York: CRC (2016).

[11] J.R. Abernathy, B.G. Greenberg, D.G. Horvitz, Estimates of induced abortion in urban North Carolina. *Demography*, **7**, 19–29 (1970).

[12] H. Zhimin, Y. Zaizai, Measure of privacy in randomized response model. *Quality and Quantity*, **46**, 1167–1180 (2012).

[13] A.Y. Kuk, Asking sensitive questions indirectly. *Biometrika*, **77**, 436–438 (1990).

[14] F.B. Adebola, A.A. Adediran, O.S. Ewemooje, Hybrid tripartite randomized response technique. *Communications in Statistics - Theory and Methods*, **46**, 11756-11763 (2017).

[15] J.L. Fleiss, B. Levin, M.C. Paik, *Statistical Methods for Rates and Proportions*, Third Edition, New York: John Wiley and Sons (2003).

[16] S. A. Eriksson, A new model for randomized response. *International Statistical Review*, **41**, 101–113 (1973).

[17] A. Young, S. Gupta, R. Parks, A binary unrelated-question RRT model accounting for untruthful responding. *Involve - Journal of Mathematics*, **12**, 1163-1173 (2019).

[18] H. Anderson, Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, **4**, 11–19 (1977).

[19] L. Ljungqvist, A unified approach to measures of privacy in randomized response models: A utilitarian perspective. *Journal of the American Statistical Association*, **88**, 97–103 (1993).

[20] J. Lanke, On the degree of protection in randomized interviews. *International Statistical Review*, **44**, 197–203 (1976).

[21] B.M. Hill, Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677–691 (1968).

[22] F.P.A. Coolen, On nonparametric predictive inference and objective Bayesiansm. *Journal of Logic, Language and Information*, **15**, 21–47 (2006).

[23] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall (1991).

[24] K. Weichselberger, Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als Umfassendes Konzept (In German).Physika, Heidelberg (2001).

[25] G.R.J. Arts, F.P.A. Coolen, Two nonparametric predictive control charts, *Journal of Statistical Theory and Practice*, **2**, 499–512 (2008).

[26] G.R.J. Arts, F.P.A. Coolen, P. Van der Laan, Nonparametric predictive inference in statistical process control, *Quality Technology and Quantitative Management*, **1**, 201–216 (2004).

[27] J. Chen, F.P.A. Coolen, T. Coolen-Maturi, On nonparametric predictive inference for asset and European option trading in the binomial tree model. *Journal of the Operational Research Society*, **70**, 1678–1691 (2019).

[28] R.M. Baker, T. Coolen-Maturi, F.P.A. Coolen, Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, **44**, 1333–1349 (2017).

[29] T. He, F.P.A. Coolen, T. Coolen-Maturi, Nonparametric predictive inference for European option pricing based on the Binomial Tree Model. *Journal of the Operational Research Society*, **70**, 1692–1708 (2019).

[30] F.P.A. Coolen, P. Coolen-Schrijner, Nonparametric predictive subset selection for proportions. *Statistics & Probability Letters*, **76**, 1675–1684 (2006).

[31] F.P.A. Coolen, P. Coolen-Schrijner. Nonparametric predictive reliability demonstration for failure-free periods. *IMA Journal of Management Mathematics*, **16**, 1–11 (2005).

[32] S. BinHimd, *Nonparametric predictive methods for bootstrap and test reproducibility*. PhD Thesis. Durham University (2014). Available at: http://npi-statistics.com.

[33] J.D. Gibbons, S. Chakraborti, *Nonparametric Statistical Inference (5th ed.)*. Chapman and Hall, Boca Raton, Florida (2011).

[34] T. Coolen-Maturi, P. Coolen-Schrijner, F.P.A. Coolen. Nonparametric predictive pairwise comparison for real-valued data with terminated tails, International Journal of Approximate Reasoning, **51**, 141-150 (2009).

[35] H.N. Alqifari, *Nonparametric Predictive Inference for Future Order Statistics*. PhD Thesis. Durham University (2017). Available at : http://npi-statistics.com.

[36] F.P.A. Coolen, F.J. Marques, Nonparametric predictive inference for test reproducibility by sampling future data orderings, *Journal of Statistical Theory and Practice*, **14**, 1–22 (2020).

[37] F.J. Marques, F.P.A. Coolen, and T. Coolen-Maturi, Approximations for the likelihood ratio statistic for hypothesis testing between two Beta distributions, *Journal of Statistical Theory and Practice*, **13**, 17 (2019).

[38] A. Simkus, F.P.A. Coolen, T. Coolen-Maturi, N.A. Karp, C. Bendtsen, Statistical reproducibility for pairwise t-tests in pharmaceutical research, *Statistical Methods in Medical Research*, **31**, 673–688 (2022).

[39] S. Chow, J. Shao, H. Wang, *Sample Size Calculations in Clinical Research*. Second Edition. New York: CRC (2008).

[40] M. Lovig, S. Khalil, S. Rahman, P. Sapra, S. Gupta, A mixture binary RRT model with a unified measure of privacy and efficiency, Communications in Statistics - Simulation and Computation, **52**, 2727–2737 (2023).

[41] M. Parker, S. Gupta, S. Khalil, A Mixture Quantitative Randomized Response Model That Improves Trust in RRT Methodology. Axioms. **13**, 11 (2024).