Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies

Andrea Ganna*, Donghwan Lee*, Erik Ingelsson and Yudi Pawitan

Submitted: 1st July 2014; Received (in revised form): 5th August 2014

Abstract

It is common and advised practice in biomedical research to validate experimental or observational findings in a population different from the one where the findings were initially assessed. This practice increases the generalizability of the results and decreases the likelihood of reporting false-positive findings. Validation becomes critical when dealing with high-throughput experiments, where the large number of tests increases the chance to observe false-positive results. In this article, we review common approaches to determine statistical thresholds for validation and describe the factors influencing the proportion of significant findings from a ‘training’ sample that are replicated in a ‘validation’ sample. We refer to this proportion as rediscovery rate (RDR). In high-throughput studies, the RDR is a function of false-positive rate and power in both the training and validation samples. We illustrate the application of the RDR using simulated data and real data examples from metabolomics experiments. We further describe an online tool to calculate the RDR using t-statistics. We foresee two main applications. First, if the validation study has not yet been collected, the RDR can be used to decide the optimal combination between the proportion of findings taken to validation and the size of the validation study. Secondly, if a validation study has already been done, the RDR estimated using the training data can be compared with the observed RDR from the validation data; hence, the success of the validation study can be assessed.

Keywords: statistical validation; rediscovery rate; false discovery rate; multiple testing; metabolomics

INTRODUCTION

Reproducibility is a hallmark of scientific research. Replication or validation studies are often undertaken to confirm that the observed findings are not artifact due to chance or uncontrolled bias. An excess of statistical significance, beyond that expected by the power of the studies, has been recently reported for animal studies of neurological diseases [1] and similar results are shown in clinical and epidemiological research [2]. These and other studies have fueled the discussion on the importance of reproducibility in biomedical science and recently pushed the US National Institutes of Health to consider measures to increase validation requirements [3].

In genetic research, validation of significant findings in external populations is a standard practice

Corresponding author. Yudi Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. Tel.: +46 8 5248 3983; E-mail: yudi.pawitan@ki.se

*These authors contributed equally to this work.

Andrea Ganna is PhD candidate at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet and Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Sweden. His doctoral research focuses on prediction model for cardiovascular diseases and high-throughput data analysis.

Donghwan Lee is assistant professor at the Department of Statistics, Ewha Womans University, Korea. His research interest is the development of statistical model and multiple testing procedure for high-throughput data analysis.

Erik Ingelsson is professor of molecular epidemiology at the Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Sweden. His research focuses on how cardiovascular disease and related conditions varies with DNA variation, RNA expression, circulating biomarkers such as proteins and metabolites, as well as with structures and functions measured by imaging.

Yudi Pawitan is a professor at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden. His research interest is in the methods development and applications of high-throughput data analysis.

© The Author 2014. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com
when performing genome-wide association studies (GWAS) [4, 5], and most journals will not publish findings from GWAS without significant evidence of validation [6]. This practice is justified by the fact that most pre-GWAS studies could not be validated and resulted in a large number of false-positive publications [2, 7]. In epidemiological research, validation studies are not as widespread as in genetic research, but they have been advocated to enhance the likelihood of research findings being true [8] and to improve estimation accuracy [9]. Although the importance of replication or validation studies is well recognized by the scientific community, there seems to be less awareness of the factors influencing the reproducibility of significant findings. Indeed, the selection of the validation study is more often driven by data availability rather than study design.

In this study, we explore the factors that influence the proportion of significant findings from a training sample that are replicated in a validation sample. We refer to this quantity as rediscovery rate (RDR) and describe its properties in multiple testing scenarios. Moreover, we examine the use of the RDR with simulated and real data from metabolomics experiments in two population-based studies.

A HYPOTHETICAL SCENARIO

It is common practice in high-throughput data analysis to compare genes, metabolites or proteins among the diseased and nondiseased groups. This is first performed in a training study. Findings that are statistically significant are often reevaluated in a validation study. As an informal illustration, in Table 1, panel A, we present a hypothetical scenario where 10 000 genes have been tested for association with the disease status. Typically, after each test is computed, the test statistics are ordered and the cutoff beyond which all genes are assumed to be associated with the diseases is determined. Associated genes are retested in a validation sample (Table 1, panel B). Tests in the validation sample are computed in the same way in the training sample and genes are claimed to be associated with the disease if the test statistics are higher than a certain threshold. Genes that are associated also in the validation sample are called ‘validated’.

In this example, $R_t = 1250$ genes are significantly associated in the training sample (the subscript $t$ means that these variables refer to the training sample) and $R_v = 663$ of those are validated (the subscript $v$ indicates the these variables refer to the validation sample). The proportion of genes falsely declared associated in the training sample or false-positive rate is $V_t/(U_t + V_t) = 450/9000 = 5\%$, and the observed power of the test or sensitivity is $S_t/(S_t + W_t) = 800/1000 = 80\%$. Similarly, in validation sample (Table 1 Panel B), the false-positive rate and the observed power are $V_v/(U_v + V_v) = 5\%$ and $S_v/(S_v + W_v) = 80\%$, respectively.

Therefore, training and validation samples have same false-positive rate and power. Moreover, the observed false discovery rate (FDR) in the validation sample is $V_t/R_t = 23/663 = 3.6\%$, which means there are only 3.6\% false positives among the significant findings. However, the proportion of significant findings from a training sample that is replicated into a validation sample is $R_v/R_t = 663/1, 250 = 53\%$, i.e. only half of the significant genes are validated.

Researchers generally want to maximize the number of validated findings, i.e. increase $R_v/R_t$, while minimizing the number of false positives among the validated findings, i.e. decrease $V_v/R_v$.

A standard approach is to use multiple-testing adjustment methods (e.g. Bonferroni and FDR) both on the training and the validation sample. For example, Low and colleagues [10], in a GWAS for intracranial aneurysm in the Japanese population, used a liberal threshold of $P < 0.0001$ in the training sample and a Bonferroni corrected $P < 0.0016$ in the validation sample. Dick and colleagues [11], in one of the first studies investigating the association between DNA methylation and body mass index, applied an FDR correction of 5\% in the training sample.

<table>
<thead>
<tr>
<th>Test result</th>
<th>Nonassociated</th>
<th>Associated</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A (Training sample)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truly nonassociated</td>
<td>$U_t = 8550$</td>
<td>$V_t = 450$</td>
<td>9000</td>
</tr>
<tr>
<td>Truly associated</td>
<td>$W_t = 200$</td>
<td>$S_t = 800$</td>
<td>1000</td>
</tr>
<tr>
<td>Total</td>
<td>8750</td>
<td>1250</td>
<td>10000</td>
</tr>
<tr>
<td>Panel B (Validation sample)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truly nonassociated</td>
<td>$U_v = 427$</td>
<td>$V_v = 23$</td>
<td>450</td>
</tr>
<tr>
<td>Truly associated</td>
<td>$W_v = 160$</td>
<td>$S_v = 640$</td>
<td>800</td>
</tr>
<tr>
<td>Total</td>
<td>587</td>
<td>663</td>
<td>1250</td>
</tr>
</tbody>
</table>

The rows describe the true state of nature and the columns the test decision based on the experimental data. Panel A, training sample: 10 000 genes are classified according to their true status and the test result. Panel B, validation sample: 1250 ($= R_v = N_v$) significant genes in the training population are reclassified in the validation study, assuming the same false-positive rate and sensitivity as in the training sample.
sample and a Bonferroni correction $P < 0.01$ in the validation. In a study of metabolites and prediabetes, Wang-Sattler and colleagues [12], used a Bonferroni correction $P < 0.00036$ in the training sample and a nominal $P < 0.05$ in the validation.

However, none of these studies provide arguments supporting the decision to use a specific multiple-correction method or a combination of methods. Researchers seem unaware of the issues related to the application of methods being developed for a single study problem to a training-validation problem.

We argue that the decision of the P-value thresholds should be tailored to take into account the characteristics of both the training and validation sample.

The article is organized as follows. In ‘Section 1’ we introduce the concept of RDR, false discovery rate in the validation sample (vFDR) and some basic notations.

In ‘Section 2’ we describe the factor influencing the RDR and vFDR and highlight potential issues related with using traditional multiple-testing correction methods both on training and validation samples. In ‘Section 3’ we describe the software to estimate the RDR and vFDR and show how this can be used with real data from metabolomics experiments. Finally, in ‘Section 4’ we discuss some practical considerations for designing validation studies or maximizing the proportion of validated true findings when the validation study has already been conducted. The article concludes with an overall discussion of the results and a detailed description of the methods, including a technical appendix.

**DEFINITION OF RDR AND VFDR**

We introduce two measures to quantify and control the proportion of validated findings. We define the RDR and the vFDR as follows (when $R_t = 0$, no validation study can be performed, so we condition on $R_t > 0$):

**Definition 1** The RDR is defined as the expected proportion of significant tests in the training sample that are also significant in the validation sample, i.e.

$$\text{RDR} = E(R_v / R_t | R_t > 0).$$

**Definition 2** The vFDR is defined to be the expected proportion of false positives among the declared significant results in the validation sample, i.e.

$$\text{vFDR} = E(V_v / R_v | R_v > 0).$$

We further define the analytical framework and introduce some basic notations to facilitate the interpretation of the results. We focus on the two independent group comparison problem (e.g. diseased and nondiseased) with equal variance using the standard $t$-test with pooled variance estimate. To estimate the RDR, we assume that the training sample and validation sample are homogeneous. Mathematical properties of the RDR and vFDR are provided in the technical Appendix.

- Let $n_0(n_{v0})$ and $n_1(n_{v1})$ be the sample sizes of the nondiseased group and diseased group in the training (validation) sample. And, let $n_t = n_{v0} + n_{v1}$ and $n_v = n_{v0} + n_{v1}$.
- Let $T_{ti}$ and $T_{vi}$ be the $t$-statistics for the $i$th gene in the training sample and validation sample ($i = 1, \ldots, N_t$), respectively.
- Given a critical value $c > 0$, we usually reject the null hypothesis if $|T_{ti}| > c$. Let $I = \{i; |T_{ti}| > c\}$. Only for $i \in I$, $T_{vi}$ is calculated. The critical value $k$ used in the validation sample might differ from the critical value in the training sample, $c$. Thus, in the validation sample, the association is declared significant if $|T_{vi}| > k$.
- $\alpha_t$ and $\alpha_v$ is the classical significance level in the training and validation sample, respectively. For example, $\alpha = 0.05$ is the commonly used threshold in a single hypothesis setting.
- $\beta_t$ and $\beta_v$ is the power in the training and validation sample, respectively.

We develop these measures in a mixture model framework. This simply means that the distribution of the effect sizes $\delta$ can be obtained from a mixture of central and noncentral $t$-distribution, each observed with a certain probability $\pi$. Effect sizes and probabilities are obtained with standard likelihood-based methods previously described [13] and reported in the technical Appendix.

In the technical Appendix we formally show that RDR and vFDR are functions of $\alpha$ and $\beta$ in both the training and validation samples.

**FACTORS INFLUENCING THE RDR**

RDR as function of false-positive rate and power

For a particular scenario, each critical value $c$ of the $t$-statistic generates two $2 \times 2$ tables similar to Table 1, panel A and B. We can consider a range of critical
values $c$ and evaluate the performances of the RDR and of the two main components: $\alpha$, or false-positive rate and $1 - \beta$, or power (Figure 1a). The RDR is a monotonically increasing function of the $c$ threshold. As the RDR increases, both the $\alpha$ and the $1 - \beta$ decrease. With a larger critical value $c$, the number of genes selected for validation decreases and the proportion of significantly validated genes increases. For example, validation of 50% of the genes in an external sample (RDR = 0.50) corresponds to a false-positive rate of 2% ($\alpha = 0.08$) and a power of 37% ($1 - \beta = 0.37$). Given a fixed sample size, larger RDR are achieved by reducing both power and false-positive rate.

**RDR using different critical values $c$**

Typically researchers do not explore the entire range of critical values $c$, but use some fixed $c$ threshold to select genes for validation. One widely used strategy in a multiple-testing setting is to control the FDR. For a $c$ threshold corresponding to an FDR of 5%, the false-positive rate is 0.09% and the power is 16% (Figure 1a). Thus, in a situation where 10,000 genes are tested, 171 genes are taken forward to validation in the external study. Of those, 135 are significant at a $P$-value of 0.05 ($\alpha = 0.05$) corresponding to an RDR of 79%.

Another strategy is to select the associated genes in the training sample by applying a Bonferroni correction.

**Figure 1:** RDR in multiple-testing setting. (a) Shows the changes in RDR (solid line), $\alpha$ level (dashed line) and power (dotted line) for changes in the critical value $c$ of the $t$-statistic in the training sample. Two vertical lines are drawn in correspondence of two multiple-testing correction methods (FDR = 5% and Bonferroni) assuming that 10,000 tests are performed. In (b) the RDR is calculated when the size of the validation sample is the same of the training (solid line), two times (dashed line) and half (dotted line) the training sample. Two scenarios, when $\pi_0 = 0.9$ and $\pi_0 = 0.99$, are shown. In (c) the RDR is calculated when the effect sizes are $(-0.2, -0.1, -0.05, 0.05, 0.1, 0.2)$ (solid line), two times (dashed line) and one-half (dotted line) the size. Two scenarios, when $\pi_0 = 0.9$ and $\pi_0 = 0.99$, are shown. In (d) the RDR is calculated when different $\alpha$ level for the validation sample are considered. Values of $\alpha_v$ are reported in the figure.
(\(x_t = 0.05/10000 = 0.000005\)). With this approach, 26 genes are selected for validation and 23 are declared validated, resulting in an RDR of 87%. The difference in the RDR between the two multiple-testing correction methods is small, while the difference in the number of validated genes is substantial: 112 more genes validated at a FDR of 5%. It is important to highlight that, among those additional genes, the vFDR is not any longer than 5% because an additional false-positive control is set by \(x_v = 0.05\). Specifically, we expect a vFDR of 0.3%, which corresponds to <1 rediscovered false-positive genes, resulting in more true-positive genes selected by controlling the FDR compared with the Bonferroni method. The selection of the critical value \(c\) in the training sample should be done bearing in mind both the RDR and the vFDR.

**Sample size and RDR**

In Figure 1a, the RDR does not reach 100%. This means that, for a certain sample size (of training and validation) and effect sizes, it is not possible to validate all the genes, no matter which critical value \(c\) is used for selection. This aspect is often underestimated. For example, using the Bonferroni approach, 26 genes are selected for validation and all of them are true positives. However, only 23 can be replicated, meaning that 3 genes that are truly associated with the disease cannot be validated given the size of the training and validation samples.

How can we validate all the genes that are truly significant in the training sample? There are three options.

1. To use a larger \(x_v\) in the validation sample (i.e. to use \(P\)-value > 0.05): this is, however, seldom done in practice. Genes are not considered validated if they are not significant in the validation sample.

2. To have a larger validation study (Figure 1b): this is probably a more feasible approach. However, in practice, the validation study is often smaller that the training because researchers want to maximize the power to discover new associations.

3. To control the vFDR instead of the \(P\)-value: similarly to what is often done in the training sample, the expected number of false positives can be controlled in the validation sample as a way to maximize the number of true positives.

**RDR and proportion of true nonsignificant genes**

In Figure 1b and c, we explore the RDR when the proportion of truly nonsignificant genes (\(\pi_0\)) is 0.9 and 0.99. The RDR is larger when \(\pi_0\) is smaller. This is, however, observed for lower \(c\) thresholds, while for larger thresholds the RDR tends to converge to the maximum, no matter what the proportion of truly significant tests. Intuitively, for large \(c\), the number of false positives is reduced toward zero (e.g. using Bonferroni correction) and therefore the validation can be achieved with the same power, independently of the proportion of truly nonnull genes. Conversely, when the number of false positives approaches zero, the probability of validating significant findings tends to be the same, regardless of the underlying number of true significant findings.

**RDR and \(k\) threshold in the validation population**

All the previous simulations were based on the assumption that the \(x_v = 0.05\). This approach seems reasonable, as it is unlikely that any researcher would consider a genetic variant being nonsignificant in the replication study as validated. It is common practice to allow an even more conservative \(P\)-value in the validation studies. In Figure 1d, we explore the RDR behavior for smaller \(x_v\), corresponding to the use of a larger \(k\) critical value. By using a smaller \(x_v\) the RDR is reduced for all possible critical values \(c\) in the training sample. Conversely, if the number of false positives taken forward for validation is already low (e.g. by using a Bonferroni correction), there is no need to apply an additional restriction to the false-positive rate by considering a low \(x_v\). For example, if the selection in the training population is done by using Bonferroni method, the number of false positives among the 26 genes selected for validation is close to zero. Therefore, any \(x_v < 1\) would only reduce the number of true positives.

**APPLICATION WITH REAL DATA**

**Software**

We have developed an online application at http://www.meb.ki.se/personal/yudpaw/rdr/ that allows one to calculate the FDR and vFDR given the results of a study.
Input:

- In step 1a and 1b the user can upload a file containing the t-statistics from the training sample comparing cases and controls or the beta and standard errors obtained, for example, from a logistic regression. It is also possible (optional) to upload the results from the validation sample to allow the calculation of the observed RDR. Alternatively, it is possible to use the data described in this article (select ‘Use example’).
- In step 1c, the user can set the number of components of the mixture model or obtain it automatically (suggested).
- In step 2, the user has two options: (i) to obtain the RDR for a fixed sample size in the validation sample or (ii) to fix the RDR and obtain the corresponding sample size.
- In step 2a, the user should specify the number of cases and controls in the training sample. Depending on the option selected in the previous step it is possible to specify the desired RDR and the case/control ratio or the number of cases and controls in the validation sample.
- In step 3, the user should specify the critical values to select features for the validation study ($\alpha_v$) and to determine which features are validated ($\alpha_o$). These values are specified as $-\log_{10} P$-values.
- In step 4, the user can decide which components to visualize. It is possible to plot the RDR, vFDR and observed RDR. Depending on the option selected in the step 2, the user will obtain a plot of these quantities as function of the critical values (in the training sample or in the validation sample) or as function of the total sample size in the validation sample.

Output:

- The upper panel of the result section reports the RDR and vFDR. If requested in step 2, the estimated sample size in the validation sample is also reported.
- The second panel reports the plot of the quantities selected in step 4.
- The third panel reports the histogram of the t-statistics and t-mixture density. This is a useful diagnostic tool to evaluate whether the mixture model is appropriate. If the histogram has a distribution that can not be well approximated by the density function, the estimation of the RDR and vRDR will be incorrect.
- The lower panel reports the parameters estimated by the t-mixture model: (i) the number of components selected, (ii) the estimated effect sizes for each component, (iii) the proportion of true null hypotheses, (iv) the proportion of each component of true alternative hypotheses.

We further created an R package called rdr, available at http://www.meb.ki.se/personal/yudpaw/, containing the following functions:

- rdr(): to obtain the RDR and vFDR as function of the significance level and power in both the training and validation samples.
- rdr.samplesize(): to compute the sample size in the validation sample for a given RDR value.
- rdr.est(): to estimate the RDR and vFDR based on the t-statistics calculated in the training sample, assuming that the test statistics follow a mixture distribution.

The detailed examples of R scripts are given in the technical Appendix.

Example with real data

We evaluate the RDR of a nontargeted metabolomics experiment comparing 5363 metabolic features in individuals with high (>3 mg/dl, defined as cases) versus low (≤3 mg/dl, defined as controls) high-sensitive C-reactive protein (hsCRP) levels in two studies: the TwinGene study, including 480 cases and 1061 controls and the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study with 275 cases and 674 controls. We use the TwinGene as the training study and the PIVUS as the validation study. A detailed description of the studies design and metabolomics experiments is given in the methods section.

Initially, we consider a situation where the validation study is not available. Because we do not know what the validation study will look like, we assume the same underlying effect sizes distribution, the same proportion of truly nonsignificant metabolic feature ($\pi_0$) and the same case–control ratio. Moreover, we fix the significance level in the validation population ($\alpha_v$) to 0.05.

The size of the training sample is specified in step 2a of the online software, while the significance level for validation is specified in step 3.
First, we want to estimate both $\pi_0$ and the underlying mixture of effect sizes in TwinGene. We apply a likelihood-based method described in the technical Appendix and previously in the literature [13] to estimate the parameters $\pi$ and $\delta/\sigma$. By minimizing the Akaike information criterion (AIC) we find that the optimal number of components for the mixture model is 5. We obtain the following estimates:

$$\hat{\pi} = \{0.49, 0.07, 0.12, 0.28, 0.04\}$$

$$\hat{\delta/\sigma} = \{0, -0.24, -0.12, -0.05, 0.18\}.$$

This can be done by selecting ‘Select nq automatically (using AIC)’ in step 1c of the online software.

Now all parameters ($\pi$, $\delta/\sigma$) are estimated and they can be used to estimate the RDR. We first show that there is a perfect overlap between the RDR estimated using the definition given in the technical Appendix and the simulated RDR (Figure 2a). The ‘simulated’ RDR is obtained by simulating training and validation samples with similar mixture of effect sizes, $\pi_0$ and case–control ratio observed in TwinGene. This can be done with the MAsim function implemented in the R-package OCplus. The RDR is then the average observed RDR over 20 simulated samples. Further, we want to calculate the RDR in TwinGene for different sample sizes of the validation study. We then plot the RDR when the sample size of the validation study is $n_v = 500, 1000$ and $1500$ (Figure 2b).

In the online software, this corresponds to select ‘RDR for fixed sample size’ in step 2 and specify the sample size of the validation study in step 2a.

As expected, the RDR increases with increasing sample size. The largest RDR improvement is obtained when the sample size increases from $n_v = 500$ to $n_v = 1000$. The RDR is a function of the critical value $c$ used in TwinGene. We explore three common methods to determine the critical value $c$: FDR = 5%, FDR = 1% and the Bonferroni correction. The Bonferroni method maximizes the RDR by reducing the number of true positives detected ($1 - \beta$, power, gray line in Figure 2b). It selects strong associations that can be easily replicated in the validation study. FDR = 5% has a higher power, correctly selecting 30% of the true positives compared with 9% of Bonferroni. The RDR is 68% when the sample size of the validation study is $n_v = 1000$. Thus, among all the selected features, 68% would be replicated at a $P$-value < 0.05. Setting a FDR = 1% as an in-between strategy, the power decreases to 20% and the RDR increases to 80%. However, it is unlikely

![Figure 2](https://example.com/figure2.png)

**Figure 2**: RDR from metabolomics study in TwinGene; 5363 metabolic features are tested. (a) Shows the estimated RDR using formula (2) (solid line) and the average simulated RDR over 20 simulations (dashed line), both based on $\pi_0$, effect sizes and case–control ratio calculated in TwinGene (training study). The estimated RDR and average simulated RDR are almost overlapped. (b) Shows the RDR when the critical value $c$ of the training study and the sample size of the validation study are varied. Solid line, dashed line and dotted line: RDR when the sample size of the validation study is fixed to $n_v = 500, 1000$ and $1500$, respectively. The gray solid line shows the power calculated in TwinGene. Three dotted–dashed lines are drawn in correspondence of three multiple-testing correction methods (FDR = 5%, FDR = 1% and Bonferroni).
that any false positive will be detected to be significantly associated in the validation study. The vFDR is expected to be 0.03%.

In Figure 3a we compare the observed RDR, using PIVUS data as the validation sample, with the estimated RDR calculated using (3). We use bootstrapping method to show the variability of the estimates. The observed RDR is larger than predicted. This is most probably due to a different underlying distribution of effect sizes in PIVUS compared with TwinGene. The plasma samples in PIVUS have been stored directly after collection, which increases the sample quality and allows the detection of stronger associations compared with TwinGene. Note, however, that the larger-than-predicted RDR constitutes a successful validation. We should be concerned if the RDR is much less than predicted. As an illustration of a situation where there is no heterogeneity between training and validation studies, we used TwinGene both as training and validation samples. To do that, we randomly split TwinGene in half and computed the observed and estimated RDR. The process was repeated 20 times. In this case, the average observed RDR is similar to the average estimated RDR (Figure 3b). The variability is larger than in Figure 3a owing to the smaller sample size.

**PRACTICAL CONSIDERATIONS**

So far, we have demonstrated the RDR as a function of the critical value $c$ in the training sample. Moreover, in most simulations, we have assumed the training and validation samples to have the same size. In practice, the critical value $c$ and the size of the validation sample can be set to maximize the RDR, while keeping the number of false positives under control. Specifically, we suggest the following:

1. To calculate the RDR before deciding the size of the validation study. In the metabolomics example above, it seems reasonable to have a validation sample size $n_v \geq 1000$, as there is a substantial RDR improvement over $n_v = 500$. If the recruitment of individuals or analysis of samples in the validation study is expensive, $n_v = 1000$ is the best choice, as 500 additional individuals do not substantially increase the RDR. As a rule of thumb, one might consider to set the RDR $= 0.8$ and obtain the estimated size for the validation sample.

2. To check the maximum RDR achievable given a certain sample size. This allows us to get a realistic expectation of our ability to replicate

---

**Figure 3**: Comparisons of observed RDR and estimated RDR in metabolomics study. In (a) the solid line shows the observed RDR obtained by using PIVUS as validation study ($n_v = 949$). The dashed line shows the estimated RDR assuming the validation study to have the same sample size of PIVUS and the same $\pi_0$, effect sizes and case−control ratio of TwinGene. The observed RDR is larger than predicted. (b) RDR calculated using half of TwinGene as training and the other half as validation. This illustrates a situation where there is no heterogeneity between training and validation studies. The RDR is plotted as a function of the critical value $c$ of the t-statistic. TwinGene is randomly split in half 20 times, and each gray line represents one of this split. The solid lines show the observed RDR and the dashed lines the estimated RDR. The thick black lines show the average RDR.
significant findings. In the metabolomics example, it is not possible to validate all the metabolic features, even when the Bonferroni correction is used to select highly significant features for validation. Importantly, this means that some true positives cannot be validated and not that the nonvalidated features are false positives!

(3) To compare the observed RDR with the expected RDR. This gives information about heterogeneity between training and validation studies. For example, the PIVUS study is more powerful than expected under the assumption of homogeneity.

(4) To decide a critical value $c$ in the training population bearing in mind that, in the validation population, additional false positives control is applied. In the metabolomics example, applying a Bonferroni correction in the training and using an additional $z_v = 0.05$ in the validation study corresponds to an expected vFDR of 0.0007%. If the aim is to maximize the number of true positives validated, it is worthwhile bringing some false positives to validation and filtered them out by controlling $z_v$. Alternatively, a control based on vFDR can be used in the validation sample. Being too strict on both sides is overly conservative.

DISCUSSION

It is a common and recommended practice in biomedical research to validate significant findings in a different population from the one where the findings were initially assessed. By evaluating the association in an additional study, the probability of getting a false positive in the training study is decreased. Also, if the training study is biased, there is a lower chance of getting the same bias in the validation study.

In this study, we describe the factors that influence the validation of significant findings and propose two quantities (RDR and vFDR) as tool to better plan the size of validation studies and/or deciding the optimal P-value thresholds. The vFDR is the proportion of false validated features among all those taken forward in the validation study, and the RDR is the expected proportion of findings validated among those declared significant in the training sample. Bogomolov and Heller [14] recently proposed a procedure to control the vFDR and described a measure, named $r$-value, to calculate the smallest FDR level at which a feature is called validated. This measure is used as feature-specific evidence toward reproducibility and can only be assessed after the validation study has been conducted. If the aim is to determine the expected success of a validation study, the RDR is a more appropriate measure.

In particular, we foresee two main applications. Firstly, if the validation population has not yet been collected, the RDR can be used to decide the optimal balance between the proportion of features taken forward to validation and the size of the validation study. Secondly, if a validation study has already been collected, the RDR estimated using the training data can be compared with the observed RDR from the validation data, providing an assessment of the success of the validation study.

In our analysis we assume that the training and validation samples are homogeneous, sharing the same underlying effect sizes and proportion of truly nonsignificant features. This assumption might not hold in practice. Heterogeneity between training and validation population is desired if the aim is to improve the generalizability of the results. In principle, if information about the underlying structure of the validation sample is available, this can be implemented in the RDR calculation. However, in the absence of this information, assuming homogeneity is the most reasonable approach. In our simulations we assume independence between features; this assumption might be violated in practice. Correlation between features does not create bias, but will increase the variability of the RDR estimates, similar to what is previously described for FDR [15].

In this article, we use the t-statistics for estimating the effect size, which is generally applicable in metabolomics experiment, GWAS and the other high-throughput studies. For other statistics S with null distribution $F$, one can first transform the statistic using $\Theta^{-1}[F(S)]$, where $\Theta(.)$ is the Gaussian distribution, so that the null distribution is Gaussian, and we can use the mixture model in this article. Improvements aimed to obtain a nonparametric RDR estimate are ongoing. We assume the t-statistics to follow the mixture distribution. The model using the mixture distribution is known to be highly flexible [16] and in principle can be well fitted to any distributional shape in real data by varying the number of mixture components. There are other methods to estimate the effect size. For example, in
GWAS, researchers usually define the effect sizes as the contribution of the single nucleotide polymorphism to genetic variance of trait. So, they use a skewed distribution (e.g. exponential distribution) as the effect size distribution [17, 18]. Future work is needed to extend the RDR and vFDR on different effect size definitions and distribution assumptions.

In summary, we explore the expected proportion required to validate significant findings in several scenarios. We formalize this concept by introducing the RDR and illustrate the benefits of using this measure for planning and assessing validation studies.

**METHODS**

**Theoretical and simulated examples**

To illustrate RDR properties we provide theoretical and simulated examples. In the multiple-testing theoretical examples, we vary the critical value $c$ of the training sample and typically fix the critical value $k$ of the validation sample so that $\alpha_v = 0.05$. We also explore scenarios with $\alpha_\ell = 0.005$ or 0.0005. The nonzero effect sizes are assumed to be a mixture of six components, i.e. $\delta = (-0.2, -0.1, -0.05, 0.05, 0.1, 0.2)$ with probability $(1 - \pi_0)/6$ and to be the same in training and validation samples. We also assume the same proportion of true null hypotheses ($\pi_0 = 0.9$ or $\pi_0 = 0.99$) in training and validation samples. The size of the training sample is $n_0 = n_{\ell1} = 500$, while we vary the size of the validation sample $n_{\ell0} = n_{\ell1} = n_0/2, n_0$ and $2n_1$.

In the simulations for the single hypothesis testing we obtained the simulated data set using the \texttt{MAsim} function implemented in the R-package \texttt{OCplus}. We fixed the effect size $\delta_i/\sigma$ to 0.5 and varied the sample size per group $n_0 = n_{\ell1}$ from 5 to 100 to obtain different RDR estimates. We also varied the effect size and fixed the sample size per group to 0 and 100. Each simulation setup was run 10 000 times. For each simulation, we got the RDR estimate, $RDR$, by using (2) and (5), and computed the median of $RDR$s for comparing with the theoretical power.

**Real-data example**

To investigate the RDR performances on real data, we studied the association between metabolites and hsCRP in two population-based studies. The TwinGene study is a longitudinal sub-study within the Swedish Twin Register, which is a population-based national register including >170 000 Swedish twins born between 1886 and 2000 [19]. In TwinGene, we performed a case-cohort design by selecting all the incident cases of coronary heart diseases, type 2 diabetes, ischemic strokes and dementia up to 31 December 2009 and a sub-cohort (controls) of 1643 individuals. Because it has been previously shown to improve the study efficiency [20], the sub-cohort was stratified on median age and sex, and for each of the four strata, we randomly selected a number of participants proportional to the corresponding number of cases in the strata. The following analysis was conducted only on the individuals from the sub-cohort because this is representative of the original population. Of them, 1541 passed the quality control and had hsCRP measured with a high-sensitivity method by SynchroN LX systems (Beckman Coulter).

The PIVUS (http://www.medsci.uu.se/pivus/pivus.htm) is a community-based study where all men and women at age 70 years living in Uppsala, Sweden, were invited to participate in 2001 [21]. Of the 2025 subjects invited, 1016 subjects participated. Blood samples were available for 972 participants selected for metabolomic profiling. Of them, 970 passed the quality control. HsCRP was measured with an ultrasensitive particle-enhanced immunoturbidimetric assay (Orion Diagnostica, Espoo, Finland).

All participants of both studies gave written informed consent, and the ethics committees of Karolinska Institutet and Uppsala University approved the study.

Metabolomics profiling was performed at the Proteomics and Metabolomics Facility, Colorado State University. Each sample was injected in nonconsecutive duplicates in a randomized manner and analyzed using ultraperformance liquid chromatography system (Waters Acquity UPLC system). Data were collected in positive ion mode. Scans were collected alternatively in MS mode at collision energy of 6 V and in idMS/MS mode using higher collision energy (15–30 V). Analytical procedures are described in detail elsewhere [22].

Bioinformatics analysis was conducted using XCMS [23], and details are described elsewhere [24], under review. In total, we identified 5363 common metabolic features in 2490 individuals without ongoing inflammatory or infectious disease ($\text{hsCRP} > 20 \text{ mg/dl}$).
TECHNICAL APPENDIX: DEFINITIONS AND RDR AND VFDR FORMULAS

The following list describes the elements used in the analysis.

1. Let \( n_0(n_{10}) \) and \( n_1(n_{11}) \) be the sample sizes of the non-diseased group and diseased group in the training (validation) sample. And, let \( n_i = n_{i0} + n_{i1} \) and \( n_v = n_{v0} + n_{v1} \).

2. Let \( T_{i0} \) and \( T_{i1} \) be the t-statistics for the \( i \)th gene in the training sample and validation sample \((i = 1, \ldots, N_i)\), respectively.

3. Given a critical value \( c > 0 \), we usually reject the null hypothesis if \(|T_{i0}| > c\). Let \( I = \{i \mid |T_{i0}| > c\} \). Only for \( i \in I \), \( T_{i0} \) is calculated. The critical value \( k \) used in the validation sample might differ from the critical value in the training sample, \( c \). Thus, in the validation sample, the association is declared significant if \(|T_{vi}| > k\).

4. Under null hypothesis of no risk difference, \( T_{i0} \) and \( T_{i1} \) are distributed according to central \( t \)-distribution with the degrees of freedom \( n_i - 2 \) and \( n_v - 2 \), respectively.

5. Under the alternative, the t-statistic has a noncentral \( t \)-distribution. For each \( i \), we assume the effect size \( D_i \) to be independent and identically distributed at discrete values \( \{\delta_0, \delta_1, \ldots, \delta_m\} \) with probabilities \( \pi_j = P(D_i = \delta_j) \) for \( j = 0, \ldots, m \). For \( j > 0 \), the nonzero effect size \( \delta_j/\sigma \) is the mean difference in standard deviation units, so ‘effect size = 1’ means a ratio of \( \sigma \) for the mean of the diseased versus the mean of nondiseased groups. The same effect sizes are assumed to be observed in the training and validation sample.

6. For each \( i \), \( T_{i0} \) and \( T_{i1} \) are assumed to be conditionally independent for given \( D_i \).

With the above assumptions, we have the distribution of \( T_{i0} \) as a mixture of the form

\[
F_i(t) = \sum_{j=0}^{m} \pi_j G_{i,j}(t),
\]

where \( G_{i,j}(t) = F(t; \sqrt{n_{i0}n_{11}}/n_i \delta_j/\sigma, n_i - 2) \) and \( \sum_{j=0}^{m} \pi_j = 1 \). Here, \( F(\cdot; \delta, \nu) \) is the \( t \)-distribution with noncentrality parameter \( \delta \) and degrees of freedom \( \nu \).

Given the critical value \( c > 0 \) for the training sample, we can compute the classical significance level, \( \alpha_i = 2[1 - G_{0,i}(c)] \). The power can be computed as \( 1 - \beta_i = \frac{1}{2} \sum_{j=1}^{m} \pi_j (1 - \beta_{j,i}) \) where \( 1 - \beta_{j,i} = 1 - G_{j,i}(c) + G_{j,i}(-c) \). Similarly we can compute these quantities for the validation sample by indicating subscript \( v \) instead of subscript \( i \), i.e. \( \alpha_v = 2[1 - G_{0,v}(c)] \) and \( 1 - \beta_v = \frac{1}{2} \sum_{j=1}^{m} \pi_j (1 - \beta_{j,v}) \) where \( 1 - \beta_{j,v} = 1 - G_{j,v}(c) + G_{j,v}(-c) \). Note that the critical value \( k \) can differ from \( c \) of the training sample.

Now we can compute the RDR and vFDR as a function of type-I error and power of both the training and validation data. Because \( R_i = \sum_{j=1}^{N_i} I(|T_{i0}| > c) \), \( R_v = \sum_{j=1}^{N_i} I(|T_{vi}| > c) \), we can show that

\[
\text{RDR} = \frac{P(|T_{i0}| > k, |T_{i1}| > c) |T_{i0}| > c)}{P(|T_{i0}| > c)}
\]

\[
= \frac{\sum_{j=0}^{m} \pi_j P(|T_{i0}| > c | D_i = \delta_j) P(|T_{i1}| > k | D_i = \delta_j)}{P(|T_{i0}| > c)}
\]

\[
= \frac{\pi_0 \alpha_i \alpha_v + \sum_{j=1}^{m} \pi_j (1 - \beta_{j,i}) (1 - \beta_{j,v})}{\pi_0 \alpha_i + (1 - \pi_0) (1 - \beta_i)}.
\]

Similarly, vFDR can be computed as

\[
\text{vFDR} = \frac{P(|T_{i0}| > k, |T_{i1}| > c, D_i = \delta_0) |T_{i0}| > c)}{P(|T_{i0}| > k, |T_{i1}| > c)}
\]

\[
= \frac{\pi_0 \alpha_i \alpha_v + \sum_{j=1}^{m} \pi_j (1 - \beta_{j,i}) (1 - \beta_{j,v})}{\pi_0 \alpha_i + (1 - \pi_0) (1 - \beta_i)}.
\]

TECHNICAL APPENDIX: ESTIMATION OF RDR AND VFDR

Because the RDR (2) and vFDR (3) formulae depend on the mixture model of the training sample (1), we need to estimate the unknown parameters \( \theta = (\pi, \delta, \sigma) \) where \( \pi = (\pi_0, \pi_1, \ldots, \pi_m) \), \( \delta = (\delta_0, \ldots, \delta_m) \) in (1). The number of components \( m \) can be selected by the AIC. Here, following [13], we apply a likelihood-based method to estimate the parameters by using the observed \( t \)-statistics in the training sample, \( T_{i0} = t_i \) \((i = 1, \ldots, N_i)\). The log-likelihood is

\[
l(\theta) = \sum_{i=1}^{N_i} \log \left( \sum_{j=0}^{m} \pi_j f_j(t_i) \right),
\]

where \( f_j(t) \) is the noncentral \( t \)-density with \( n_i - 2 \) degrees of freedom and noncentrality parameter
\( \sqrt{n_0n_1/n_0+n_1} \delta / \sigma \). To get the maximum likelihood estimates \( \hat{\theta} \) by maximizing (4), the EM algorithm is commonly used. However, to overcome its slow convergence, [13] has already derived a much faster algorithm based on prebinning the data into a relatively fine grid of intervals.

Given the complete set of parameter estimates \( \hat{\theta} \) and critical values \( c \) and \( k \), we can theoretically proceed to estimate the RDR (2) or vFDR (3) using the estimated type-I error, power, i.e.

\[
\text{RDR} = \text{RDR}|_{\theta=0}, \quad \text{and vFDR} = \text{vFDR}|_{\theta=0}. \tag{5}
\]

**TECHNICAL APPENDIX: R SCRIPT EXAMPLES**

The rdr R-package can be downloaded at http://www.meb.ki.se/personal/yudpaw/. The following R code (rdr.est function) can be used to calculate the RDR in a hypothetical scenario with 10,000 t-tests from 40 individuals (20 cases and 20 controls) in the training and validation sample. We applied a threshold of \(-\log_{10} P\)-value = 5 and \(-\log_{10} P\)-value = 2 in the training and validation sample, respectively. Similarly, the rdr.samplesize function can be used to calculate the validation sample size given a fixed RDR.

```r
library(rdr)
library(OCplus)

## Inputs
ng<-10000 ## number of tests
nt0<-nt1<-nv0<-nv1<-20 ## sample size for training and validation studies
c.t<-5; c.v<-2 ## cutoffs for t-test

## Vector of t-statistics in training sample - create difference between cases and controls
tstat<-rt(ng, df=10)

## RDR estimation using t-statistics of the training study
RDR.est<-rdr.est(tstat, ng=3, c.t=c.t, c.v=c.v, nt0=nt0, nt1=nt1, nv0=nv0, nv1=nv1)

## Given RDR, compute the sample size of the validation study
rdr.samplesize(RDR=0.8, c.t=c.t, c.v=c.v, nt0=nt0, nt1=nt1, cc.ratio=1, D=RDR.est$D,p0=RDR.est$p0, p1=RDR.est$p1)
```

**Key Points**

- Validation studies increase the generalizability of the results and decrease the possibility to report false-positive findings.
- Current multiple testing correction methods (e.g. Bonferroni and FDR) only focus on controlling the false positives in the training sample without considering the properties of the validation sample.
- RDR is the expected proportion of significant findings from a training sample that are replicated in a validation sample.
- If the validation study has not yet been performed, the RDR can be used to decide the optimal combination between the proportion of findings taken to validation and the size of the validation study.
- If a validation study has already been done, the RDR estimated using the training data can be used to assess the success of the validation study, by comparing the observed RDR from the validation data.

**FUNDING**

This work was partially founded by research grant from the Swedish Heart-Lung Foundation and Swedish Research Council, and partially supported by the 2014 Ewha Womans University Research Grant.

**References**


