A rigorous method for integrating multiple heterogeneous databases in genetic studies

József Bukszár[1] and Edwin JCG van den Oord[1]

[1] Center for Biomarker Research and Personalized Medicine, School of Pharmacy, Virginia Commonwealth University

Correspondence should be addressed to Edwin JCG van den Oord (ejvandenoord@vcu.edu)

ABSTRACT

The large number of existing databases provides a freely available independent source of information with a considerable potential to increase the likelihood of identifying genes for complex diseases. We developed a flexible framework for integrating such heterogeneous databases into novel large scale genetic studies and implemented the methods in a freely-available, user-friendly R package called MIND. For each marker, MIND computes the posterior probability that the marker has effect in the novel data collection based on the information in all available data. MIND 1) relies on a very general model, 2) is based on the mathematical formulas that provide us with the exact value of the posterior probability, and 3) has good estimation properties because of its very efficient parameterization. For an existing data set, only the ranks of the markers are needed, where ties among the ranks are allowed. Through simulations, cross-validation analyses involving 18 GWAS, and an independent replication study of 6,544 SNPs in 6,298 samples we show that MIND 1) is accurate, 2) outperforms marker selection for follow up studies based on *p*-values, and 3) identifies effects that would otherwise require replication of over 20 times as many markers.

AUTHOR SUMMARY

The large number of existing databases provides a freely available independent source of information with a considerable potential to increase the likelihood of identifying genes for complex diseases. We developed a flexible framework for integrating such heterogeneous databases into novel large scale genetic studies and implemented the methods in a freely-available, user-friendly R package called MIND. For each marker, MIND computes an estimate of the (posterior) probability that the marker has effect in the novel data collection based on the

information in all available data. For an existing data set, only the ranks of the markers are needed to be known, where ties among the ranks are allowed. MIND 1) relies on a realistic model that takes confounding effects into account, 2) is based on the mathematical formulas that provide us with the exact value of the posterior probability, and 3) has good estimation properties because of its very efficient parameterization. Simulation, validation, and a replication study in independent samples show that MIND is accurate and greatly outperforms marker selection without using existing data sets.

INTRODUCTION

During the past decade, databases related to the genetic basis of complex diseases have grown dramatically. Typical examples are gene expression data repositories, meta-analyses of genome-wide linkage scans, published candidate gene association studies, disease-specific biochemical pathways, and genome-wide association studies (GWAS). These databases provide a freely available independent source of information. Integrating this information in novel studies has great potential to increase the likelihood of identifying disease genes. First, the use of existing information may increase statistical power and reduce the risk of false discoveries through improving the prior probability that a marker is associated with the disease. Second, integrating data generated by other technologies may also reduce platform-specific errors and increase confidence in the robustness of the findings when multiple lines of evidence point to the same association. Third, because data integration considers multiple data sources, it may improve the understanding of disease mechanisms by informing the broader context in which disease genes operate[1].

Data integration may be particularly critical in large scale genetic studies of complex diseases. The reason is that rather than a few markers with large effects, many markers with small effects may be involved. Large sample sizes may therefore be required to find true positives while controlling false discoveries where the cost per sample in these high dimensional investigations is typically high. As a result, economic feasibility may interfere with designing adequately powered studies. Furthermore, with the exception of traits that are routinely measured in control groups of genetic studies (e.g. smoking[2-4]), for many disorders and outcomes studied (e.g. drug response) very large sample size may simply not be available. The use of

existing information may then become the only readily available option to detect small effects in a cost-efficient manner.

Because of their volume and heterogeneity, information from existing databases can no longer be integrated intuitively by investigators. This explains efforts towards developing more systematic data integration methods[5-12]. One limitation of these methods is that they lack a solid statistical basis. For example, most methods produce a cumulative measure of the biological relevance of genes after combining information across multiple sources. However, because it is usually very hard to asses the quality of that overall score, it is unclear how to use these scores in a way that information from different databases is used and combined optimally.

In this article we present a rigorous and flexible framework for integrating multiple heterogeneous existing data sets into novel studies aimed at identifying genes affecting complex diseases. We implemented the method in a freely-available R package called MIND (Mathematically-based Integration of heterogeNeous Data), that allows researchers to perform all analyses discussed in this paper through a single command line with 8 parameters. MIND can integrate existing data sets generated by any kind of technology (e.g., expression arrays, proteomics, GWAS) or activity (e.g., actual data collection, literature search, construction of disease-specific biochemical networks). Furthermore, external data may provide information at any genetic level ranging from individual variants (e.g., SNPs), genes (e.g., literature search), groups of genes (e.g., pathways), or entire chromosomal segments (e.g., linkage studies or targeted next-generation sequencing).

The end product of MIND is an estimate of the compound local true discovery rate (cℓTDR), which is the posterior probability that a genetic marker has an effect based on the information in the novel data collection and the existing data sets. The adjective “compound”

indicates that the cℓTDR capitalizes on (i.e. compounds) all disease relevant information present in the external data sets. The cℓTDR is a posterior probability because it also takes the results from the novel data collection into account. Finally, the term "local" reflects that the cℓTDR provides marker specific information. In scenarios where researchers are interested in groups of markers (e.g. pathways, top results) the cℓTDRs can simply be summed across all markers. The resulting cumulative cℓTDR is then the expected number of markers with effect in that group.

A very important feature of MIND is that, accurately modeling the data integration process and properties of data bases (e.g. number of effects differs across data sets), it relies on a solid mathematical foundation that provide us with the exact posterior probability that a marker has effect in the novel data collection based on the information in all available data sets. This ensures that MIND is more accurate than any heuristic or ad-hoc method. Furthermore, while the exact value of cℓTDR depends on many unknown parameters, most of the unknows can be collapsed into a single parameter for which we developed a precise estimator. As a result, the deviation of the estimate of the cℓTDR from its real value is only due to sample fluctuation, which was also verified by simulation.

A noteworthy property of MIND is that it merely requires that the information in existing data sets can be ranked. This ensures general applicability as ranks can almost always be calculated. For example, one could count the number of times genes co-occur with a specific disease in the literature or use only two ranks indicating whether a gene is implicated or not. Ranks also provide a robust method if there are concerns about the distributional assumptions of the test statistics in the external data sets.

We demonstrate MIND through simulations, cross-validation analyses involving 18 GWAS, and an independent replication study of 6,544 SNPs in 6,298 samples. The markers included in the replication study were selected based on a meta-analysis of the 18 GWAS. The results obtained with the markers selected by MIND are compared with a traditional *p*-value based SNP selection.

METHODS

The goal of MIND is to identify markers that are associated with a complex disease based on the test statistic values in the novel data collection (NDC) and on their ranks in the existing data sets (EDSs). **Figure 1** displays a schematic overview of the preparatory data transformation as well as the three major steps of the method. First, the existing data sets need to be transformed to the genetic level of the novel data collection if needed (e.g. assign the rank of a gene after an existing literature search to each SNP in that gene in the novel GWAS). The three major steps are 1) Compute for each existing data set the prior probabilities that markers are associated with the disease, 2) Combine the individual sets of prior probabilities into a single set of prior probabilities, and 3) Compute the cℓTDR for each marker.

Before discussing each of these three steps, we note for every EDS we only need the rank of their genetic units, whereas for the NDC we need the null and the alternative p.d.f. of the test statistics, f_0 and f_1, as well as the number of alternative genetic units in the NDC, m_1^{**}, or the estimates of them. Although estimating m_1^{**}, f_0 and f_1 is not part of our framework, we developed the estimators of m_1^{**}, f_0 and f_1 for the scenario where the test statistic is approximately normally distributed or its distribution is a mixture of normal distributions in the NDC. Furthermore, MIND allows genetic units to be different across the data sets involved. For

instance, if we have gene expression data, GWAS and linkage data as EDS, their genetic units are gene, SNP, and chromosomal segment, respectively. To handle these different units we first transform the EDSs into data sets based on the genetic unit of the NDC, which we call the test unit. For instance, a gene-based EDS can be transformed into SNP-based EDS by assigning to each SNP the smallest EDS rank (or p-value) of the genes that contain the SNP.

In the rest of this section we describe the three major steps. For each step we also present the mathematical formula based on which the step is carried out. The mathematical proofs for the formulas are provided in the Supplemental Material, Appendix.

Step 1: Obtaining prior probabilities of test units for each EDS: First we need two concepts to quantify information. We define the *information parameter* of an EDS to the NDC as

$$\kappa = m_1^{overlap} / m_1 - m_1^* / m, \tag{1}$$

where m denotes the number of test units that are both in the EDS and in the NDC, m_1 is the number of test units alternative in the EDS, m_1^* is the number of test units in the EDS that are alternative in the NDC, and $m_1^{overlap}$ is the number of test units that are alternative in the EDS and in the NDC. We will call an EDS *informative to* the NDC if its information parameter is positive. Note that κ is positive if, and only if, the number of test units alternative both in the NDC and the EDS is larger than it would by chance, i.e. when the alternative label would be randomly assigned to the test units of the EDS.

For test unit i in an EDS, we define *the contribution of test unit i from the EDS to the NDC* as

$$co(i) = (m\gamma(r_i) - m_1)\kappa / m_0, \tag{2}$$

where r_i is the rank of test unit i in the EDS, $\gamma(r)$ denotes the probability that a test unit ranked r in the EDS is alternative in the EDS, and $m_0=m-m_1$.

Based on the information in an EDS, for the prior probability that test unit *i* is alternative in the NDC, $\gamma^*(i)$, we have that

$$\gamma^*(i) = \begin{cases} co(i) + m_1^{**}/m^{**} & \text{if test unit } i \text{ is in the EDS} \\ m_1^{**}/m^{**} & \text{if test unit } i \text{ is not in the EDS} \end{cases} \quad (3)$$

where $co(i)$ is the contribution of test unit *i* from the EDS to the NDC, m^{**} and m_1^{**} is the number of test units and the number of alternative test units in the NDC, respectively (for the proof see Theorem 3 and Corollary 4 in **Appendix**). We remark that the information parameter being 0 implies 0 contributions for all test units, which results in $\gamma^*(i)= m_1^{**}/m^{**}$ for every test unit. Note that this is exactly what we would have in case of no prior information. Another property of the $\gamma^*(i)$ formula is that, as all the contributions of an EDS sum up to zero on the test units of the EDS (see Lemma 8 in **Appendix**), $\gamma^*(i)$ sum up to m_1^{**} on the test units for every EDS. In other words, using an EDS merely redistributes the total amount of prior probabilities among the test units.

The contribution of a test unit depends on 3 factors: 1) the rank of the test unit in the EDS, 2) the information parameter of the EDS and 3) the effect sizes in the EDS (see (2)). These latter two, intuitively speaking, stretch out the contributions, and hence amplify the redistribution of the prior probabilities. Indeed, larger information parameter or average effect size of the EDS make the contributions differ from each other more within the EDS. It is an advantage of our method, however, that we do not need to know the information parameter and the average effect size separately to obtain the prior probabilities, because only their combined effects matter, which we can estimate from the data.

To obtain (3) we utilized that in practice, we can approximate m_1^*/m by m_1^{**}/m^{**}, where the rationale is that the group of test units in the NDC we have EDS information for should

contain proportionally as many alternatives as the entire NDC does (see Corollary 4 in **Appendix**). On the other hand, if we think that the above assumption is violated, then we may be able to estimate m_1^* in the same way as m_1^{**} is estimated, and use Theorem 3 to obtain the prior probability estimates.

Step 2: Combining the sets of prior probabilities into a single set of prior probabilities: Once we have the prior probabilities for every EDS, we calculate the combined prior odd $\beta_i^{(\text{combined prior})} = \gamma_i^{(\text{combined prior})} / \left(1 - \gamma_i^{(\text{combined prior})}\right)$ that test unit i is alternative in the NDC by

$$\beta_i^{(\text{combined prior})} = \frac{m_1^{**}}{m_0^{**}} \prod_{j=1}^{k} \frac{m_0^{**} \gamma^{*j}(i)}{m_1^{**} \left(1 - \gamma^{*j}(i)\right)}, \tag{4}$$

where $\gamma^{*j}(i)$ is the jth EDS-based prior probability that test unit i is alternative in the NDC, and $m_0^{**} = m^{**} - m_1^*$ (see eq. 13 in **Appendix**). Note that if we have no prior information for a test unit in any EDS, then from the formula in (4) we obtain that the combined odd of this test unit is m_1^{**} / m_0^{**}, which is exactly what we supposed to have in the case of no prior information. Moreover, according to the formula in (4), the combined odd of a test unit is proportional to the average of the prior odds of the test unit across the EDSs, where by the average we mean the geometrical mean. If a test unit performs better than a test unit with no information in an EDS (odd = m_1^{**} / m_0^{**}), then its odd in that EDS will have a positive (increasing) impact on its combined odd, and vice versa, i.e. if a test unit performs worse than a test unit with no information in an EDS, then its odd in that EDS will have a negative (decreasing) impact on its combined odd.

Step 3: Computing cℓTDR for each test unit: The cℓTDR of test unit i can be written as

$$c\ell TDR(i) = \frac{\beta_i^{(\text{combined prior})} f_1(t_i)}{f_0(t_i) + \beta_i^{(\text{combined prior})} f_1(t_i)}, \tag{5}$$

where f_0 and f_1 is the null and alternative p.d.f. in the NDC, respectively, and where t_i is the observed test statistic value of test unit *i* in the NDC (see Claim 24 in **Appendix**). In summary, combining equations in (3), (4) and (5) we obtain the cℓTDR of a test unit as a function of m_1^{**}, f_0, f_1, and the contributions of test units from each EDS. Instead of the terms cℓTDR depends on, we will use their estimates to obtain estimate of the cℓTDR. In the next section we present a method that estimates the contributions.

Estimating the contributions

As mentioned above, in order to estimate cℓTDR by our formulas we need to estimate the contributions of test units from an EDS to the NDC, defined in (2). Because we use the same procedure to estimate contributions for each EDS, throughout this subsection we assume that we have a single EDS, which we will refer to as the EDS. As we focus on the test units in the NDC, it is irrelevant whether the EDS contains test units not in the NDC or not, so for the sake of simplicity, we assume that the test units the EDS contains are also in the NDC. For estimating the contributions we will use the statistic

$$O_{d,M} = \#\left\{j \ : \ |t_j| \geq d, r_j \leq M\right\} - \#\left\{j \ : \ |t_j| \geq d\right\} M/m, \tag{6}$$

where #*A* denotes the number of elements in set *A*, t_j is the NDC test statistic of test unit *j*, and r_j is the rank of test unit *j* in the EDS. In **Appendix** (Theorem 25) we proved that for any positive integer $M \leq m$ and real number $d \geq 0$ we have that

$$E\left(O_{d,M}\right) = \left(F_0(d) - F_1(d)\right) \sum_{j, r_j \leq M} co(j), \tag{7}$$

where F_0 and F_1 is the null and alternative c.d.f. in the NDC, respectively. Based on eq (7), first we calculate a rough estimate of the cumulative contribution, defined as $CO(M) = \sum_{j, r_j \le M} co(j)$ by

$$\overline{CO}(M) = \frac{1}{\#D} \sum_{d \in D} \frac{O_{d,M}}{F_0(d) - F_1(d)}, \tag{8}$$

where D is a set of the positive real numbers, and $\#D$ denotes the number of elements in D. As the contribution of the test unit whose rank is r in the EDS can be obtained as

$$co(r) = CO(r) - CO(r-1),$$

we can calculate the contribution estimates from estimates of the cumulative contribution. To ensure that test units with smaller (better) ranks have larger contribution estimates, we need to use a cumulative estimate that is a concave function of M. For this we construct a concave function of M that fits $M \rightarrow \overline{CO}(M)$ well (see Section 1.3 in **Appendix** for details).

RESULTS

Accuracy

To study the accuracy of MIND, we simulated 500 studies with 3 existing data sets and a novel data collection consisting of one million markers of which 5,500 had a small effect (see **Supplemental Material** for details). In **Figure 2**, we show the estimates of the cumulative cℓTDR, defined as the sum of the k largest cℓTDRs as a function of k. The cumulative cℓTDR at k equals the expected number of markers with effect among the k markers with the largest cℓTDRs. For the sake of comparison, in the figure we also show the corresponding curves where no existing data sets were used to compute the cℓTDR estimates. Each curve in Figure 2 is the average of the corresponding curves in the 500 simulation studies. The fact that the lines overlap

perfectly implies that on average the estimated cumulative cℓTDRs are very precise indicators of the number of markers with effect in the novel data. Moreover, we found that the cumulative cℓTDR differs from the number of markers with effect among the markers with largest cℓTDRs by less than 9.6% of the number of markers with effect among the markers with largest cℓTDRs in 99% of the simulations studies, and the percentage difference gets smaller as the number of selected markers increases (see **Supplemental Material** for details). This shows that the estimated cumulative cℓTDR is an accurate predictor of the number of markers with effect among the markers with largest cℓTDRs . Comparing in **Figure 2** results for the cℓTDR when the existing data sets were used versus when no existing data sets were used shows how data integration increases the proportion of markers with effect among markers selected by their cℓTDR.

Illustration with empirical data

We illustrate our framework using a meta-analysis of 18 schizophrenia GWAS studies comprising a total of 21,953 cases and controls. Even after including study-specific principal components to control for stratification, the meta-analyses suggested the presence of many SNPs with very small effects (consistent with a previous publication[13]). To distinguish and select these very small effects from the markers with no effects, we used MIND. Nine external data sets with potential relevance for schizophrenia were tested for information content 1) schizophrenia candidate genes[14], 2) the top bins from a meta-analysis of linkage scans[15], 3) results from an expression array meta-analysis using post-mortem brain tissue from schizophrenia cases[16], 4) a global proteomic analysis in post-mortem prefrontal brain tissues[17], 5) CNVs associated with schizophenia[18], 6) disease genes in the OMIM database[19], 7) gene length, 8) gene expression

quantitative trait loci (eQTLs)[20-25], and 9) human orthologs of murine genes showing association with behavioral phenotypes relevant to neuropsychiatric outcomes[26]. Six (above data sets 1, 2, 3, 6, 8, and 9) out of these 9 external data sets appeared informative for the GWAS meta-analyses and were included in subsequent analyses. We note that our finding that eQTL data are informative for GWAS is consistent with other reports in the literature[27].

In **Figure 3** we use the (meta-analyses of) expression data to graphically illustrate informativeness. To each GWAS SNP that was ±50kb of a gene in the eQTL dataset, we assigned the rank of the *p*-value of that gene in the expression data. We picked the smallest *p*-value if there were multiple *p*-values per gene. A total of 441,392 GWAS SNPs could be assigned a rank. The x-axis shows the top *j* SNPs according to their rank in the expression data set. The purple line gives the relation as observed in the data, and the many thin grey lines show results from 1,000 generated existing data sets obtained by randomly permuting the ranks of genes in the expression data set. The figure shows that up to about the first 40,000 SNPs, SNPs that are in genes that rank higher in the expression data also have better *p*-values in the GWAS and that this pattern is unlikely to occur by chance.

As an initial "internal" validation, we compared the 5,000 SNPs with the best *p*-values versus the 5,000 SNPs with the best cℓTDRs in terms of the heterogeneity of effects and gene ontology (see **Supplemental Material).** The I^2 index[28] was used to study heterogeneity that could, for example, be increased due to technical errors affecting results from only one or a few GWAS. Integrating data generated by other technologies may reduce the effect of such technical errors. We therefore hypothesized that compared to *p*-value selected SNPs, the cℓTDR selects SNPs that show more consistent effects across the 18 GWAS in our meta-analysis. Results in **Supplemental Figure 2** confirm that this is indeed the case. The gene ontology analysis is

motivated by the observation that most biological functions seem to be carried out by co-regulated "modules" (e.g. pathways, complexes)[29]. Some of these modules could possibly be pathogenic, implying that disease genes may share gene ontology terms. For this specific analysis we only integrated the empirical external data sets (e.g. linkage analyses, expression array) to avoid that differences were introduced by using databases that we (partially) generated using biological knowledge (e.g. candidate gene studies). Results (see **Supplemental Figure 3**) support for the notion that data integration more successfully identifies gene ontology terms thereby improving our understanding of disease mechanisms.

Validation

We validated the ability of our data integration method to identify markers with effects via 1) simulation studies, 2) cross-validation, and 3) and actual replication study of 6,544 SNPs in a sample independent of the 18 GWAS studies that included 6,298 subjects from 1,811 nuclear families. In the simulation studies we used the same parameters as used in the accuracy section above. For cross-validation, presenting a more realistic test case (e.g. actual effects sizes, artifacts, LD among markers), we selected subsets from all 18 GWAS in such a way that the sample size available for selecting SNPs was 85-90% of the total sample size. The remaining studies were used for replication/cross-validation. For each of the 575 unique cross-validation combinations, we selected the 5,000 SNPs with the smallest *p*-values and the 5,000 SNPs with the best cℓTDR after integrating our six informative existing data sets. The replication study involving genotyping of 6,544 SNPs in independent samples was conducted using a custom Illumina iSelect chip. About half of the SNPs were selected based on having the smallest *p*-values and the other half based on having the best cℓTDRs.

Results are shown in **Figure 4a**, **b**, and **c**. All three panels converge to the same conclusions. First, considering the cℓTDR (green and blue dots) always gives better results compared to SNP selection based on *p*-values alone. Second, although the success of MIND decreases as *p*-values increase, in many instances it still successfully identifies SNPs with effects that have large *p*-values with ranks >100,000 in the GWAS meta-analysis. Even if costs to follow up that many SNPs would not be an issue, it may still not produce equally good results because, compared to the much smaller set of cℓTDR selected SNPs, many more tests would need to be performed in the replication study. Third, these cℓTDR selected SNPs with *p*-value ranks >100,000 in the meta-analysis replicate as well as or better than SNPs with small *p*-values but poor cℓTDRs.

DISCUSSION

We developed a mathematically rigorous and flexible framework for integrating heterogeneous databases into large-scale genetic studies, and implemented our method in a freely available user-friendly R package called MIND. Through simulations, cross-validation analyses involving 18 GWAS, and independent replication of 6,544 SNPs in 6,298 samples we show that MIND 1) is accurate, 2) outperforms marker selection for follow up studies based on *p*-values, and 3) is able to identify effects that would otherwise require replicating over 20 times more markers.

Although a main application of MIND involves integrating existing data in a novel data collection, it is applicable in other scenarios as well. For example, MIND can rank genes according to their relevance to a disease using only existing databases. However, it is typically difficult to assess the quality of such prioritization scores; our framework provides an estimate of the probability that a gene is associated with the disease of interest. This clear interpretation allows for more informed decisions about which genes to select for further study. A second

example involves questions related to similarities among multiple high dimensional data sets. For example, if we have datasets for different diseases in the same population, MIND can be used to study co-morbidity where a high concordance would indicate a substantial overlap in disease etiology. Alternatively, if we have datasets for the same disease in different populations, MIND would shed light on the overlap in the genetic disease architecture of the different populations.

We should stress that MIND can handle novel data collections of any kind. Next-generation DNA sequencing (NGS) has the potential to accelerate genetic research. However, because costs for NGS are still high and power to detect the (cumulative) effects of all rare variants low[30], data integration could play an important role. Indeed, several methods have already been proposed that test for association using weights based on predictions of functional effects of (rare) variants (e.g. [31-32]). However, as these weights do not take the strength of disease relevant information into account, our method could be used to further optimize these tests. A second example is that NGS enables a comprehensive analysis of not just genomes but also transcriptomes and methylomes. MIND offers the possibility to integrate all these different sources of information to improve statistical power, increase confidence in the robustness of the findings when multiple lines of evidence converge to the same genetic factors, and inform the broader context in which the disease genes operate.

SOFTWARE

MIND has been made freely available as an R package at http://www.people.vcu.edu/~jbukszar/.

ACKNOWLEDGMENTS

J.B. and EvdO are supported by grants R01HG004240; R01MH078069; and HG004240-02S1.

1 Zhong, H., Yang, X., Kaplan, L. M., Molony, C. & Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet* **86**, 581-591, doi:S0002-9297(10)00102-3 [pii]

10.1016/j.ajhg.2010.02.020 (2010).

2 Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**, 441-447, doi:ng.571 [pii]

10.1038/ng.571 (2010).

3 Thorgeirsson, T. E. *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**, 448-453, doi:ng.573 [pii]

10.1038/ng.573 (2010).

4 Liu, J. Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-440, doi:ng.572 [pii]

10.1038/ng.572 (2010).

5 Yu, W., Wulf, A., Liu, T., Khoury, M. J. & Gwinn, M. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases. *BMC Bioinformatics* **9**, 528, doi:1471-2105-9-528 [pii]

10.1186/1471-2105-9-528 (2008).

6 Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* **37**, W305-311, doi:gkp427 [pii]

10.1093/nar/gkp427 (2009).

7 Radivojac, P. *et al.* An integrated approach to inferring gene-disease associations in

humans. *Proteins* **72**, 1030-1037, doi:10.1002/prot.21989 (2008).

8 Franke, L. *et al.* Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet* **78**, 1011-1025, doi:S0002-9297(07)63922-6 [pii]

10.1086/504300 (2006).

9 Gaulton, K. J., Mohlke, K. L. & Vision, T. J. A computational system to select candidate genes for complex human traits. *Bioinformatics* **23**, 1132-1140, doi:btm001 [pii]

10.1093/bioinformatics/btm001 (2007).

10 George, R. A. *et al.* Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* **34**, e130, doi:gkl707 [pii]

10.1093/nar/gkl707 (2006).

11 Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**, 537-544, doi:nbt1203 [pii]

10.1038/nbt1203 (2006).

12 Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* **22**, 773-774, doi:btk031 [pii]

10.1093/bioinformatics/btk031 (2006).

13 Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:nature08185 [pii]

10.1038/nature08185 (2009).

14 Allen, N. C. *et al.* Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat Genet* **40**, 827-834, doi:ng.171 [pii]

10.1038/ng.171 (2008).

15 Ng, M. Y. *et al.* Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry* **14**, 774-785, doi:mp2008135 [pii]

10.1038/mp.2008.135 (2009).

16 Higgs, B. W., Elashoff, M., Richman, S. & Barci, B. An online database for brain disease research. *BMC Genomics* **7**, 70, doi:1471-2164-7-70 [pii]

10.1186/1471-2164-7-70 (2006).

17 Martins-de-Souza, D. *et al.* Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia. *Eur Arch Psychiatry Clin Neurosci* **259**, 151-163, doi:10.1007/s00406-008-0847-2 (2009).

18 Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet* **25**, 528-535, doi:S0168-9525(09)00202-9 [pii]

10.1016/j.tig.2009.10.004 (2009).

19 McKusick, V. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588--604 (2007).

20 Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107, doi:07-PLBI-RA-4030 [pii]

10.1371/journal.pbio.0060107 (2008).

21 Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494-1499, doi:ng.2007.16 [pii]

10.1038/ng.2007.16 (2007).

22 Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**,

1217-1224, doi:ng2142 [pii]

10.1038/ng2142 (2007).

23 Veyrieras, J. B. *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* **4**, e1000214, doi:10.1371/journal.pgen.1000214 (2008).

24 Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772, doi:nature08872 [pii]

10.1038/nature08872 (2010).

25 Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777, doi:nature08903 [pii]

10.1038/nature08903 (2010).

26 Konneker, T. *et al.* A searchable database of genetic evidence for psychiatric disorders. *Am J Med Genet B Neuropsychiatr Genet* **147B**, 671-675, doi:10.1002/ajmg.b.30802 (2008).

27 Nicolae, D. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *Plos Genet* **6**, doi:10.1371/journal.pgen.1000888 (2010).

28 Higgins, J. P., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557-560 (2003).

29 Alon, U. Biological networks: the tinkerer as an engineer. *Science* **301**, 1866-1867 (2003).

30 Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* **11**, 773-785, doi:nrg2867 [pii]

10.1038/nrg2867 (2010).

31 Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832-838, doi:S0002-9297(10)00207-7 [pii]

10.1016/j.ajhg.2010.04.005 (2010).

32 Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384, doi:10.1371/journal.pgen.1000384 (2009).

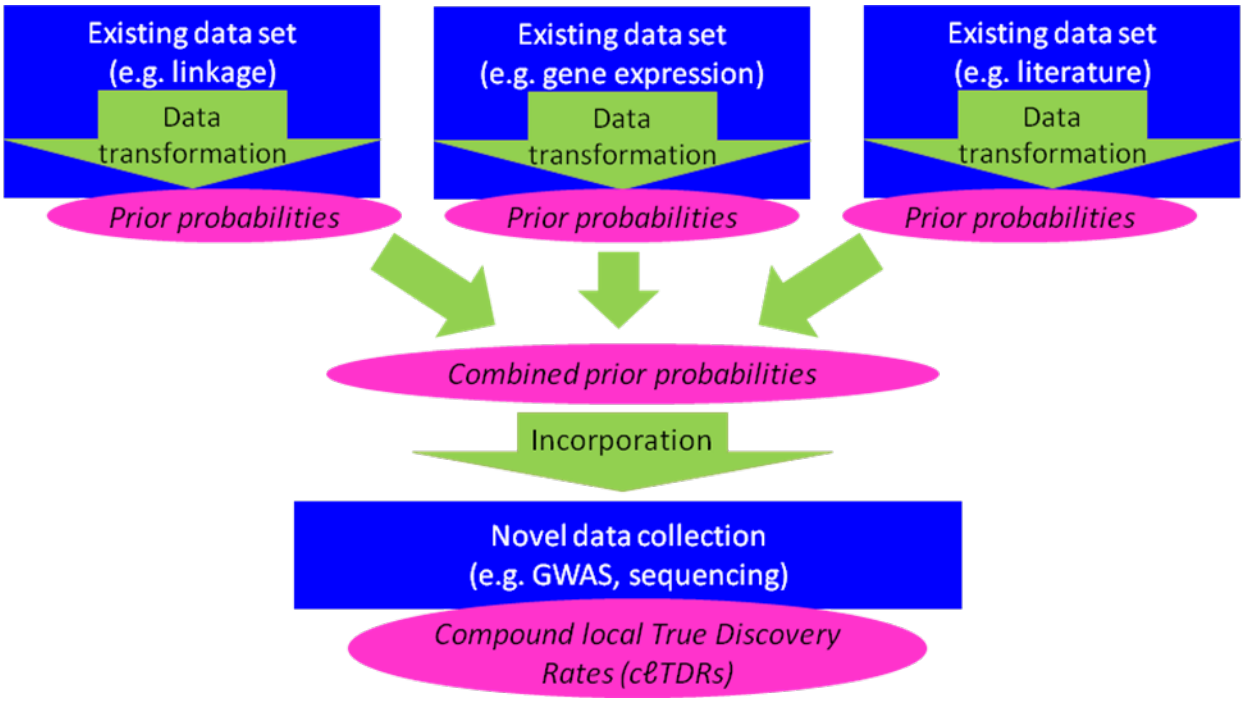
FIGURE LEGENDS

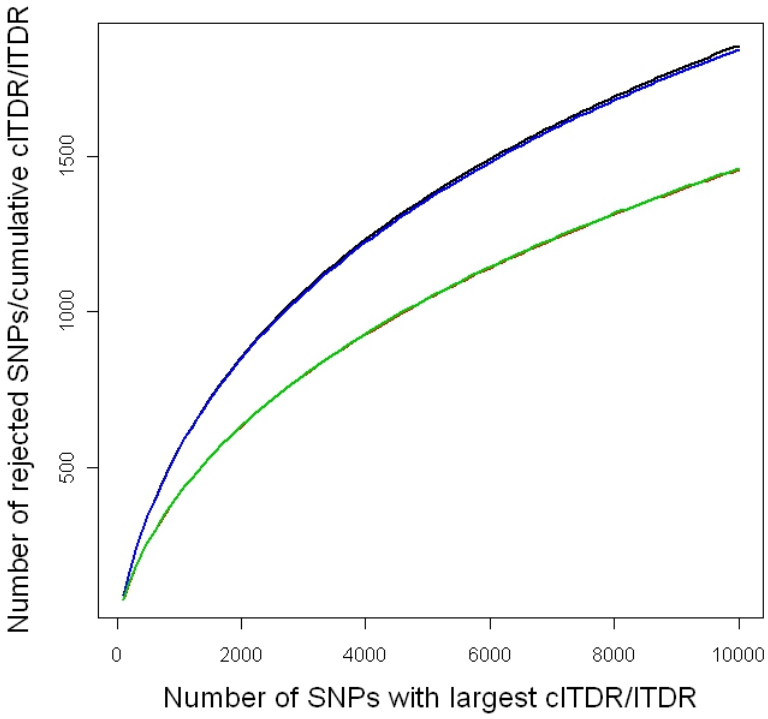
Figure 1 A schematic overview of the MIND data integration framework.

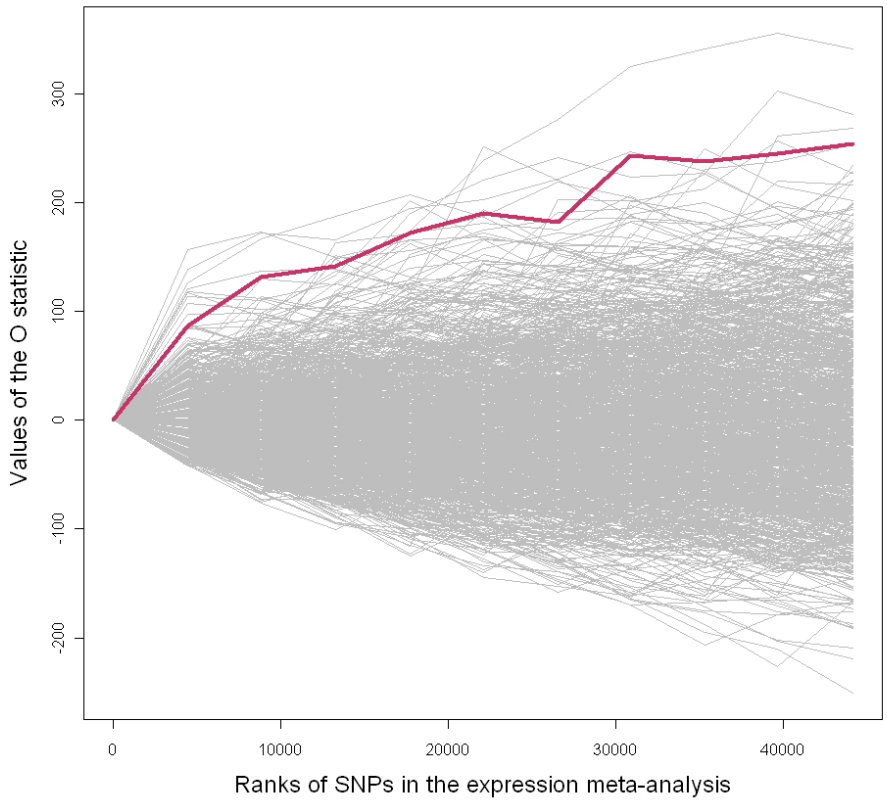
Figure 2 The estimated cumulative cℓTDR (black) as well as the number of markers with effect among the markers with the largest cℓTDRs (blue) curves are plotted. We also plotted the corresponding curves, where no existing data sets were used to compute the cℓTDR estimates (red and green). Each curve is the average of the corresponding curves in the 500 simulation studies.

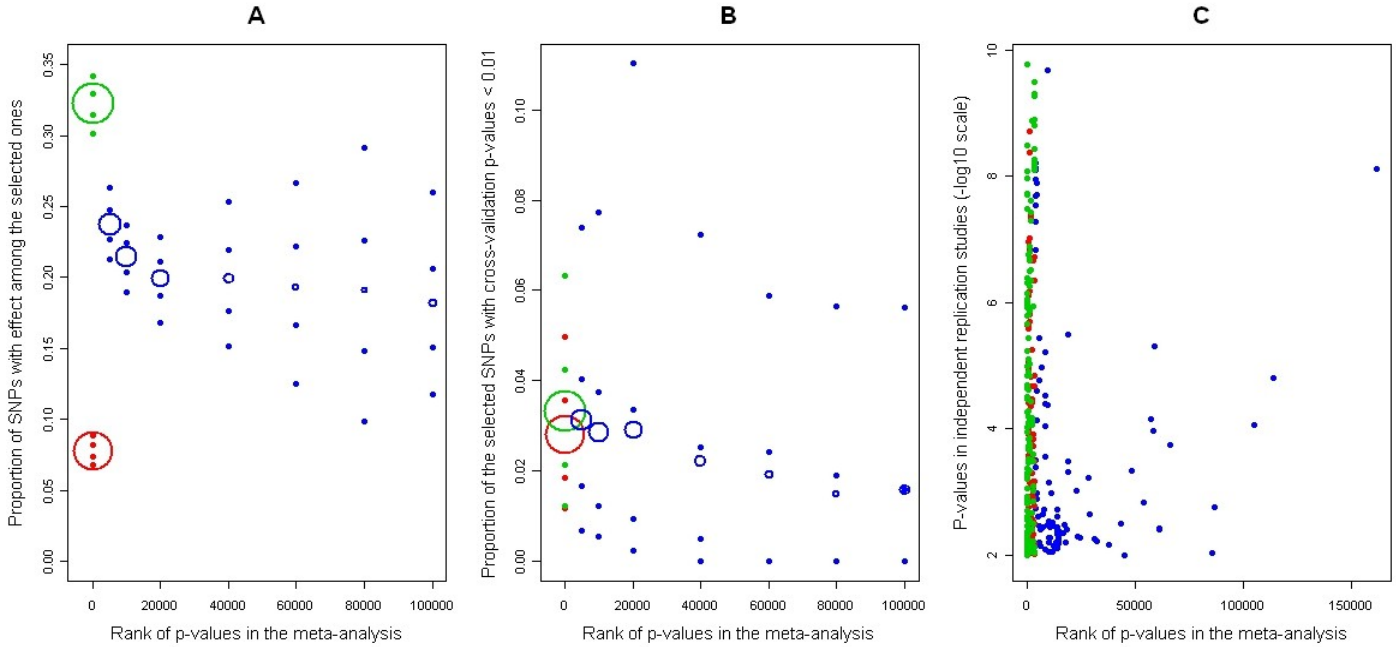
Figure 3 Enrichment in GWAS *p*-values (y-axis) in the set of SNPs that are in genes with higher ranks in the expression data. The purple line represents the expression data set, and each of the thin grey lines shows results from 1,000 data sets generated by random permutation.

Figure 4 Performance of *p*-value versus cℓTDR based selection by simulation (A), cross-validation (B), and replication (C). All panels: *p*-value based results are red, cℓTDR based results are blue, overlap between both methods is green, and x-axis is rank of p-values in the meta-analysis. Panels A and B: For each x-axis interval, the .05, .25, .75 and .95 quantiles of the proportion of SNPs with effect among those selected (A) or proportion of SNPs with cross-validation *p*-value less than 0.01 among those selected (B) are reported. Center of the circles are located at the mean of the proportions, while the area of the circle is proportional to the number of SNPs selected. Panel C: all *p*-values less than 0.01 in the replication study are shown.








A
Proportion of SNPs with effect among the selected ones
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
20000
40000
60000
80000
100000
Rank of p-values in the meta-analysis
B
Proportion of the selected SNPs with cross-validation p-values < 0.01
0.00
0.02
0.04
0.06
0.08
0.10
0
20000
40000
60000
80000
100000
Rank of p-values in the meta-analysis
C
P-values in independent replication studies (-log10 scale)
2
4
6
8
10
0
50000
100000
150000
Rank of p-values in the meta-analysis

Supplemental material "A rigorous method for integrating multiple heterogeneous databases in genetic studies"

1. Simulation study

In order to study how accurately the cumulative cℓTDR predicts the number of selected SNPs that have effect in the novel data collection, we used 500 simulations. In each simulation, we generated 3 existing data sets and a novel data collection. For the novel data collection we simulated 1,000,000 test statistic values, 5,500 of which had effect. To generate an existing data set we randomly chose 500,000 of the 1,000,000 NDC SNPs to be matched and simulated test statistic values for them, 50,000 of which had effect. The "alternative in the EDS" label was randomly assigned to SNPs in such a way that the number of SNPs alternative both in the NDC and EDS was 2,200. We calculated the ranks of the test statistics in the EDSs, which were used for the data integration. The statistic values were drawn from the normal distribution with variance 1. The mean of the normal distribution was 1.6 and 2.0 for the SNPs with effect in the existing data sets and the novel data collection, respectively, and the mean was 0 for the null SNPs for every data set. As estimating the null and alternative distribution of the statistics as well as the number of SNPs with effect in the novel data collection is not part of our method, we used the 'real' functions and number of SNPs with effect for our procedures.

The average of estimated cumulative cℓTDR (=after data integration) and ℓTDR (=before data integration) curves as well as the number of SNPs with effect in the NDC in the top SNPs for cℓTDR and ℓTDR – based selection are plotted in the left panel of Figure S1 (identical to Figure 1 in the article). Each curve is the average of the corresponding curves in the 500 simulation studies. The figure suggests that the cumulative cℓTDR and the cumulative ℓTDR curves are unbiased predictors of the number of SNPs with effect in the NDC that were selected by the corresponding method. To further study the accuracy of these predictors, we calculated the percentage difference between the cumulative cℓTDR/ℓTDR and the number of SNPs with effect selected by cℓTDR/ℓTDR, that is the absolute value of the difference between the two expressed with the percentile of the number of SNPs with effect selected by cℓTDR/ℓTDR. The quantiles of the percentage differences in the 500 simulation studies are plotted in the right panel in Figure S1. For instance, the continuous black curve in the figure shows that in 99% of the simulation studies the percentage difference between the cumulative cℓTDR and the number of selected SNPs with effect in the NDC was always less than 9.6%, and less than 7.5% if more than 5,000 SNPs were selected. The 50%, 75%, 90% and 99% quantile curves of the percentage difference between the cumulative cℓTDR/ℓTDR and the actual number of selected test units show that percentage difference 1) gets smaller as the number of selected markers gets larger and 2) is smaller for cℓTDR selection than for ℓTDR selection when the number of selected markers is small and comparable otherwise. We conclude that the estimated cumulative cℓTDR is a good predictor of the number of markers with effect among the selected ones and may be limited only by the imperfection of the estimate of the null and alternative distribution of the statistics and the number of markers with effect in the novel data collection, which estimate is, however, not part of our method.

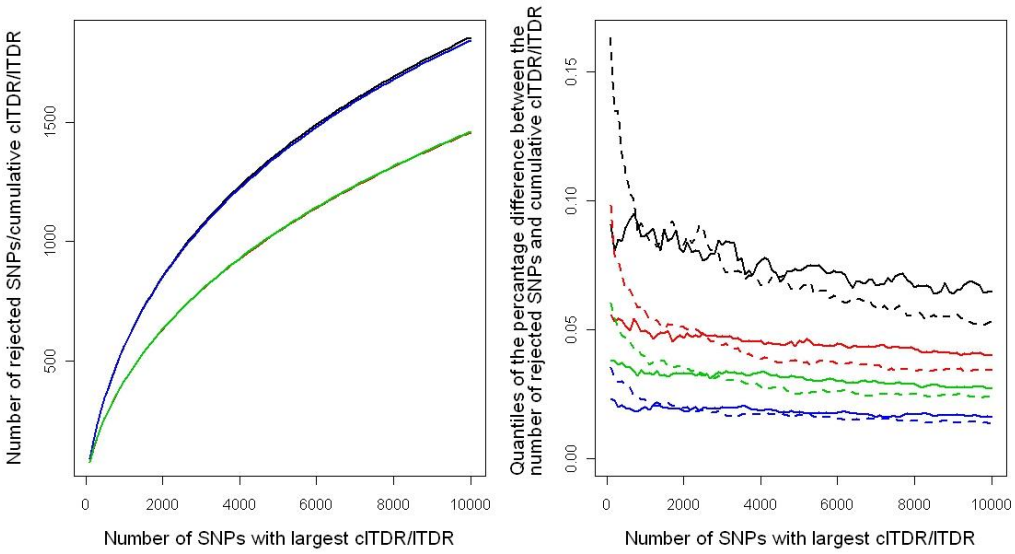


Figure s1 Left panel: The estimated cumulative ℓTDR/cℓTDR (red and black) as well as the number of SNPs with effect among the ℓTDR/ cℓTDR selected ones (blue and green) curves are plotted. Each curve is the average of the corresponding curves in the in the 500 simulation studies. Right panel: Multiple quantiles of the absolute value of percentage difference between the estimated cumulative ℓTDR/ cℓTDR and the number of SNPs with effect among the cℓTDR selected ones are plotted. Black, red, green and blues curves represent the 99%, 90%, 75% and 50% quantiles, respectively, while continuous and dashed curves represent cℓTDR and ℓTDR selection,respectively.

16.4%, 9.9%, 6.1%, respectively, for the 99%, 90% and 75% confidence intervals.

2. Data sets and QC

GWAS meta-analysis: In our empirical example we used a meta-analysis we performed involving 18 schizophrenia GWAS studies. After stringent QC 1,085,772 (imputed) SNPs were available for 21,953 subjects (11,185 cases and 10,768 controls). To account for possible population stratification effects within each of the GWAS studies, we included the first 3 principal components obtained with EigenSoft[14] plus any additional principal components if they significantly ($p < 0.05$) predicted case-control status.

External data sets: Our external data sets included 1) schizophrenia candidate genes from the SZgene data base[15] that summarizes the results of 1,617 studies reporting on 952 candidate genes, 2) the top bins from a meta-analysis of 32 independent genome-wide linkage scans that included 3,255 pedigrees with 7,413 genotyped cases affected (see Table 2)[16], 3) results from an expression array meta-analysis of 12 controlled studies across 6 different microarray platforms using brain tissue from schizophrenia, bipolar, and controls (about 35 subjects in each group)[17], 4) a global proteomic analysis in post-mortem prefrontal brain tissues of 9 schizophrenic patients and 7 controls[18], and 5) replicated and significant CNVs (see Table 2[19]) from 10 studies. Other data sets involve features of disease genes in general such as 1) genes present in the OMIM database, 2)

gene length, disease genes are suggested to be longer[20], 3) SNPs that are strongly associated with variation in transcript abundance in the following tissues: liver, cortex and large B-Cell lymphomas using the eQTL browser at U. Chicago[21-26], and 4) human orthologs of murine genes showing association with behavioral phenotypes relevant to neuropsychiatric outcomes[27].

3. Details results and items in the text

Test results for informativeness: Test results for informativeness are shown in **Table S1** and indicate that six out of the 11 existing data sets appeared informative for the GWAS meta-analyses. The non-informative existing data sets are the ones with small samples sizes or not directly related to schizophrenia.

Table S1. Information tests for existing data sets

Data type	Source	Unit	Informative?
Meta-analysis expression studies			
Schizophrenia	Stanley	Gene	Yes
Bipolar	Stanley	Gene	No
Candidate genes			
Schizophrenia	SZgene database	Gene	Yes
Bipolar	SLEP + literature	Gene	No
Disease genes	OMIM	Gene	Yes
Meta-analysis schizophrenia linkage studies	Nga et al. (2009).	Region	Yes
Human (neurological) genes with mouse orthologs	SLEP	Gene	Yes
Candidate schizophrenia CNV regions	Sebat et al. (2009)	Region	No
Gene length	ENSEMBL database	Gene	No
Schizophrenia proteomics	de-Souza et al (2009)	Gene	No
Expression QTLs	Browser at U. Chicago	SNP	Yes

Figure 2: Figure 2 allows a visual inspection of informativeness. The x-axis shows the top *j* SNPs according to their rank in the expression data set. The y-axis shows a measure of enrichment, indicating whether GWAS *p*-values of SNPs are better for the genes that rank higher in the expression data. More precisely, if p_i indicates the *p*-value of SNP *i* in the GWAS, r_i indicates the rank of SNP *i* in the expression data, and *j* is an arbitrary a cut-off for the rank of SNPs in the expression data, then enrichment (O statistic) is defined as

$$O(j) = \#\{p_i : p_i \le 0.01, r_i \le j\} - (\#\{p_i : p_i \le 0.01\})\frac{j}{m}$$

where *m* is the number of SNPs in the expression data. Clearly, $O(j)$ being positive suggests enrichment of GWAS p-values smaller than 0.01 among the SNPs ranked j or better in the expression data. The purple line in Figure 2 gives the observed relation and the many thin lines show results from 1,000 generated existing data sets obtained by randomly permuting the ranks of genes in the expression data set. The results show that up to the first ~40,000 SNPs, SNPs that are in genes that rank higher in the expression data also have better *p*-values in the GWAS and that this pattern is unlikely to occur by chance.

Heterogeneity: To study heterogeneity we used the index[4] $I^2 = 100\% \times (Q - df)/Q$, where (Cochran's) Q is computed by summing the squared deviations of each study's estimate from the overall meta-analytic estimate, weighting each study's contribution in the same manner as in the meta-analysis. I^2 describes the percentage of total variation across studies due to heterogeneity rather than chance. Larger values show increasing heterogeneity and the (lower bound) value of 0% indicates no heterogeneity.

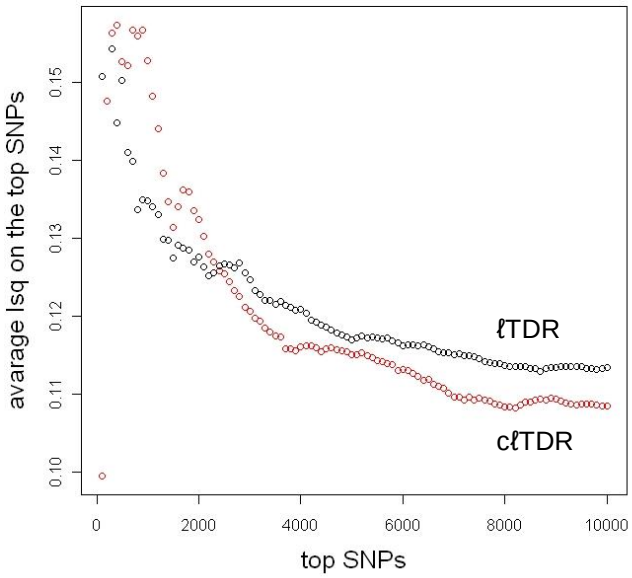


Figure s2 Heterogeneity of SNP effects before and after data integration. The y-axis shows the average I^2 index and the x-axis the top SNPs ranked in ascending order using either the ℓTDR (red) or cℓTDR (black).

Results in **Figure S2** show the average I^2 of the top SNPs, with the number of SNPs indicated on the x-axis, selected before and after data integration. As the number of SNPs used to calculate the average I^2 is small at the left hand side of the figure, the results fluctuate initially. However, as the number of SNPs increases, the average I^2 values becomes better for the cℓTDR compared to the ℓTDR, implying that data integration results in the selection of SNPs with more consistent effects across the 18 GWAS studies.

GO analyses: We performed gene ontology (GO) analyses to establish whether genes selected through data integration show differences in terms of GO themes. For these analyses we only used empirical data sets (linkage analyses, expression array) to avoid that enrichment was introduced by using external data sets that are based on biological knowledge (e.g. candidate gene studies). We first selected SNPs based on having a good (n=1,435) with a larger posterior probability of belonging to the group of SNPs with small effects than the other two groups. Then we identified the genes in which these SNPs were located and subjected those SNPs to a GO analysis to search for biological themes. A similar number of genes were selected using the top ℓTDR results. As an additional control groups, we also selected a similar number of genes from the bottom (i.e. highest values) of the cℓTDR and ℓTDR distributions.

For the GO analyses we used GOEAST[2], a web based software toolkit. The ontology covers three domains: *cellular component*, the parts of a cell or its extracellular environment; *molecular function*, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and *biological process*, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms. GOEAST uses an exact

(hypergeometric) test to evaluate the null hypothesis that genes are picked at random from the total gene population.

Figure S3 shows the results for the top (blue line) and bottom (green line) genes selected before (**Fig. a**) and after (**Fig. b**) data integration. The y-axis shows the *p*-values and the X-axis shows the number of genes with *p*-values smaller than 0.1 and then sorted in ascending order. Compared to the control group of genes selected from the bottom of the cℓTDR and ℓTDR distributions, there are many more genes with *p*-values smaller

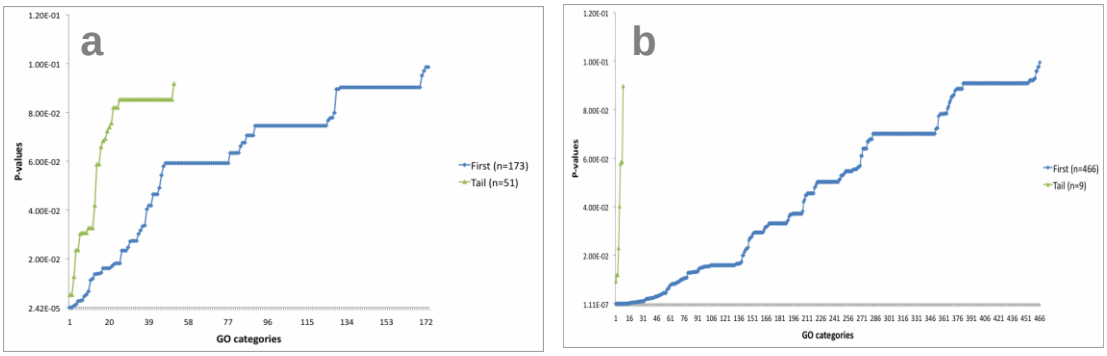


Figure S3. GO analysis on the top (blue line) and bottom (green line) genes selected before (Fig a) and after (Fig b) data integration. The y-axis shows the *p*-values testing the null hypothesis that the selected genes were picked at random from the total gene population and the X-axis shows the number of genes with *p*-values smaller than 0.1 and then sorted in ascending order.

than 0.1 and the tests also indication much smaller *p*-values. Furthermore, this pattern is much more pronounced after the data integration. Thus, our top results differ from the bottom results in terms of gene ontology and data integration tends to increase the distinction between top and bottom genes.

1. Higgs BW, Elashoff M, Richman S, Barci B. An online database for brain disease research. *BMC Genomics.* 2006;7:70.

2. Zheng Q, Wang XJ. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.* Jul 1 2008;36(Web Server issue):W358-363.

3. Cochran WG. The combination of estimates from different experiments. *Biometrics.* 1954;10:101-129.

4. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557-560.

Appendix for "A rigorous method for integrating multiple heterogeneous databases in large scale genetic studies"

1.1 Mathematical formulas for the exact computation of cℓTDR

Step 1: Obtaining prior probabilities of test units for each EDS

The goal of this section is to derive the equation of prior probability (Theorem 3). Throughout this section we assume that we have a single existing data set. For brevity we will use the notation γ_r instead of $\gamma(r)$ for the probability that a test unit ranked r in the existing data set is alternative in the existing data set.

Theorem 1 *Suppose we have only one existing data set. Denote the number of test units of NDC that are also in the existing data set as m. Out of these m test units, denote the number of those alternative in the NDC, the existing data set and in both data sets as m_1^*, m_1, and $m_1^{overlap}$, respectively, and let $m_0 = m - m_1$. Then we have that*

$$\gamma^*(i) = \pi\gamma(r_i) + \nu(1-\gamma(r_i)) = \frac{m_1^{overlap}}{m_1}\gamma(r_i) + \frac{m_1^* - m_1^{overlap}}{m_0}(1-\gamma(r_i)), \tag{1}$$

where r_i is the rank of test unit i in the existing data set and

$$\pi = \Pr\left(H_i^{(NDC)} = 1 \mid H_i^{(EDS)} = 1\right) = \frac{m_1^{overlap}}{m_1}$$
$$\nu = \Pr\left(H_i^{(NDC)} = 1 \mid H_i^{(EDS)} = 0\right) = \frac{m_1^* - m_1^{overlap}}{m_0},$$

where $H_i^{(NDC)} = 1$ or 0 if test unit i is alternative or null in the NDC, respectively, and $H_i^{(EDS)} = 1$ or 0 if test unit i is alternative or null in the existing data set, respectively.

Proof. By definition

$$\gamma^*(i) \overset{def}{=} \Pr\left(H_i^{(\mathrm{NDC})} = 1 \mid S = s\right),$$

where $S = s$ represents our information from the existing data set. Applying the well-known identity $\Pr(B \mid C) = \sum_i \Pr(B \mid A_i, C)\Pr(A_i \mid C)$,where $\{A_i\}$ is a partition of the probability space, we obtain that

$$\gamma^*(i) = \Pr\left(H_i^{(\mathrm{NDC})} = 1 \mid H_i^{(\mathrm{EDS})} = 1, S = s\right)\Pr\left(H_i^{(\mathrm{EDS})} = 1 \mid S = s\right) +$$
$$\Pr\left(H_i^{(\mathrm{NDC})} = 1 \mid H_i^{(\mathrm{EDS})} = 0, S = s\right)\Pr\left(H_i^{(\mathrm{EDS})} = 0 \mid S = s\right) =$$
$$\Pr\left(H_i^{(\mathrm{NDC})} = 1 \mid H_i^{(\mathrm{EDS})} = 1\right)\Pr\left(H_i^{(\mathrm{EDS})} = 1 \mid S = s\right) + \Pr\left(H_i^{(\mathrm{NDC})} = 1 \mid H_i^{(\mathrm{EDS})} = 0\right)\Pr\left(H_i^{(\mathrm{EDS})} = 0 \mid S = s\right) =$$
$$\pi\gamma(r_i) + \nu(1-\gamma(r_i)).$$

■

Lemma 2 *For test unit i in the existing data set we have that*

$$\gamma^*(i) = co(i) + \frac{m_1^*}{m},$$

where

$$co\left(i\right)=\frac{1}{m_0}\left(m\gamma\left(r_i\right)-m_1\right)\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right),$$

and r_i is the rank of test unit i in the existing data set, m is the number of test units of NDC that are also in the existing data out of which m_1^*, m_1, and $m_1^{overlap}$, is the number of test units alternative in the NDC, the existing data set and in both data sets, respectively.

Proof. For brevity we will denote r_i as j, hence $\gamma\left(r_i\right)\equiv\gamma\left(j\right)\equiv\gamma_j$. From (1) we have that

$$\gamma^*\left(i\right)=\pi\gamma\left(r_i\right)+\nu\left(1-\gamma\left(r_i\right)\right)=\pi\gamma_j+\nu\left(1-\gamma_j\right)=$$

$$\frac{m_1^{overlap}}{m_1}\gamma_j+\frac{m_1^*-m_1^{overlap}}{m_0}\left(1-\gamma_j\right)=$$

$$\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*-m_1^{overlap}}{m_0}\right)\gamma_j+\frac{m_1^*-m_1^{overlap}}{m_0}=$$

$$\frac{1}{m_0}\left[\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\gamma_j+m_1^*-m_1^{overlap}\right]=$$

$$\frac{1}{m_0}\left[\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\left(\gamma_j-\frac{m_1}{m}\right)+\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\frac{m_1}{m}+m_1^*-m_1^{overlap}\right]=$$

$$\frac{1}{m_0}\left[\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\left(\gamma_j-\frac{m_1}{m}\right)+m_1^{overlap}-m_1^*\frac{m_1}{m}+m_1^*-m_1^{overlap}\right]=$$

$$\frac{1}{m_0}\left[\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\left(\gamma_j-\frac{m_1}{m}\right)+m_1^*\frac{m_0}{m}\right]=$$

$$\frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1}m-m_1^*\right)\left(\gamma_j-\frac{m_1}{m}\right)+\frac{m_1^*}{m}=$$

$$\frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right)\left(m\gamma_j-m_1\right)+\frac{m_1^*}{m}=co\left(i\right)+\frac{m_1^*}{m}.$$

■

Theorem 3 *For test unit i in the NDC we have that*

$$\gamma^*(i)=\begin{cases}co\left(i\right)+\frac{m_1^*}{m} & \text{if test unit } i \text{ is in the existing data set}\\ \frac{m_1^{(NDC)}-m_1^*}{m^{(NDC)}-m} & \text{if test unit } i \text{ is NOT in the existing data set,}\end{cases}$$

where m is the number of test units of NDC that are also in the existing data set, out of which m_1^*, m_1, and $m_1^{overlap}$ is the number of test units alternative in the NDC, the existing data set and in both data sets, respectively, $m^{(NDC)}$ and $m_1^{(NDC)}$ is the number of test units and the number of alternative test units in the NDC.

Proof. The statement of the theorem directly follows from Lemma 2 for test units in the existing data set. The number of test units in the NDC and not in the existing data set is $m^{(\mathrm{NDC})}-m$, $m_1^{(\mathrm{NDC})}-m_1^*$ of which is alternative. Consequently, $\gamma^*(i)=\frac{m_1^{(\mathrm{NDC})}-m_1^*}{m^{(\mathrm{NDC})}-m}$ for test unit i that is not in the existing data set, because we have no prior information for this test unit from the existing data set. ■

Corollary 4 *Under the reasonable assumption that the concentration of the alternative test units is the same inside and outside the region covered by the existing data set, i.e.* $\frac{m_1^*}{m} = \frac{m_1^{(NDC)}}{m^{(NDC)}}$, *we have that*

$$\gamma^*(i) = \begin{cases} co\,(i) + \frac{m_1^{(NDC)}}{m^{(NDC)}} & \text{if test unit } i \text{ is in the existing data set} \\ \frac{m_1^{(NDC)}}{m^{(NDC)}} & \text{if test unit } i \text{ is NOT in the existing data set,} \end{cases}$$

and $co\,(i)$ *can be calculated as*

$$co\,(i) = \frac{1}{m_0}\left(m\gamma\left(r_i\right) - m_1\right)\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^{(NDC)}}{m^{(NDC)}}\right).$$

Proof. The statement of the corollary follows from Theorem 3 and from that $\frac{m_1^*}{m} = \frac{m_1^{(\mathrm{NDC})}}{m^{(\mathrm{NDC})}}$ implies $\frac{m_1^{(\mathrm{NDC})} - m_1^*}{m^{(\mathrm{NDC})} - m} = \frac{m_1^*}{m} = \frac{m_1^{(\mathrm{NDC})}}{m^{(\mathrm{NDC})}}$. ■

Some properties of rank-based probability, γ, prior probability, γ^*, and the contribution First we derive formulas for rank-based probabilities that we will use to obtain formulas for prior probabilities and contribution.

Theorem 5 *Suppose that* $X_1, ..., X_{m_0}$ *are identically (not necessarily independently) distributed random variables, representing the true null statistics, and suppose that* $Y_1, ..., Y_{m_1}$ *are identically (not necessarily independently) distributed random variables, representing the true alternative statistics. Denote the kth largest random variable from* $\{X_1, ..., X_{m_0}, Y_1, ..., Y_{m_1}\}$ *as* Z_k. *Then for any fixed* i, *the probability that* Y_i *is the kth largest test statistic value,* Z_k, *is*

$$\Pr\left(Y_i = Z_k\right) = \frac{\gamma_k}{m_1}$$

and the probability that X_i *is the kth largest test statistic value is*

$$\Pr\left(X_i = Z_k\right) = \frac{1 - \gamma_k}{m_0},$$

where γ_k *is the probability that* Z_k *is alternative.*

Proof.

$$\gamma_k = \Pr\left(Z_k \text{ is alternative}\right) = \textstyle\sum_{j=1}^{m_1} \Pr\left(Y_j = Z_k\right) = m_1 \Pr\left(Y_i = Z_k\right),$$

from which the first statement follows. The second statement can be proven similarly. ■

Corollary 6 *As a consequence of the above theorem we have that*

$$\textstyle\sum_{k=1}^{m} \gamma_k = m_1, \tag{2}$$

where $m = m_1 + m_0$.

Proof. For any fixed i, the events $\{Y_i = Z_k\}_{k=1}^m$ is a partition of the the probability space, we have that

$$\textstyle\sum_{k=1}^{m} \Pr\left(Y_i = Z_k\right) = \sum_{k=1}^{m} \frac{\gamma_k}{m_1} = 1,$$

which implies the statement of the corollary. ■

Remark 7 *Note that the analogous statement in Corollary 6 for parametric calculation is not valid. That is if the parametric formula*

$$\gamma_j^{par} = \Pr\left(H_j = 1 \mid T = t_j\right) = \frac{m_1 f_\Psi\left(t_j\right)}{m_0 f_0\left(t_j\right) + m_1 f_\Psi\left(t_j\right)}. \tag{3}$$

is used to calculate the probability that a test unit is alternative in the existing data set, where f_0 and f_Ψ are the null and the alternative p.d.f., then $\sum_{k=1}^{m}\gamma_k^{par}$ may not be equal with m_1. However, it is easy to see that

$$\sum_{k=1}^{m} E\left(\gamma_k^{par}\right) = m_1.$$

Lemma 8 *For any existing data set we have that*

$$\sum_{i=1}^{m} co\left(i\right) = 0,$$

where m is the number of test units in the existing data set.

Proof.

$$\sum_{i=1}^{m} co\left(i\right) = \sum_{i=1}^{m} \frac{1}{m_0}\left(m\gamma\left(r_i\right) - m_1\right)\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^*}{m}\right) = \frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^*}{m}\right)\sum_{i=1}^{m}\left(m\gamma\left(r_i\right) - m_1\right) =$$

$$\frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^*}{m}\right)\left(m\sum_{j=1}^{m}\gamma\left(j\right) - m_1\right) = \frac{m}{m_0}\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^*}{m}\right)\left(m_1 - m_1\right) = 0,$$

where we used (2). ■

Corollary 9 *If we have a single existing data set, then*

$$\sum_{i=1}^{m^{(NDC)}} \gamma^*(i) = m_1^{(NDC)},$$

where $m^{(NDC)}$ and $m_1^{(NDC)}$ is the number of test units and the number of alternative test units in the NDC, respectively. That is the sum of the prior probabilities is the same as the one without prior information.

Proof. We have that

$$\sum_{i=1}^{m^{(\mathrm{NDC})}} \gamma^*(i) = \sum_{\substack{i \text{ is a test unit in} \\ \text{the existing data set}}} \gamma^*(i) + \sum_{\substack{i \text{ is not a test unit in} \\ \text{the existing data set}}} \gamma^*(i) = \sum_{i=1}^{m}\left(co\left(i\right) + \frac{m_1^*}{m}\right) + \sum_{i=1}^{m^{(\mathrm{NDC})}-m}\left(\frac{m_1^{(\mathrm{NDC})} - m_1^*}{m^{(\mathrm{NDC})} - m}\right) =$$

$$\left\{\sum_{i=1}^{m} co\left(i\right) = 0\right\} = m_1^* + \left(m_1^{(\mathrm{NDC})} - m_1^*\right) = m_1^{(\mathrm{NDC})}.$$

■

Claim 10 *We have that*

$$\frac{mm_1}{m_0}\kappa \le m_1^{overlap}, \tag{4}$$

and equality holds if and only if $m_1^{overlap} = m_1^*$.

Proof.

$$\frac{mm_1}{m_0}\kappa = \frac{mm_1}{m_0}\left(\frac{m_1^{overlap}}{m_1} - \frac{m_1^*}{m}\right) = \frac{m}{m_0}m_1^{overlap} - \frac{m_1}{m_0}m_1^* = \frac{m}{m_0}m_1^{overlap} - \frac{m_1}{m_0}m_1^{overlap} + \frac{m_1}{m_0}m_1^{overlap} - \frac{m_1}{m_0}m_1^* =$$

$$\left(\frac{m}{m_0} - \frac{m_1}{m_0}\right)m_1^{overlap} - \frac{m_1}{m_0}\left(m_1^* - m_1^{overlap}\right) = m_1^{overlap} - \frac{m_1}{m_0}\left(m_1^* - m_1^{overlap}\right).$$

Clearly,

$$m_1^{overlap} - \frac{m_1}{m_0}\left(m_1^* - m_1^{overlap}\right) \le m_1^{overlap},$$

and equality holds if and only if $m_1^{overlap} = m_1^*$. ■

Existing data sets with ties Suppose we have t groups of test units and the ranks of all test units in a group are identical. Let R_i be the number of test units whose rank is the ith largest one or smaller for $i = 1, ..., t$, and let $R_0 = 0$. Clearly, $R_i - R_{i-1}$ is the number of test units in the ith group for all $i = 1, ..., t$. Denote the probability that a test unit in the ith group is alternative in the existing data set as Γ_i.

Claim 11 *Γ_i can be calculated by*

$$\Gamma_i = \Pr\left(Z_r \text{ is alternative} \mid R_{i-1} < r \leq R_i\right) =$$

$$\frac{1}{R_i - R_{i-1}} \sum\nolimits_{r=R_{i-1}+1}^{R_i} \gamma(r) = \frac{1}{R_i - R_{i-1}} \left\{R_i \Pr\left(Z_r \text{ is alternative} \mid r \leq R_i\right) - R_{i-1} \Pr\left(Z_r \text{ is alternative} \mid r \leq R_{i-1}\right)\right\},$$

where Z_r and $\gamma(r)$ are defined above.

Corollary 12 *Denote the probability that a test unit in the ith group is alternative in the NDC as Γ_i^*. Then we have that*

$$\Gamma_i^* = \frac{m_1^*}{m} + \frac{1}{R_i - R_{i-1}} \sum\nolimits_{r_j=R_{i-1}+1}^{R_i} co\left(j\right) = \frac{m_1^*}{m} + \frac{1}{R_i - R_{i-1}} \left(CO\left(R_i\right) - CO\left(R_{i-1}\right)\right), \tag{5}$$

where we define $CO\left(R\right) = \sum_{r_j=1}^{R} co\left(j\right)$.

Proof. We prove first that

$$\frac{1}{R_i - R_{i-1}} \sum\nolimits_{j=R_{i-1}+1}^{R_i} \gamma_j^* = \frac{1}{R_i - R_{i-1}} \sum\nolimits_{r_j=R_{i-1}+1}^{R_i} \frac{m_1^{overlap}}{m_1} \gamma\left(r_j\right) + \frac{m_1^* - m_1^{overlap}}{m_0} \left(1 - \gamma\left(r_j\right)\right) =$$

$$\frac{m_1^{overlap}}{m_1} \Gamma_i + \frac{m_1^* - m_1^{overlap}}{m_0} \left(1 - \Gamma_i\right) = \Gamma_i^*,$$

where the last step can be proved analogously to (1). Then it readily follows from Lemma 2 that

$$\Gamma_i^* = \frac{1}{R_i - R_{i-1}} \sum\nolimits_{r_j=R_{i-1}+1}^{R_i} \gamma_j^* = \frac{m_1^*}{m} + \frac{1}{R_i - R_{i-1}} \sum\nolimits_{r_j=R_{i-1}+1}^{R_i} co\left(j\right).$$

■

The statement analogous to the one in Corollary 6 is also valid:

Claim 13 *The sum of the prior probabilities will be m_1, i.e.*

$$\sum\nolimits_{i=1}^{t} \left(R_i - R_{i-1}\right) \Gamma_i = m_1.$$

Proof.

$$\sum\nolimits_{i=1}^{t} \left(R_i - R_{i-1}\right) \Gamma_i = \sum\nolimits_{i=1}^{t} \left(R_i - R_{i-1}\right) \Pr\left(Z_r \text{ is alternative} \mid R_{i-1} < r \leq R_i\right) =$$

$$\sum\nolimits_{i=1}^{t} \left\{R_i \Pr\left(Z_r \text{ is alternative} \mid r \leq R_i\right) - R_{i-1} \Pr\left(Z_r \text{ is alternative} \mid r \leq R_{i-1}\right)\right\} =$$

$$\left\{R_0 = 0\right\} = R_t \Pr\left(Z_r \text{ is alternative} \mid r \leq R_t\right) = \left\{R_t = m\right\} = m \Pr\left(Z_r \text{ is alternative} \mid r \text{ is anything}\right) = m \frac{m_1}{m} = m_1.$$

■

Claim 14 *Suppose that $X_1, ..., X_{m_0}$ are i.i.d. random variables with F_0 and $Y_1, ..., Y_{m_1}$ are i.i.d. random variables with F_Ψ. Suppose that $X_1, ..., X_{m_0}, Y_1, ..., Y_{m_1}$ are independent, and denote the rth largest random variable among them as Z_r. Then the probability that Y_s is in the ith (largest) group of test units is*

$$\Pr\left(Y_s \text{ is in the } i\text{th group}\right) = \frac{\Gamma_i}{m_1}$$

and the probability that X_s is in the ith (largest) group of test units is

$$\Pr\left(X_s \text{ is in the } i\text{th group}\right) = \frac{1-\Gamma_i}{m_0},$$

where Γ_i is the probability that Z_k is alternative given it is in the ith group.

Proof. The probability that Y_s is in the ith (largest) group of test units is

$$\Pr\left(Y_s \text{ is in the } i\text{th group}\right) = \frac{1}{R_i - R_{i-1}} \sum\nolimits_{j=R_{i-1}+1}^{R_i} \frac{\gamma_j}{m_1} = \frac{\Gamma_i}{m_1}$$

and the probability that X_s is in the ith (largest) group of test units is

$$\Pr\left(X_s \text{ is in the } i\text{th group}\right) = \frac{1}{R_i - R_{i-1}} \sum\nolimits_{j=R_{i-1}+1}^{R_i} \frac{1-\gamma_j}{m_0} = \frac{1}{m_0} - \frac{1}{R_i - R_{i-1}} \sum\nolimits_{j=R_{i-1}+1}^{R_i} \frac{\gamma_j}{m_0} = \frac{1}{m_0} - \frac{\Gamma_i}{m_0} = \frac{1-\Gamma_i}{m_0}.$$

■

Remark 15 *Note that the distribution in Theorem (5) and that in Claim 14) is formally the same. As a result, for multiple instances we do not need to distinguish the case ties from that of no ties.*

Step 2: Combining the sets of prior probabilities into a single set of prior probabilities

Definition 16 *Suppose we have k existing data sets with (ranks of) test statistic values $S_1^i, ..., S_m^i$. The combined prior probability that a test unit is alternative in the novel data collection based on the information in the existing data sets is defined as*

$$\gamma_j^{(combined\ prior)} \overset{def}{=} \Pr\left(H_j^{(NDC)} = 1 \mid S_j^i = s_j^i, i = 1, ..., k\right).$$

Theorem 17 *Suppose we have k existing data sets, and $S_1^i, ..., S_m^i$ are the test statistic values or the ranks of the test statistic values in the ith existing data set, $i = 1, ..., k$, where some S_j^i may be missing. Denote the number of alternative test units in the ith existing data set as m_1^i. Then we have that*

$$\gamma_j^{(combined\ prior)} = \frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i\right)\right] \Pr\left(H_j^{(NDC)} = 1, H_j^{(i)} = \delta_i, i = 1, ..., k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i\right)\right] \Pr\left(H_j^{(i)} = \tau_i, i = 1, ..., k\right)} \tag{6}$$

where m_1 and m_0 is the number of alternative and null test units, respectively, in the novel data collection. Notation $\{0,1\}^k$ means all the 0-1 vectors of length k. Moreover,

$$\beta_j^{(combined\ prior)} = \frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i\right)\right] \Pr\left(H_j^{(NDC)} = 1, H_j^{(i)} = \delta_i, i = 1, ..., k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i\right)\right] \Pr\left(H_j^{(NDC)} = 0, H_j^{(i)} = \tau_i, i = 1, ..., k\right)}, \tag{7}$$

where $\beta_j^{(combined\ prior)} := \gamma_j^{(combined\ prior)} / \left(1 - \gamma_j^{(combined\ prior)}\right)$.

Proof. We have that

$$\gamma_j^{(\text{combined prior})} \overset{def}{=} \Pr\left(H_j^{(\text{NDC})}=1 \mid S_j^i=s_j^i, i=1,...,k\right) \overset{*}{=}$$

$$\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \Pr\left(H_j^{(\text{NDC})}=1 \mid H_j^{(i)}=\delta_i, i=1,...,k\right) \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid S_j^i=s_j^i, i=1,...,k\right) =$$

$$\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \left\{ \Pr\left(H_j^{(\text{NDC})}=1 \mid H_j^{(i)}=\delta_i, i=1,...,k\right) \right.$$

$$\left. \frac{\Pr\left(S_j^i=s_j^i, i=1,...,k \mid H_j^{(i)}=\delta_i, i=1,...,k\right) \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k} \Pr\left(S_j^i=s_j^i, i=1,...,k \mid H_j^{(i)}=\tau_i, i=1,...,k\right) \Pr\left(H_j^{(i)}=\tau_i, i=1,...,k\right)} \right\} =$$

$$\left\{ \Pr\left(H_j^{(\text{NDC})}=1 \mid H_j^{(i)}=\delta_i, i=1,...,k\right) \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k\right) = \Pr\left(H_j^{(\text{NDC})}=1,\ H_j^{(i)}=\delta_i, i=1,...,k\right) \right\} =$$

$$\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \frac{\Pr\left(S_j^i=s_j^i, i=1,...,k \mid H_j^{(i)}=\delta_i, i=1,...,k\right) \Pr\left(H_j^{(\text{NDC})}=1,\ H_j^{(i)}=\delta_i, i=1,...,k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k} \Pr\left(S_j^i=s_j^i, i=1,...,k \mid H_j^{(i)}=\tau_i, i=1,...,k\right) \Pr\left(H_j^{(i)}=\tau_i, i=1,...,k\right)} =$$

$$\frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i=s_j^i \mid H_j^{(i)}=\delta_i\right)\right] \Pr\left(H_j^{(\text{NDC})}=1,\ H_j^{(i)}=\delta_i, i=1,...,k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k} \left[\prod_{i=1}^k \Pr\left(S_j^i=s_j^i \mid H_j^{(i)}=\tau_i\right)\right] \Pr\left(H_j^{(i)}=\tau_i, i=1,...,k\right)}.$$

At $*$ we used that

$$\Pr(B \mid C) = \textstyle\sum_i \Pr(B \mid A_i, C) \Pr(A_i \mid C)$$

if A_i, $i=1,2...$ is a partition of the probability space. Moreover, we used the reasonable assumption that $\Pr\left(H_j^{(\text{NDC})}=1 \mid H_j^{(i)}=\delta_i, S_j^i=s_j^i, i=1,...,k\right) = \Pr\left(H_j^{(\text{NDC})}=1 \mid H_j^{(i)}=\delta_i, i=1,...,k\right)$. ■

As it is unknown in practice how the sets of alternative test units in the existing data sets and novel data collection overlap, the probabilities $\Pr\left(H_j^{(i)}=\tau_i, i=1,...,k\right)$ and $\Pr\left(H_j^{(\text{NDC})}=1,\ H_j^{(i)}=\delta_i, i=1,...,k\right)$ in (6) are unknown. Therefore, we need to use some mild assumptions to provide some practically useful and sufficiently accurate methods to combine prior probabilities from several existing data sets.

Theorem 18 *Suppose we have k existing data sets. Suppose that*

$$\begin{aligned} \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(NDC)}=1\right) &= \textstyle\prod_{i=1}^k \Pr\left(H_j^{(i)}=\delta_i \mid H_j^{(NDC)}=1\right) a^{\sum_{t=1}^k (1-\delta_t)} \\ \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(NDC)}=0\right) &= \textstyle\prod_{i=1}^k \Pr\left(H_j^{(i)}=\delta_i \mid H_j^{(NDC)}=0\right) b^{\sum_{t=1}^k \delta_t}, \end{aligned} \tag{8}$$

hold for every $(\delta_1,...,\delta_k)\in\{0,1\}^k$, where $0\le a,b\le 1$ and we define $0^0=1$. Then

$$\beta_j^{(combined\ prior)} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)} a}{\gamma_j^{(i)}\left(1-\pi^{(i)}\right) b + \left(1-\gamma_j^{(i)}\right)\left(1-\nu^{(i)}\right)} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma^{*i}(j) - (1-a)\left(1-\gamma_j^{(i)}\right)\nu^{(i)}}{1-\gamma^{*i}(j) - \gamma_j^{(i)}(1-b)\left(1-\pi^{(i)}\right)}, \tag{9}$$

where m_1 is the number of alternative test units in the novel data collection, $m_0 = m - m_1$,

$$\pi^{(i)} = \Pr\left(H_j^{(NDC)} = 1 \mid H_j^{(i)} = 1\right)$$
$$\nu^{(i)} = \Pr\left(H_j^{(NDC)} = 1 \mid H_j^{(i)} = 0\right)$$

for $i = 1, ..., k$,

$$\gamma^{*i}(j) = \pi^{(i)}\gamma_j^{(i)} + \nu^{(i)}\left(1 - \gamma_j^{(i)}\right)$$

is the prior probability that test unit j is alternative in the NDC based on the ith EDS (see (1)), and $\gamma_j^{(i)}$ is the probability that test unit j is alternative the ith EDS, i.e.

$$\gamma_j^{(i)} = \Pr\left(H_j^{(i)} = 1 \mid S_j^i = s_j^i\right) = \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)},$$

where $S_1^i, ..., S_m^i$ are the test statistic values or the ranks of the test statistic values in the ith existing data set, $i = 1, ..., k$.

Proof. Applying criterion in (8) for the formula in (7) we obtain that

$$\beta_j^{(\text{combined prior})} = \frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i\right)\right]\Pr\left(H_j^{(\text{NDC})} = 1, H_j^{(i)} = \delta_i, i = 1, ..., k\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i\right)\right]\Pr\left(H_j^{(\text{NDC})} = 0, H_j^{(i)} = \tau_i, i = 1, ..., k\right)} =$$

$$\frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i\right)\right]\Pr\left(H_j^{(i)} = \delta_i, i = 1, ..., k \mid H_j^{(\text{NDC})} = 1\right)\Pr\left(H_j^{(\text{NDC})} = 1\right)}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i\right)\right]\Pr\left(H_j^{(i)} = \tau_i, i = 1, ..., k \mid H_j^{(\text{NDC})} = 0\right)\Pr\left(H_j^{(\text{NDC})} = 0\right)} =$$

$$\frac{\Pr\left(H_j^{(\text{NDC})} = 1\right)}{\Pr\left(H_j^{(\text{NDC})} = 0\right)}\frac{\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i\right)\right]\prod_{i=1}^k \Pr\left(H_j^{(i)} = \delta_i \mid H_j^{(\text{NDC})} = 1\right)a^{\sum_{t=1}^k(1-\delta_t)}}{\sum_{(\tau_1,...,\tau_k)\in\{0,1\}^k}\left[\prod_{i=1}^k \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i\right)\right]\prod_{i=1}^k \Pr\left(H_j^{(i)} = \tau_i \mid H_j^{(\text{NDC})} = 0\right)b^{\sum_{t=1}^k\tau_t}} =$$

$$\frac{\Pr\left(H_j^{(\text{NDC})} = 1\right)}{\Pr\left(H_j^{(\text{NDC})} = 0\right)}$$

$$\prod_{i=1}^k \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1 \mid H_j^{(\text{NDC})} = 1\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0 \mid H_j^{(\text{NDC})} = 1\right)a}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1 \mid H_j^{(\text{NDC})} = 0\right)b + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0 \mid H_j^{(\text{NDC})} = 0\right)} =$$

$$\left[\frac{\Pr\left(H_j^{(\text{NDC})} = 0\right)}{\Pr\left(H_j^{(\text{NDC})} = 1\right)}\right]^{k-1}$$

$$\prod_{i=1}^k \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1, H_j^{(\text{NDC})} = 1\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0, H_j^{(\text{NDC})} = 1\right)a}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1, H_j^{(\text{NDC})} = 0\right)b + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0, H_j^{(\text{NDC})} = 0\right)} =$$

$$\left[\frac{\Pr\left(H_j^{(\text{NDC})} = 0\right)}{\Pr\left(H_j^{(\text{NDC})} = 1\right)}\right]^{k-1}\prod_{i=1}^k \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)\Pr\left(H_j^{(\text{NDC})} = 1 \mid H_j^{(i)} = 1\right) +}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)\Pr\left(H_j^{(\text{NDC})} = 0 \mid H_j^{(i)} = 1\right)b +}$$

$$\frac{+\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right)\Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid H_j^{(i)} = 0\right)a}{+\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right)\Pr\left(H_j^{(\mathrm{NDC})} = 0 \mid H_j^{(i)} = 0\right)} =$$

$$\left[\frac{\Pr\left(H_j^{(\mathrm{NDC})} = 0\right)}{\Pr\left(H_j^{(\mathrm{NDC})} = 1\right)}\right]^{k-1}$$

$$\prod_{i=1}^{k}\frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)\pi^{(i)} + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right)\nu^{(i)}a}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)\left(1-\pi^{(i)}\right)b + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right)\left(1-\nu^{(i)}\right)} =$$

$$\left\{\gamma_j^{(i)} = \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)}\right\} =$$

$$\left[\frac{\Pr\left(H_j^{(\mathrm{NDC})} = 0\right)}{\Pr\left(H_j^{(\mathrm{NDC})} = 1\right)}\right]^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{\gamma_j^{(i)}\left(1-\pi^{(i)}\right)b + \left(1-\gamma_j^{(i)}\right)\left(1-\nu^{(i)}\right)} = \left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{\gamma_j^{(i)}\left(1-\pi^{(i)}\right)b + \left(1-\gamma_j^{(i)}\right)\left(1-\nu^{(i)}\right)},$$

which proves the first equality in (9). Moreover, we have that

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{\gamma_j^{(i)}\left(1-\pi^{(i)}\right)b + \left(1-\gamma_j^{(i)}\right)\left(1-\nu^{(i)}\right)} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{1-\left\{\gamma_j^{(i)} - \gamma_j^{(i)}b + \left(\gamma_j^{(i)}\pi^{(i)}b + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}\right)\right\}} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{1-\left\{\gamma_j^{(i)}\left(1-b+\pi^{(i)}b\right) + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}\right\}} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)} - (1-a)\left(1-\gamma_j^{(i)}\right)\nu^{(i)}}{1-\left\{\gamma_j^{(i)}\left(1-b+\pi^{(i)}b-\pi^{(i)}\right) + \gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}\right\}} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)} - (1-a)\left(1-\gamma_j^{(i)}\right)\nu^{(i)}}{1-\left\{\gamma_j^{(i)}(1-b)\left(1-\pi^{(i)}\right) + \gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}\right\}} =$$

$$\left\{\gamma^{*i}(j) = \pi^{(i)}\gamma_j^{(i)} + \nu^{(i)}\left(1-\gamma_j^{(i)}\right)\right\} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma^{*i}(j) - (1-a)\left(1-\gamma_j^{(i)}\right)\nu^{(i)}}{1-\left\{\gamma_j^{(i)}(1-b)\left(1-\pi^{(i)}\right) + \gamma^{*i}(j)\right\}} =$$

$$\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma^{*i}(j) - (1-a)\left(1-\gamma_j^{(i)}\right)\nu^{(i)}}{1-\gamma^{*i}(j) - \gamma_j^{(i)}(1-b)\left(1-\pi^{(i)}\right)},$$

which proves the second equality in (9). ■

Corollary 19 *Suppose we have k existing data sets. Suppose that*

$$\begin{array}{c} \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(NDC)}=1\right)=\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=\delta_i \mid H_j^{(NDC)}=1\right) \\ \Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(NDC)}=0\right)=\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=\delta_i \mid H_j^{(NDC)}=0\right), \end{array} \tag{10}$$

for every $(\delta_1,...,\delta_k)\in\{0,1\}^k$, then

$$\beta_j^{(combined\ prior)}=\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma^{*i}(j)}{1-\gamma^{*i}(j)}=\frac{m_1}{m_0}\prod_{i=1}^{k}\frac{m_0\gamma^{*i}(j)}{m_1\left(1-\gamma^{*i}(j)\right)}, \tag{11}$$

where m_1 is the number of alternative test units in the novel data collection, $m_0=m-m_1$, and $\gamma^{*i}(j)$ is the prior probability that test unit j is alternative in the NDC based on the ith EDS.

Proof. The condition in (10) is equivalent to (8) with $a=b=1$. Therefore, substituting $a=b=1$ in (9) we obtain that

$$\beta_j^{(\text{combined prior})}=\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma^{*i}(j)}{1-\gamma^{*i}(j)}=\frac{m_1}{m_0}\prod_{i=1}^{k}\frac{m_0\gamma^{*i}(j)}{m_1\left(1-\gamma^{*i}(j)\right)}.$$

■

Corollary 20 *If*

$$H_j^{(NDC)}=1\Leftrightarrow H_j^{(1)}=1\Leftrightarrow ...\Leftrightarrow H_j^{(k)}=1, \tag{12}$$

then

$$\beta_j^{(combined\ prior)}=\left(\frac{m_0}{m_1}\right)^{k-1}\prod_{i=1}^{k}\frac{\gamma^{*i}(j)}{1-\gamma^{*i}(j)}=\frac{m_1}{m_0}\prod_{i=1}^{k}\frac{m_0\gamma^{*i}(j)}{m_1\left(1-\gamma^{*i}(j)\right)},$$

where m_1 is the number of alternative test units in the novel data collection, $m_0=m-m_1$, and $\gamma^{*i}(j)$ is the prior probability that test unit j is alternative in the NDC based on the ith EDS.

Proof. First we need to see that the condition in (12) is equivalent to (8) for $a=b=0$. Indeed, substituting $a=b=0$ in (8) we obtain that

$$\Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(\text{NDC})}=1\right)=\begin{cases}\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=1 \mid H_j^{(\text{NDC})}=1\right) & \text{if } \delta_i=1 \text{ for every } i \\ 0 & \text{otherwise}\end{cases}$$

and

$$\Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(\text{NDC})}=0\right)=\begin{cases}\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=0 \mid H_j^{(\text{NDC})}=0\right) & \text{if } \delta_i=0 \text{ for every } i \\ 0 & \text{otherwise.}\end{cases}$$

As

$$1=\textstyle\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k}\Pr\left(H_j^{(i)}=\delta_i, i=1,...,k \mid H_j^{(\text{NDC})}=1\right)=\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=1 \mid H_j^{(\text{NDC})}=1\right),$$

we have that $\Pr\left(H_j^{(i)}=1 \mid H_j^{(\text{NDC})}=1\right)=1$ for every i, hence $H_j^{(\text{NDC})}=1\Longrightarrow H_j^{(i)}=1$ for every i. Similarly as

$$1=\textstyle\sum_{(\delta_1,...,\delta_k)\in\{0,1\}^k}\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=\delta_i \mid H_j^{(\text{NDC})}=0\right)=\prod_{i=1}^{k}\Pr\left(H_j^{(i)}=0 \mid H_j^{(\text{NDC})}=0\right),$$

we have that $\Pr\left(H_j^{(i)}=0 \mid H_j^{(\text{NDC})}=0\right)=1$ for every i, hence $H_j^{(\text{NDC})}=0\Longrightarrow H_j^{(i)}=0$ for every i. These two together implies $H_j^{(\text{NDC})}=1\Leftrightarrow H_j^{(i)}=1$ for every i.

Also the condition in (12) implies $\pi^{(i)} = \Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid H_j^{(i)} = 1\right) = 1$ and $\nu^{(i)} = \Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid H_j^{(i)} = 0\right) = 0$. Therefore, substituting $a = b = 0$, $\pi^{(i)} = 1$ and $\nu^{(i)} = 0$ in (9) we obtain that

$$\beta_j^{(\text{combined prior})} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^{k} \frac{\gamma_j^{(i)}\pi^{(i)} + \left(1-\gamma_j^{(i)}\right)\nu^{(i)}a}{\gamma_j^{(i)}\left(1-\pi^{(i)}\right)b + \left(1-\gamma_j^{(i)}\right)\left(1-\nu^{(i)}\right)} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^{k} \frac{\gamma_j^{(i)}}{\left(1-\gamma_j^{(i)}\right)} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^{k} \frac{\gamma^{*i}(j)}{1-\gamma^{*i}(j)},$$

where the last equation holds because $\gamma^{*i}(j) = \pi^{(i)}\gamma_j^{(i)} + \nu^{(i)}\left(1-\gamma_j^{(i)}\right) = \gamma_j^{(i)}$, as $\pi^{(i)} = 1$ and $\nu^{(i)} = 0$. This completes the proof of the Corollary. ■

Remark 21 *Note that the conditions in (10) and (12) represent the two extrema of (8), and in both cases the combined odds can be calculated as*

$$\beta_j^{(combined\ prior)} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^{k} \frac{\gamma^{*i}(j)}{1-\gamma^{*i}(j)} = \frac{m_1}{m_0} \prod_{i=1}^{k} \frac{m_0\gamma^{*i}(j)}{m_1\left(1-\gamma^{*i}(j)\right)} = \frac{m_1}{m_0} \prod_{i=1}^{k} \frac{m_0}{m_1}\beta^{*i}(j) \tag{13}$$

where m_1 is the number of alternative test units in the novel data collection, $m_0 = m - m_1$, $\gamma^{*i}(j)$ is the prior probability that test unit j is alternative in the NDC based on the ith EDS, and the odd $\beta^{*i}(j)$ is defined as $\beta^{*i}(j) = \gamma^{*i}(j) / \left(1-\gamma^{*i}(j)\right)$. Moreover, the terms in the product in (9) can be approximated with $\beta^{*i}(j)$ even if (10) and (12) do not hold, suggesting that the formula in (13) is reasonable even for the general case. For the general formula (6) the structure of how the sets of test units alternative in the EDSs as well as the set of test units alternative in the NDC overlap each other need to be known, which may be difficult to estimate.

Remark 22 *Recall that from (1) we have that*

$$\gamma^{*i}(j) = \pi^{(i)}\gamma_j^{(i)} + \nu^{(i)}\left(1-\gamma_j^{(i)}\right),$$

where

$$\pi^{(i)} = \Pr\left(H_j^{(NDC)} = 1 \mid H_j^{(i)} = 1\right)$$
$$\nu^{(i)} = \Pr\left(H_j^{(NDC)} = 1 \mid H_j^{(i)} = 0\right)$$

for $i = 1, ..., k$,(see (1)), and $\gamma_j^{(i)}$ is the probability that test unit j is alternative the ith EDS, i.e.

$$\gamma_j^{(i)} = \Pr\left(H_j^{(i)} = 1 \mid S_j^i = s_j^i\right) = \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right)\Pr\left(H_j^{(i)} = 0\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right)\Pr\left(H_j^{(i)} = 1\right)},$$

where $S_1^i, ..., S_m^i$ are the test statistic values or the ranks of the test statistic values in the ith existing data set, $i = 1, ..., k$.

Step 3: Computing cℓTDR for each test unit

The compound ℓTDR ($c\ell TDR$) of a test unit is defined as the posterior probability that the test unit is alternative in the novel data collection based on the information we have from the existing data sets and the novel data collection. The mathematical definition of the $c\ell TDR$ of a test unit is the following.

Definition 23 *The compound ℓTDR ($c\ell TDR$) of test unit j is defined as*

$$c\ell TDR(j) = \Pr\left(H_j^{(NDC)} = 1 \mid T = t_j, S_j^{(i)} = s_j^{(i)}, i = 1, ..., k\right),$$

where t_j is the observed test statistic value of test unit j in the novel data collection and $S_j^{(i)}$ is the test statistic value or the rank of the test statistic value of test unit j in the ith existing data set, $i = 1, ..., k$.

Claim 24 *The $c\ell TDR$ of a test unit can be calculated as*

$$c\ell TDR(j) = \frac{\gamma_j^{(combined\ prior)} f_1(t)}{f_0(t)\left(1 - \gamma_j^{(combined\ prior)}\right) + \gamma_j^{(combined\ prior)} f_1(t)} = \frac{\beta_j^{(combined\ prior)} f_1(t_j)}{f_0(t_j) + \beta_j^{(combined\ prior)} f_1(t_j)}, \tag{14}$$

where f_0 and f_1 is the null and alternative p.d.f. in the novel data collection, respectively, $\gamma_j^{(combined\ prior)}$ is the combined prior probability (from the existing data sets) that test unit j is alternative in the novel data collection, and $\beta_i^{(combined\ prior)} = \gamma_i^{(combined\ prior)} / \left(1 - \gamma_i^{(combined\ prior)}\right)$.

Proof. We have that

$$c\ell TDR(j) = \Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid T = t_j, S = s\right) =$$

$$\frac{\Pr\left(T = t_j \mid H_j^{(\mathrm{NDC})} = 1\right)\Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid S = s\right)}{\Pr\left(T = t_j \mid H_j^{(\mathrm{NDC})} = 0\right)\Pr\left(H_j^{(\mathrm{NDC})} = 0 \mid S = s\right) + \Pr\left(T = t_j \mid H_j^{(\mathrm{NDC})} = 1\right)\Pr\left(H_j^{(\mathrm{NDC})} = 1 \mid S = s\right)} =$$

$$\frac{\gamma_j^{(\mathrm{combined\ prior})} f_1(t_j)}{f_0(t_j)\left(1 - \gamma_j^{(\mathrm{combined\ prior})}\right) + \gamma_j^{(\mathrm{combined\ prior})} f_1(t_j)} = \frac{\beta_j^{(\mathrm{combined\ prior})} f_1(t_j) /}{f_0(t_j) + \beta_j^{(\mathrm{combined\ prior})} f_1(t_j)}.$$

■

The estimator of the $c\ell$TDR can be obtained by substituting $\beta_j^{(\mathrm{combined\ prior})}$, $f_0(t)$ and $f_1(t)$ and with their estimates in (14).

1.2 Estimating the contributions

Theorem 25 *Denote the number of test units that are both in the NDC and the existing data set as m. For a positive integer $M \leq m$ and real $d \geq 0$ we have that*

$$E(O_{d,M}) = (F_0(d) - F_1(d)) \sum_{j, r_j \leq M} co(j), \tag{15}$$

where

$$O_{d,M} = Q_{d,M} - \frac{M}{m} m'(d)$$

and $Q_{d,M}$ and $m'(d)$ are defined as

$$Q_{d,M} = \#\{j : |t_j| \geq d, r_j \leq M\} \text{ and } m'(d) = \#\{j : |t_j| \geq d\},$$

and $co(j)$ denotes the contribution of test unit j.

Proof. Denote the set of test units alternative in the EDS and the ones alternative in the NDC as E_1 and N_1, respectively. Moreover, denote the set of test units that are null in the EDS and the ones null in the NDC as E_0 and N_0, respectively. We have that

$$E\left(Q_{d,M} \cap E_1 \cap N_1\right) = m_1^{overlap}\left(1 - F_1\left(d\right)\right)\sum_{i=1}^{M}\frac{\gamma_i}{m_1} = \left\{\Gamma_1 = \frac{1}{M}\sum_{i=1}^{M}\gamma_i\right\} = m_1^{overlap}\left(1 - F_1\left(d\right)\right)\frac{M}{m_1}\Gamma_1,$$

$$E\left(Q_{d,M} \cap E_1 \cap N_0\right) = \left(m_1 - m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\sum_{i=1}^{M}\frac{\gamma_i}{m_1} = \left(m_1 - m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_1}\Gamma_1$$

$$E\left(Q_{d,M} \cap E_0 \cap N_1\right) = \left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\sum_{i=1}^{M}\frac{1-\gamma_i}{m_0} =$$

$$\left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\frac{1}{m_0}\left(M - \sum_{i=1}^{M}\gamma_i\right) = \left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\frac{1}{m_0}\left(M - M\Gamma_1\right) =$$

$$\left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right),$$

$$E\left(Q_{d,M} \cap E_0 \cap N_0\right) = \left(m - m_1 - m_1^{*} + m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\sum_{i=1}^{M}\frac{1-\gamma_i}{m_0} =$$

$$\left(m - m_1 - m_1^{*} + m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right).$$

Therefore, we have that

$$E\left(Q_{d,M}\right) = E\left(Q_{d,M} \cap E_1 \cap N_1 + Q_{d,M} \cap E_1 \cap N_0 + Q_{d,M} \cap E_0 \cap N_1 + Q_{d,M} \cap E_0 \cap N_0\right) =$$

$$m_1^{overlap}\left(1 - F_1\left(d\right)\right)\frac{M}{m_1}\Gamma_1 + \left(m_1 - m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_1}\Gamma_1 +$$

$$\left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right) + \left(m - m_1 - m_1^{*} + m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right).$$

Moreover, we have that

$$E\left(\frac{M}{m}m'(d)\right) = \frac{M}{m}m_1^{*}\left(1 - F_1\left(d\right)\right) + \frac{M}{m}\left(m - m_1^{*}\right)\left(1 - F_0\left(d\right)\right).$$

Combining the above two we obtain

$$E\left(Q_{d,M} - \frac{M}{m}m'(d)\right) =$$

$$m_1^{overlap}\left(1 - F_1\left(d\right)\right)\frac{M}{m_1}\Gamma_1 + \left(m_1 - m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_1}\Gamma_1 +$$

$$\left(m_1^{*} - m_1^{overlap}\right)\left(1 - F_1\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right) + \left(m - m_1 - m_1^{*} + m_1^{overlap}\right)\left(1 - F_0\left(d\right)\right)\frac{M}{m_0}\left(1 - \Gamma_1\right) -$$

$$\frac{M}{m}m_1^{*}\left(1 - F_1\left(d\right)\right) - \frac{M}{m}\left(m - m_1^{*}\right)\left(1 - F_0\left(d\right)\right) =$$

$$\left(1 - F_1\left(d\right)\right)\left\{m_1^{overlap}\frac{M}{m_1}\Gamma_1 + \left(m_1^{*} - m_1^{overlap}\right)\frac{M}{m_0}\left(1 - \Gamma_1\right) - \frac{M}{m}m_1^{*}\right\} +$$

$$\left(1 - F_0\left(d\right)\right)\left\{\left(m_1 - m_1^{overlap}\right)\frac{M}{m_1}\Gamma_1 + \left(m - m_1 - m_1^{*} + m_1^{overlap}\right)\frac{M}{m_0}\left(1 - \Gamma_1\right) - \frac{M}{m}\left(m - m_1^{*}\right)\right\} =$$

$$\left(1 - F_1\left(d\right)\right)M\left\{\left(m_1^{overlap}\frac{1}{m_1} - \left(m_1^{*} - m_1^{overlap}\right)\frac{1}{m_0}\right)\Gamma_1 + \left(m_1^{*} - m_1^{overlap}\right)\frac{1}{m_0} - \frac{1}{m}m_1^{*}\right\} +$$

$$(1-F_0(d))M\left\{\left[\left(m_1-m_1^{overlap}\right)\frac{1}{m_1}-\left(m-m_1-m_1^*+m_1^{overlap}\right)\frac{1}{m_0}\right]\Gamma_1+\right.$$
$$\left.\left(m-m_1-m_1^*+m_1^{overlap}\right)\frac{1}{m_0}-\frac{1}{m}\left(m-m_1^*\right)\right\}=$$
$$(1-F_1(d))M\left\{\left(m_1^{overlap}\frac{m}{m_1m_0}-\frac{m_1^*}{m_0}\right)\Gamma_1+\frac{m_1^*m_1}{mm_0}-\frac{m_1^{overlap}}{m_0}\right\}+$$
$$(1-F_0(d))M\left\{\left[\left(m_1-m_1^{overlap}\right)\frac{m}{m_1m_0}-\frac{(m-m_1^*)}{m_0}\right]\Gamma_1+(m-m_1^*)\frac{m_1}{mm_0}-\frac{m_1-m_1^{overlap}}{m_0}\right\}=$$
$$(1-F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1}-m_1^*\right)\Gamma_1+\frac{m_1^*m_1}{m}-m_1^{overlap}\right\}+$$
$$(1-F_0(d))\frac{M}{m_0}\left\{\left[\left(m_1-m_1^{overlap}\right)\frac{m}{m_1}-(m-m_1^*)\right]\Gamma_1+(m-m_1^*)\frac{m_1}{m}-m_1+m_1^{overlap}\right\}=$$
$$(1-F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1}-m_1^*\right)\Gamma_1+\frac{m_1^*m_1}{m}-m_1^{overlap}\right\}+$$
$$(1-F_0(d))\frac{M}{m_0}\left\{\left[m-m_1^{overlap}\frac{m}{m_1}-m+m_1^*\right]\Gamma_1+m_1-m_1^*\frac{m_1}{m}-m_1+m_1^{overlap}\right\}=$$
$$(1-F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1}-m_1^*\right)\Gamma_1+\frac{m_1^*m_1}{m}-m_1^{overlap}\right\}+$$
$$(1-F_0(d))\frac{M}{m_0}\left\{\left[m_1^*-m_1^{overlap}\frac{m}{m_1}\right]\Gamma_1-m_1^*\frac{m_1}{m}+m_1^{overlap}\right\}=$$
$$(F_0(d)-F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1}-m_1^*\right)\Gamma_1+\frac{m_1^*m_1}{m}-m_1^{overlap}\right\}=$$
$$(F_0(d)-F_1(d))\frac{M}{m_0}\left\{\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right)m\Gamma_1+\left(\frac{m_1^*}{m}-\frac{m_1^{overlap}}{m_1}\right)m_1\right\}=$$
$$(F_0(d)-F_1(d))\frac{M}{m_0}\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right)\{m\Gamma_1-m_1\}=$$
$$(F_0(d)-F_1(d))(mM\Gamma_1-Mm_1)\left[\frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right)\right]=\left\{\Gamma_1=\frac{1}{M}\sum\nolimits_{j=1}^{M}\gamma(j)\right\}$$
$$(F_0(d)-F_1(d))\sum\nolimits_{j=1}^{M}(m\gamma(j)-m_1)\left[\frac{1}{m_0}\left(\frac{m_1^{overlap}}{m_1}-\frac{m_1^*}{m}\right)\right]=(F_0(d)-F_1(d))\sum\nolimits_{j=1}^{M}co(j),$$

which concludes the proof of the theorem. ■

1.3 Smoother

In this section we present a method that smooths the $M \to \widetilde{CO}(M)$ to a concave curve. In particular, we will choose the best fitting curve from the family of all $M \to CO(M)$ curves that can be attained by a suitable selection of parameters. We emphasize that our goal is not to find accurate estimates of the parameters, but to estimate the contributions accurately by a curve-fitting method. As a matter of fact, we may obtain similar curves by infinite many choices of parameter sets, thus, it is possible to approximate the true $M \to CO(M)$ curve very well while the parameters of the approximating curve may be far from the real parameter values.

From (15) and the definition of $\widetilde{CO}(M)$ and the contribution we have that

$$E\left(\widetilde{CO}(M)\right) = \sum_{j,r_j \le M} co(j) = \frac{m\kappa}{m_0} \sum_{j,r_j \le M} \left(\gamma(r_j, m_1, \Psi) - \frac{m_1}{m}\right), \tag{16}$$

where the term on the right-hand side depends on three unknown parameters, κ, Ψ and m_1, as $m_0 = m - m_1$. We will use the approximation

$$\gamma(r; m_1, \Psi) \approx \frac{m_1}{\widetilde{m}_1} \gamma(r; \widetilde{m}_1, \Psi) \tag{17}$$

where $\widetilde{m}_1$ is an arbitrary choice, say our guess for the number of alternatives in the existing data set. By applying approximation (17) to the term on the right-hand side in (16), we obtain

$$E\left(\widetilde{CO}(M)\right) \approx \frac{m}{\widetilde{m}_1} \frac{\kappa m_1}{m_0} \sum_{j,r_j \le M} \left(\gamma(r; \widetilde{m}_1, \Psi) - \frac{\widetilde{m}_1}{m}\right). \tag{18}$$

where the term on the right-hand side depends on only two unknown parameters, $\kappa^* = \kappa m_1/m_0$ and Ψ, thus, it will be denoted as $CO(M; \kappa^*, \Psi)$. By a curve-fitting method, we find $\overline{\kappa^*}$ and $\overline{\Psi}$ that provides the best fitting curve to $M \to \widetilde{CO}(M)$ from the family of curves $\{M \to CO(M; \kappa^*, \Psi), 0 < \kappa^*, \Psi\}$. Finally, $CO\left(M; \overline{\kappa^*}, \overline{\Psi}\right)$ will be our estimate of the cumulative contribution, $CO(M) = \sum_{j=1}^{M} co(j)$. In the course of the algorithm, the term $\gamma(r; \widetilde{m}_1, \Psi)$ is computationally evaluated by the method described in section 1.3. In principle, we could use (20) or (21) to compute γs, however, the numerical integration required is computationally very intensive, especially for large $\widetilde{m}_1$, say 1,000 or larger. We remark that $\overline{\kappa^*}$ and $\overline{\Psi}$ are not meant to be the estimator of κ^* and Ψ, as a matter of fact they may be far from the real κ^* and Ψ. Note that this is not a problem as long as $CO\left(M; \overline{\kappa^*}, \overline{\Psi}\right)$ is an accurate estimator of the cumulative contribution. We remark that we have an estimator of Ψ, which plugged in (20) provides the estimator $\widehat{\kappa^*}$ by a curve-fitting method. From (4) we have that $m\widehat{\kappa^*}$ is a tight lower bound estimate of $m_1^{overlap}$.

Computation of the rank-based probability that a test unit has an effect in the existing data set, γ

In this subsection first we present an algorithm that computes the probability that a test unit is alternative in a data set, γ, utilizing the rank of the test statistic of the test unit in the data set. For completeness we will also present the formulas of γ, although using the algorithm is computationally much faster and not less accurate than applying numerical integrals to evaluate the formulas. We assume that the c.d.f. of the test statistic values is F_0 and F_Ψ under the null and the alternative hypothesis, respectively. First we deal with the case when there are no ties in the ranking, then we deal with the case when there are ties in the ranking. Ties occur when the test units are sorted in a couple of categories, and our prior information does not distinguish between test units in the same category. For instance, we have only two categories if we have a candidate gene list, the test units that are on the list and those that are not on the list.

More formally, suppose that $X_1, ..., X_{m_0}$ are identically distributed random variables with c.d.f. F_0, and $Y_1, ..., Y_{m_1}$ are identically distributed random variables with F_Ψ. Suppose that $X_1, ..., X_{m_0}, Y_1, ..., Y_{m_1}$ are independent, and denote the rth largest random variable among them as Z_r. The probability that Z_r is an alternative will be denoted as $\gamma(r; m_1, \Psi)$, or shortly γ_r. In this section we give an algorithm and formulas in order to compute γ_r.

Algorithmic computation of the probability that a test unit is alternative in the a data set based on its rank, γ Based on the following theorem, $\sum_{i=1}^{k}\gamma_i$ can be computed algorithmically, from which γ_is can readily be calculated.

Theorem 26 *For an integer $k = 1, ..., m$ we have that*

$$\sum_{i=1}^{k}\gamma_i = \begin{cases} a & if \quad k < m \\ m_1 & if \quad k = m, \end{cases}$$

where a is the solution of the equation

$$m_1 M\left(\frac{k-a}{m_0}\right) = a,$$

where $M(p)$ denotes the c.d.f. of the alternative p-values, i.e. $M(p) = 1 - F_1\left(F_0^{-1}(1-p)\right)$. Moreover, $a \in (\max(k - m_0, 0), \min(k, m_1))$.

Proof. For $k = m$, the statement of the theorem follows from (2). Therefore, for the rest of the proof we can assume that $k < m$. By definition of γ_i, we have that

$$\sum_{i=1}^{k}\gamma_i = E(\#\text{alternatives in the top } k \text{ test statistic values}) = E\left(\#\left(p : p \in P_{alt},\ p < \frac{k-a}{m_0}\right)\right), \tag{19}$$

where P_{alt} denotes the set of p-values of true alternative test units and a is selected in such a way that $a = E\left(\#\left(p : p \in P_{alt},\ p < \frac{k-a}{m_0}\right)\right)$. The equation in (2) holds because the expected number of alternative and null p-values smaller than $\frac{k-a}{m_0}$ is a and $k-a$, respectively. Also, as we have that

$$E\left(\#\left(p : p \in P_{alt},\ p < \frac{k-a}{m_0}\right)\right) = m_1 \Pr\left(p < \frac{k-a}{m_0} \mid p \in P_{alt}\right) = m_1 M\left(\frac{k-a}{m_0}\right),$$

where $M(p)$ denotes the c.d.f. of alternative p-values. In order to obtain $\sum_{i=1}^{k}\gamma_i$ we need to solve

$$m_1 M\left(\frac{k-a}{m_0}\right) = a$$

for a.

From (2) we have that

$$0 < a = \sum_{i=1}^{k}\gamma_i < m_1,$$

for $k = 1, ..., m-1$, and, clearly $a = \sum_{i=1}^{k}\gamma_i < \sum_{i=1}^{k}1 = k$, thus, $a < \min(k, m_1)$. As $m_1 > a = E\left(\#\left(p : p \in P_{alt},\ p < \frac{k-a}{m_0}\right)\right)$, we have that $\frac{k-a}{m_0} < 1$, which implies $k - m_0 < a$. As $0 < a$, we have that $\max(k - m_0, 0) < a$, which completes the proof of the theorem. ■

Mathematical formulas of the probability that a test unit is alternative in the a data set based on its rank, γ Now we give formulas of γ for the case when there are ties,and the case when there are no ties among the ranks.

Theorem 27 *Suppose that $X_1, ..., X_{m_0}$ are i.i.d. random variables with F_0 and $Y_1, ..., Y_{m_1}$ are i.i.d random variables with F_Ψ. Suppose that $X_1, ..., X_{m_0}, Y_1, ..., Y_{m_1}$ are independent, and denote the rth largest random variable among them as Z_r. Then the probability that Z_r is an alternative can be calculated as*

$$\gamma(r; m_1, \Psi) := \Pr(Z_r \text{ is alternative}) =$$

$$m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=\max(0, r-m_0-1)}^{\min(m_1-1, r-1)} \binom{m_1-1}{j} (1 - F_\Psi(x))^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} (1 - F_0(x))^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\} f_\Psi(x)\, dx. \tag{20}$$

To prove the theorem we need the following lemma.

Lemma 28 *The marginal density function of Z_r intersected with the event that Z_r is alternative can be obtained as*

$$g_r(x) := \Pr(Z_r = x;\ Z_r \text{ is alternative}) =$$

$$m_1 f_\Psi(x) \sum_{i=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{i} (1-F_\Psi(x))^i F_\Psi^{m_1-1-i}(x) \binom{m_0}{r-1-i} (1-F_0(x))^{r-1-i} F_0^{m_0-(r-1-i)}(x).$$

Proof.

$$g_r(x) = \lim_{\delta\to 0} \frac{G_r(x+\delta) - G_r(x)}{\delta} = \lim_{\delta\to 0} \frac{\Pr\left(Z_{(r)} \in (x, x+\delta)\right)}{\delta} =$$

$$\lim_{\delta\to 0} \frac{\Pr\left(\#\{Z_j < x\} = n-r; \Pr(Y \in (x,x+\delta)); \#\{Z_j > x+\delta\} = r-1\right)}{\delta} =$$

$$\lim_{\delta\to 0} \frac{m_1 \Pr(Y \in (x,x+\delta))}{\delta} \sum_{i=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \Pr(\{\#\{Y_j > x+\delta\} = i;\ \#\{Y_j < x\} = m_1-1-i\} \cap$$

$$\{\#\{X_j > x+\delta\} = r-1-i;\ \#\{X_j < x\} = m_0-(r-1-i)\}) =$$

$$\lim_{\delta\to 0} \frac{m_1 \Pr(Y \in (x,x+\delta))}{\delta} \sum_{i=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{i} (1-F_\Psi(x))^i F_\Psi^{m_1-1-i}(x) \binom{m_0}{r-1-i} (1-F_0(x))^{r-1-i} F_0^{m_0-(r-1-i)}(x) =$$

$$m_1 f_\Psi(x) \sum_{i=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{i} (1-F_\Psi(x))^i F^{m_1-1-i}(x) \binom{m_0}{r-1-i} (1-F_0(x))^{r-1-i} F_0^{m_0-(r-1-i)}(x).$$

■

Proof of the Theorem. Utilizing the lemma we have that

$$\gamma_r = \Pr(Z_r \text{ is alternative}) = \int_{-\infty}^{\infty} \Pr(Z_r = x;\ Z_r \text{ is alternative})\, dx = \int_{-\infty}^{\infty} g_r(x)\, dx =$$

$$m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} (1-F_\Psi(x))^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} (1-F_0(x))^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\} f_\Psi(x)\, dx,$$

which completes the proof of the theorem. ■

Theorem 29 *With the reasonable assumption that $m_1 \le m_0$ we have that*

$$\Pr(Z_i \text{ is alternative} \mid i \le R) =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} b(j, m_1-1, 1-F_\Psi(x))\, B(\min(m_0, R-1-j), m_0, 1-F_0(x)) \right\} f_\Psi(x)\, dx, \tag{21}$$

where $b(x,n,p)$ and $B(x,n,p)$ are the binomial density and the distribution function values, respectively, at point x.

Proof.

$$\Pr(Z_i \text{ is alternative} \mid i \le R) = \frac{\sum_{r=1}^{R} \gamma_r}{R} =$$

$$\frac{1}{R} \sum_{r=1}^{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} (1-F_\Psi(x))^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} (1-F_0(x))^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\, dx =$$

$$\frac{1}{R}\sum_{r=1}^{R} m_1 \int_{-\infty}^{\infty} ...dx = \frac{1}{R}\sum_{r=1}^{\min(m_0,R)} m_1 \int_{-\infty}^{\infty} ...dx + \chi\left(R \geq m_0+1\right)\frac{1}{R}\sum_{r=m_0+1}^{R} m_1 \int_{-\infty}^{\infty} ...dx = *, \tag{22}$$

where $\chi(A)$ is the indicator function, i.e. $\chi(A)=1$ if A holds, and $\chi(A)=0$ otherwise. First we calculate the first sum. The rationale is that for $r \leq \min(m_0, R)$ we have $\max(0, r-m_0-1)=0$.

$$\frac{1}{R}\sum_{r=1}^{\min(m_0,R)}$$

$$m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\} f_\Psi(x)\, dx =$$

$$\frac{1}{R}\sum_{r=1}^{\min(m_0,R)}$$

$$m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\} f_\Psi(x)\, dx = *$$

for the change of the sum

$$\sum_{r=1}^{R}\sum_{j=0}^{\min(m_1-1,r-1)} = \sum_{r=1}^{\min(m_1,R)}\sum_{j=0}^{r-1} + \chi\left(R>m_1\right)\sum_{r=m_1+1}^{R}\sum_{j=0}^{m_1-1} =$$

$$\sum_{r=1}^{\min(m_0,R)}\sum_{j=0}^{\min(m_1-1,r-1)} = \sum_{r=1}^{\min(m_1,R)}\sum_{j=0}^{r-1} + \chi\left(R>m_1\right)\sum_{r=m_1+1}^{\min(m_0,R)}\sum_{j=0}^{m_1-1} =$$

$$\begin{cases} \sum_{r=1}^{R}\sum_{j=0}^{r-1} & \text{if} \quad R \leq m_1 \\ \sum_{r=1}^{m_1}\sum_{j=0}^{r-1} + \sum_{r=m_1+1}^{\min(m_0,R)}\sum_{j=0}^{m_1-1} & \text{if} \quad R > m_1 \end{cases} = \{\text{sum change}\} =$$

$$\begin{cases} \sum_{j=0}^{R-1}\sum_{r=j+1}^{R} & \text{if} \quad R \leq m_1 \\ \sum_{j=0}^{m_1-1}\sum_{r=j+1}^{m_1} + \sum_{j=0}^{m_1-1}\sum_{r=m_1+1}^{\min(m_0,R)} & \text{if} \quad R > m_1 \end{cases} =$$

$$\begin{cases} \sum_{j=0}^{R-1}\sum_{r=j+1}^{R} & \text{if} \quad R \leq m_1 \\ \sum_{j=0}^{m_1-1}\sum_{r=j+1}^{\min(m_0,R)} & \text{if} \quad R > m_1 \end{cases} = \{m_1 \leq m_0\} = \sum_{j=0}^{\min(m_1,R)-1}\sum_{r=j+1}^{\min(m_0,R)}$$

$$* = \frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1}\sum_{r=j+1}^{\min(m_0,R)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\, dx =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \sum_{r=j+1}^{\min(m_0,R)} \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\, dx = \{s = r-1-j\} =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \left[\sum_{s=0}^{\min(m_0,R)-1-j} \binom{m_0}{s} \left(1-F_0(x)\right)^s F_0^{m_0-s}(x) \right] \right\} f_\Psi(x)\, dx =$$

$$\frac{1}{R} m_1 \int_{0}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) B\left(\min(m_0,R)-1-j, m_0, 1-F_0(x)\right) \right\} f_\Psi(x)\, dx.$$

For the second sum in (22), as $r \geq m_0+1$ implies $\max(0, r-m_0-1) = r-m_0-1$, we have that

$$\frac{1}{R}\sum_{r=m_0+1}^{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=\max(0,r-m_0-1)}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\, dx =$$

$$\frac{1}{R}\sum_{r=m_0+1}^{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=r-m_0-1}^{\min(m_1-1,r-1)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\,dx = *$$

for the change of the sum we use that $m_1 \leq m_0$

$$\sum_{r=m_0+1}^{R}\sum_{j=r-m_0-1}^{\min(m_1-1,r-1)} = \sum_{r=m_0+1}^{R}\sum_{j=r-m_0-1}^{m_1-1} = \{\text{sum change}\} = \sum_{j=0}^{m_1-1}\sum_{r=m_0+1}^{\min(j+m_0+1,R)}$$

$$* = \frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1}\sum_{r=m_0+1}^{\min(j+m_0+1,R)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{r-1-j} \left(1-F_0(x)\right)^{r-1-j} F_0^{m_0-(r-1-j)}(x) \right\}$$

$$f_\Psi(x)\,dx = \{s = r - m_0 - 1\} =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1}\sum_{s=0}^{\min(j,R-m_0-1)} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \binom{m_0}{s+m_0-j} \left(1-F_0(x)\right)^{s+m_0-j} F_0^{j-s}(x) \right\}$$

$$f_\Psi(x)\,dx =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1} \binom{m_1-1}{j} \left(1-F_\Psi(x)\right)^j F_\Psi^{m_1-1-j}(x) \left[\sum_{s=0}^{\min(j,R-m_0-1)} \binom{m_0}{s+m_0-j} \left(1-F_0(x)\right)^{s+m_0-j} F_0^{j-s}(x) \right] \right\}$$

$$f_\Psi(x)\,dx =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) \left[B\left(\min\left(m_0, R-1-j\right), m_0, 1-F_0(x)\right) - B\left(m_0-j-1, m_0, 1-F_0(x)\right) \right] \right\} f_\Psi(x)\,dx.$$

Putting the two sums together, from (22) we have that

$$\frac{1}{R} m_1 \int_{0}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) B\left(\min\left(m_0, R\right)-1-j, m_0, 1-F_0(x)\right) \right\} f_\Psi(x)\,dx +$$

$$\chi\left(R \geq m_0+1\right) \frac{1}{R} m_1$$

$$\int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) \left[B\left(\min\left(m_0, R-1-j\right), m_0, 1-F_0(x)\right) - B\left(m_0-j-1, m_0, 1-F_0(x)\right) \right] \right\} f_\Psi(x)\,dx.$$

If $R \geq m_0 + 1$, then we have that

$$\frac{1}{R} m_1 \int_{0}^{\infty} \left\{ \sum_{j=0}^{m_1-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) B\left(m_0-1-j, m_0, 1-F_0(x)\right) \right\} f_\Psi(x)\,dx +$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) \left[B\left(\min\left(m_0, R-1-j\right), m_0, 1-F_0(x)\right) - B\left(m_0-j-1, m_0, 1-F_0(x)\right) \right] \right\} f_\Psi(x)\,dx =$$

$$\frac{1}{R} m_1 \int_{-\infty}^{\infty} \left\{ \sum_{j=0}^{m_1-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) B\left(\min\left(m_0, R-1-j\right), m_0, 1-F_0(x)\right) \right\} f_\Psi(x)\,dx.$$

If $R \leq m_0$, then we have that

$$\frac{1}{R} m_1 \int_{0}^{\infty} \left\{ \sum_{j=0}^{\min(m_1,R)-1} b\left(j, m_1-1, 1-F_\Psi(x)\right) B\left(R-1-j, m_0, 1-F_0(x)\right) \right\} f_\Psi(x)\,dx.$$

■

1.4 Overview of the algorithm that computes the estimates of $c\ell TDR$

First we give the algorithm supposing that we have no ties in the ranks of the existing data sets, then we show how this needs to be modified for the existing data in which there are ties in the ranks.

1. First we need to estimate $F_0(d)$ and $F_1(d)$ in the novel data collection. (Note that here both $F_0(d)$ and $F_1(d)$ can be a mixture distribution, i.e. $F_0(d)$ and $F_1(d)$ merely denotes the null and alternative distribution in the novel data collection.)

2. We estimate the cumulative contribution, $CO(M) = \sum_{j=1}^{M} co(j)$ by first calculating

$$\widetilde{CO}(M) = \frac{\frac{1}{|D|}\sum_{d\in D}(O_{d,M})}{F_0(d) - F_1(d)} \tag{23}$$

for $M = 1, ..., m$, where D is a set of the positive real numbers, $|D|$ is the number of elements in D. Then we apply a smoother method (see section 1.3) to fit $M \to \widetilde{CO}(M)$ curve with a concave $M \to \widehat{CO}(M)$ a concave function of M.

3. From $\widehat{CO}(M)$ we calculate the estimator

$$\widehat{co}(i) = \widehat{CO}(i) - \widehat{CO}(i-1)$$

for $i = 1, ..., m$, where we define $\widehat{CO}(0) = 0$.

4. Then $\widehat{co}(i)$ is used to calculate

$$\widehat{\gamma}_i^* = \widehat{co}(i) + \frac{m_1^*}{m}$$

for every existing data set.

5. Then we use

$$\widehat{\beta}_i^{(\text{combined prior})} = \left(\frac{m_0^*}{m_1^*}\right)^{k-1} \prod_{i=1}^{k} \frac{\widehat{\gamma}_i^{*\,j}}{1 - \widehat{\gamma}_i^{*\,j}} \tag{24}$$

to calculate the estimates of combined prior odd for test unit i (see formula 13), where m_0^*/m_1^* is the ratio of the null and alternative test units (markers) in the novel data collection, and $\widehat{\gamma}_i^{*\,j}$ is the prior probability estimate of test unit i from the jth existing data set obtained in step 4.

6. The estimate of $c\ell TDR$ of test unit i will be calculated by

$$c\ell TDR(i) = \frac{\beta_i^{(\text{combined prior})} f_1(t)/f_0(t)}{1 + \beta_i^{(\text{combined prior})} f_1(t)/f_0(t)} = \frac{\beta_i^{(\text{combined prior})} f_1(t)}{f_0(t) + \beta_i^{(\text{combined prior})} f_1(t)},$$

where f_0 and f_1 is the null and alternative p.d.f. in the novel data collection.

If there are ties among the ranks in an EDS, then Step 2 and 3 are modified for that EDS in the following way.

Suppose we have t groups of test units and the ranks of all test units in a group are identical, but they are different across the groups. Let R_j be the number of test units whose rank is the jth smallest one or smaller than that for $j = 1, ..., t$. Then we calculate $\widetilde{CO}(M)$ only for $M = R_1, ..., R_t$ in Step 2, and in Step 3 the estimator $\widehat{co}$ is obtained as

$$\widehat{co}(i) = \frac{1}{R_j - R_{j-1}}\left(\widehat{CO}(R_j) - \widehat{CO}(R_{j-1})\right)$$

for every test unit i in group j, $j = 1, ..., t$, where we define $R_0 = 0$ and $\widehat{CO}(0) = 0$ for the sake of simplicity (see (5) for justification).

We remark that in order to decrease computational burden, this modification in Step 2 and 3 can be used for the case of no ties as well. Note that in case of no ties, the choice of $R_1 < ... < R_t$ is not determined by the groups of ties in the rank of test units in the EDS, but the desired accuracy of the contribution estimator.