

MATHEMATICALLY-BASED  
INTEGRATION OF HETEROGENEOUS  
DATA

József Bukszár<sup>a</sup>      Edwin J. C. G. van den Oord<sup>b</sup>

RRR 16-2011, AUGUST 17, 2011

RUTCOR  
Rutgers Center for  
Operations Research  
Rutgers University  
640 Bartholomew Road  
Piscataway, New Jersey  
08854-8003  
Telephone: 732-445-3804  
Telefax: 732-445-5472  
Email: [rrr@rutcor.rutgers.edu](mailto:rrr@rutcor.rutgers.edu)  
<http://rutcor.rutgers.edu/~rrr>

---

<sup>a</sup>Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA

<sup>b</sup>Center for Biomarker Research and Personalized Medicine, Virginia Commonwealth University, Richmond, VA 23298, USA

RUTCOR RESEARCH REPORT  
RRR 16-2011, AUGUST 17, 2011

# MATHEMATICALLY-BASED INTEGRATION OF HETEROGENEOUS DATA

József Bukszár      Edwin J. C. G. van den Oord

**Abstract.** We present the exact formula, and based on that an estimate of the posterior probability that a hypothesis is true alternative in a multiple-hypothesis set-up utilizing information from several external data sets. Each external data set is a ranked list of a subset of the hypotheses, where for the ranks we merely assume that a hypothesis that is alternative in the external data set is more likely to have smaller rank than a hypothesis that is null in the external data set. An alternative hypothesis may be null in some or all of the external data sets, and also a null hypothesis may be alternative in some or all of the external data sets. The work is motivated by the problem of identifying biomarkers in a genetic data set that are associated with a complex disease utilizing several heterogeneous data sets.

---

**Acknowledgements:** This work was supported by NIH grant R01 HG004240.

# 1 Introduction

Given  $m$  hypotheses,  $H_1, \dots, H_m$ ,  $m_1$  of which are alternative and the remaining  $m_0 = m - m_1$  are null. We have observed test statistics,  $t_1, \dots, t_m$ , called primary data/statistics, where  $t_i$  is drawn from distribution  $F_1$  if  $H_i$  is alternative and  $t_i$  is drawn from distribution  $F_0$ , if  $H_i$  is null. Both  $F_1$  and  $F_0$  are supposed to be known, however, we do not know which hypotheses are alternative and which are null. In addition, we have information from multiple independent external data sets (EDSs). The information from an EDS means that we have ranks on a subset of hypotheses. Each hypothesis in an EDS is either alternative or null in the EDS. For the ranks we merely assume that a hypothesis that is *alternative in the EDS* is more likely to have smaller rank in the EDS than a hypothesis that is *null in the EDS*. An alternative hypothesis may be null in some or all of the EDSs, and also a null hypothesis may be alternative in some or all of the EDSs. We can utilize information from an EDS only if it is *informative*. An EDS is defined to be *informative* if the number of alternative hypotheses that are also alternative in the EDS is larger than it would be expected by chance, i.e. when the label "alternative in the EDS" would be randomly assigned to the hypotheses in the EDS. We will refer to the fact that we have information for hypothesis  $i$  from an EDS as "the hypothesis  $i$  is in that EDS".

We will refer to the probability that a hypothesis is alternative based on the information we have in the primary statistics and external data sets as the *compound posterior probability* (cpp). We present an exact mathematical formula of the cpp, and we will also give an estimate of the cpp of a hypothesis. We suppose that the null and alternative p.d.f.,  $f_1$  and  $f_0$ , as well as  $m_1$  are known. There are numerous estimators of the null and alternative p.d.f., c.d.f. and the number of alternative hypotheses, that can readily be used in our framework.

Our work is motivated by the problem of identifying genetic units in a genetic (primary) data set that are associated with a complex disease. As collecting genetic data is costly, analysis on the primary data set only is likely to be underpowered. On the other hand, the number of publicly available databases of complex diseases has grown dramatically and this trend is likely to continue. Therefore, it is a natural idea to utilize information from the publicly available data sets. However, these data sets may be of different type (gene expression, linkage, genotyping data etc.), thus, traditional meta-analysis techniques, such as taking weighted average of statistics in different data sets, cannot be applied. On the other hand, our framework can be applied as long as the primary data set has test statistics, as we merely assume that we have ranked data (priority list) from an external source (data set).

In order to use our framework, the following technical step needs to be carried out. The data sets we wish to involve in the analysis need to be converted into data sets that have the same genetic unit, called test unit. For instance, if we decide the test unit to be a SNP, then we need to convert the gene-based data sets (e.g., gene expression study) into SNP-based data. A simple approach in this instance would be to assign the same (rank of) p-value to each SNP in the gene, or the smallest p-value of the genes if the SNP is in more than one gene. Each test unit in the primary data set is identified with a hypothesis.

A hypothesis is alternative in a data set if it has a statistical effect in that data set. A statistical effect can be either a real effect or a confounding effect. A confounding effect may occur due to stratification or other artifacts. Note that we account for the confounding effects as in our model an alternative hypothesis may be null in some or all of the EDSs, and also a null hypothesis may be alternative in some or all of the EDSs. We emphasize that allowing different null/alternative status for a hypothesis/test unit across multiple data sets is also important because of the sheer difference between the number of alternative hypotheses in different type of data sets. For instance, the proportion of differentially expressed genes can be as high as 30% in a microarray data set, whereas the proportion of SNPs that show significance is typically less than 1% in a GWAS.

## 2 METHOD

### 2.1 The mathematical formula of the cpp

The goal of this section is to present the mathematical formula of the cpp. This will be done in the following 3 steps, where in each step we present a formula for

1. the prior probability that a hypothesis is alternative based on the information in an EDS,
2. the (combined) prior probability that a hypothesis is alternative based on the information in all EDSs available, and
3. the cpp.

#### 2.1.1 Step 1: The formula of the prior probability that a hypothesis is alternative based on the information in an EDS

In this section we present the mathematical formula for the exact value of the prior probability that a hypothesis is alternative based on the information in an EDS. While this prior probability is based on multiple unknown parameters, almost all of these unknown parameters can be aggregated into a single term, which we will call the *contribution of the EDS* to the particular hypothesis. Interestingly, the contribution and the number of alternative hypotheses,  $m_1$ , together are sufficient to calculate an approximation of the prior probability of a hypothesis. First we need some definitions.

**Definition 1** *The information parameter of an EDS is defined as*

$$\kappa = \frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}} - \frac{m_1^*}{m^{\text{eds}}}, \quad (1)$$

where  $m^{\text{eds}}$  is the number of hypotheses in the EDS,  $m_1^{\text{eds}}$  is the number of hypotheses that are alternative in the EDS,  $m_1^*$  is the number of alternative hypotheses that are in the EDS,

and, finally,  $m_1^{overlap}$  is the number of alternative hypotheses that are also alternative in the EDS.

Note that  $\kappa > 0$  if and only if the number of alternative hypotheses that are also alternative in the EDS is larger than expected by chance, i.e. when the label "alternative in the EDS" is randomly assigned to the hypotheses in the EDS. We will call an EDS *informative* if its information parameter is positive.

**Definition 2** For hypothesis  $i$  in an EDS, we define the contribution of the EDS to hypothesis  $i$  as

$$co(i) = \frac{\kappa}{m_0^{eds}} (m^{eds} \gamma(r_i) - m_1^{eds}) \quad (2)$$

where  $r_i$  is the rank of hypothesis  $i$  in the EDS,  $\gamma(r)$  is the probability that a hypothesis ranked  $r$  in the EDS is alternative in the EDS,  $\kappa$  is the information parameter of the EDS,  $m_1^{eds}$  is the number of hypotheses alternative in the EDS, and  $m_0^{eds} = m^{eds} - m_1^{eds}$ .

It is interesting to remark, that the total contribution of an EDS to all hypotheses is 0, i.e.  $\sum_{i \in EDS} co(i) = 0$ , where the summation is on the hypotheses that are in the EDS (see Lemma 10 in Appendix).

**Notation 3** We will use the notation  $\gamma_i^*$  for the prior probability that a hypothesis  $i$  is alternative based on the information in the EDS.

Based on the information in an EDS, for the prior probability that hypothesis  $i$  is alternative,  $\gamma_i^*$ , we have that

$$\gamma_i^* = \begin{cases} co(i) + \frac{m_1^*}{m^{eds}} & \text{if hypothesis } i \text{ is in the EDS} \\ \frac{m_1 - m_1^*}{m - m^{eds}} & \text{if hypothesis } i \text{ is not in the EDS,} \end{cases} \quad (3)$$

where  $co(i)$  is the contribution of the EDS to hypothesis  $i$  defined in (2),  $m$  is the number of hypotheses,  $m^{eds}$  is the number of hypotheses in the EDS,  $m_1$  is the number of alternative hypotheses, and  $m_1^*$  is the number of alternative hypotheses that are in the EDS (for the proof see Theorem 6 in Appendix). We remark that the information parameter being 0 implies 0 contributions for all hypotheses, which results in  $\gamma_i^* = m_1^*/m^{eds}$  for every hypothesis in the EDS. Note that this is exactly what we would have in case of no prior information. Another property of the formula in (3) is that, as all the contributions of an EDS sum up to zero on the hypotheses of the EDS (see Lemma 10 in Appendix),  $\gamma_i^*$  sum up to  $m_1$  on the hypotheses for every EDS. In other words, using an EDS redistributes the total amount of prior probabilities among the hypotheses.

The contribution of a hypothesis mainly depends on 3 factors: 1) the rank of the hypothesis in the EDS, 2) the information parameter of the EDS and 3) the extent to which a hypothesis that is alternative in the EDS is more likely to have smaller rank than a hypothesis that is null in the EDS (the "effect size" of the hypotheses alternative in the EDS). These

latter two, intuitively speaking, stretch out the spectrum of 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 (contribution) matter, which we can estimate from the data.

In practice we can approximate  $m_1^*/m^{eds}$  by  $m_1/m$ , where the rationale is that the group of hypotheses we have EDS information for should contain proportionally as many alternatives as the whole set of hypotheses does (see Corollary 7 in Appendix). This yields

$$\gamma_i^* \approx \begin{cases} co(i) + \frac{m_1}{m} & \text{if hypothesis } i \text{ is in the EDS} \\ \frac{m_1}{m} & \text{if hypothesis } i \text{ is \textbf{not} in the EDS} \end{cases} \quad (4)$$

(see Corollary 7 in Appendix). The advantage of (4) over (3) is that only the contribution and  $m_1$  need to be known/estimated. On the other hand, if we think that the above assumption is violated, we can still estimate  $m_1^*$  in the same way as  $m_1$  is estimated, and use the formula in (3) to obtain the estimate of  $\gamma_i^*$  for practical purpose.

### 2.1.2 Step 2: The (combined) prior probability that a hypothesis is alternative based on the information in all EDSs available

Once we have the prior probability estimates from every external data set, we can compute the combined prior odd  $\beta_i^{(\text{combined prior})} = \gamma_i^{(\text{combined prior})} / (1 - \gamma_i^{(\text{combined prior})})$  that hypothesis  $i$  is alternative for every  $i$  by

$$\beta_i^{(\text{combined prior})} = \frac{m_1}{m_0} \prod_{j=1}^k \frac{m_0 \gamma_i^{*j}}{m_1 (1 - \gamma_i^{*j})}, \quad (5)$$

where  $\gamma_i^{*j}$  is the  $j$ th external data set-based estimate of the prior probability that hypothesis  $i$  is alternative,  $m_1$  is the number of alternative hypotheses, and  $m_0 = m - m_1$  (see Corollary 16 in Appendix). We remark that the exact general formula of the combined prior odd is given in Theorem 12 in Appendix. The exact general formula, however, depends on how the set of alternative hypotheses and the sets of alternative hypotheses that are alternative in the EDSs overlap. This overlap structure is, however, very difficult if not impossible to estimate in practice. On the other hand, we proved that the exact general formula reduces to the formula in (5) for two important scenarios. One of these two scenarios is where the events that hypothesis  $j$  is alternative in different EDS is independent with the probability measure  $\Pr(\cdot | H_j = 1)$  and  $\Pr(\cdot | H_j = 0)$  for every  $j$ , and the other scenario is where a hypothesis is alternative if and only if it is alternative in every EDS. These two scenarios represent two extrema, and the formula in (5) is likely to be sufficiently accurate for realistic scenarios (for details see Remark 17 in Appendix). Another pleasant feature of the formula in (5) is that we obtain  $\gamma_i^{(\text{combined prior})} = \gamma_i^*$  as a special case if we have a single EDS, i.e. (5) is always accurate if we have only one EDS. We conclude that based on only the prior

probabilities of the EDS (5) is the most accurate formula of the combined prior odd, which is even exact for some scenarios.

Note that if we have no prior information for hypothesis  $i$  in any EDS, then from (3) we have that  $\gamma_i^{*j} = m_1/m$  for every  $j$ , which plugged in (5) implies  $\beta_i^{(\text{combined prior})} = m_1/m_0$ . Note that this is exactly what we supposed to obtain for a hypothesis we have no prior information for. Moreover, according to the formula in (5), the combined odd of a hypothesis is proportional to the geometrical mean of the prior odds of the hypothesis across the EDSs. If a hypothesis performs better than a hypothesis 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 hypothesis performs worse than a hypothesis with no information in an EDS, then its odd in that EDS will have a negative (decreasing) impact on its combined odd.

### 2.1.3 Step 3: The compound posterior probability (cpp)

The compound posterior probability that hypothesis  $i$  is alternative based on the information we have from the EDSs and the test statistics can be formally written as

$$cpp(i) = \Pr \left( H_i = 1 \mid T_i = t_i, S_i^{(j)} = s_i^{(j)}, i = 1, \dots, k \right),$$

where  $H_i = 1$  represents that hypothesis  $i$  is alternative,  $t_i$  is the observed test statistic value of hypothesis  $i$  and  $s_i^{(j)}$  is the observed rank of the hypothesis  $i$  in the  $j$ th EDS,  $j = 1, \dots, k$ . It is easy to show that the cpp of hypothesis  $i$  can be obtained as

$$cpp(i) = \frac{\beta_i^{(\text{combined prior})} f_1(t)}{f_0(t) + \beta_i^{(\text{combined prior})} f_1(t)}, \quad (6)$$

where  $f_0(t)$  and  $f_1(t)$  is the null and alternative p.d.f. of the statistics, respectively, and  $\beta_i^{(\text{combined prior})}$  is obtained by (5).

In summary, the cpp of hypotheses can be obtained by the three steps described above, using (3), (5) and in Appendix (6), based on  $m_1$  ( $m_1^*$ ),  $f_0(t)$  and  $f_1(t)$ , and the contributions of hypotheses from each external data sets. Instead of the terms cpp depends on, we will use their estimates. In the next section we present a method that estimates the contributions.

## 2.2 Estimate of the cpp via estimating the contributions

In section 2.1 we presented formulas of how the compound posterior probability, cpp, can be obtained from the number of alternative hypotheses, the null and the alternative p.d.f., and from the contributions of the EDSs to the hypotheses. As mentioned before, we will use the estimates of these parameters in the formulas in order to obtain the estimates of the cpp. In this section we present methods that estimate the contribution.

Throughout this subsection we assume that we have a single external data set (EDS), which we will refer to as the EDS. For estimating the contributions of the EDS to the

hypotheses we will use the statistic

$$O_{d,M} = |\{j : |t_j| \geq d, r_j \leq M\}| - \frac{M}{m^{eds}} |\{j : |t_j| \geq d\}|, \quad (7)$$

where  $|a|$  denotes the number of elements of  $a$  if  $a$  is a set, and  $|a|$  denotes the absolute value of  $a$  if  $a$  is a real, as usual,  $t_j$  is the observed statistic of hypothesis  $j$ ,  $r_j$  is the rank of hypothesis  $j$  in the external data set, and  $m^{eds}$  denotes the number of hypotheses in the EDS. In the Appendix (Theorem 18) we proved that for a positive integer  $M \leq m^{eds}$  and real number  $d \geq 0$  we have that

$$E(O_{d,M}) = (F_0(d) - F_1(d)) \sum_{j, r_j \leq M} co(j), \quad (8)$$

where  $O_{d,M}$  is defined in (7),  $F_0(d)$  and  $F_1(d)$  is the null and alternative c.d.f., respectively. First we estimate the cumulative contribution defined as  $CO(M) = \sum_{j, r_j \leq M} co(j)$ , then from the estimate of  $CO(M)$  computed for multiple values of  $M$ , we can readily obtain the contribution estimates. In order to estimate  $CO(M)$ , first we calculate the "rough" estimates  $\widetilde{CO}(M)$ , from which we will obtain the estimate  $\widehat{CO}(M)$  of the cumulative contributions. Based on (8) we calculate the "rough" cumulative estimates as

$$\widetilde{CO}(M) = \frac{1}{|D|} \sum_{d \in D} \frac{O_{d,M}}{F_0(d) - F_1(d)}, \quad (9)$$

where  $D$  is a set of positive real numbers, and  $|D|$  denotes the number of elements in  $D$ . If we have ties, i.e. hypotheses with the same rank in the EDS, then we proceed in the following way. Suppose we have  $t$  groups of hypotheses and the ranks of all hypotheses in a group are identical, but the ranks are different across the groups. Let  $R_j$  be the number of hypotheses whose rank is the  $j$ th smallest one or smaller than that for  $j = 1, \dots, t$ . We calculate  $\widetilde{CO}(M)$  for  $M = R_j$ ,  $j = 1, \dots, t$ . If we have no ties among the ranks, we may still use the same procedure with  $R_j$ ,  $j = 1, \dots, t$  that we choose rather than those dictated by the data.

Note that the two main properties of the contributions, notably that  $co(i)$  is a decreasing function of the rank of hypothesis  $i$  in the EDS (Claim 8 in Appendix) and that the sum of all contributions of an EDS is 0 (see Lemma 10 in Appendix), is equivalent to that  $M \mapsto CO(M)$  is a unimodal concave function with  $CO(m^{eds}) = 0$ . Therefore, to maintain the two main properties of the contributions, we find the curve that fits the best to  $M \mapsto \widetilde{CO}(M)$  in the family of the unimodal concave functions that take 0 at  $m^{eds}$ , or in a certain subfamily of these functions. The function obtained this way will be the cumulative contribution estimate, and it will be denoted as  $M \mapsto \widehat{CO}(M)$ . Then 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 hypothesis  $i$  in group  $j$ ,  $j = 1, \dots, t$ , where we define  $R_0 = 0$  and  $\widehat{CO}(0) = 0$  for the sake of simplicity. Then we use  $\widehat{co}(i)$  to calculate the prior probability estimates  $\widehat{\gamma}_i^*$  for

every hypothesis  $i$  by (4) or in (3) if we think  $\widehat{m}_1^*/m^{eds}$  differs from  $\widehat{m}_1/m$  significantly. We replace  $\gamma_i^{*j}$  with their estimates in (5) to obtain estimates of  $\beta_i^{(\text{combined prior})}$ , that we use in (6) with the estimates of  $f_0(t)$  and  $f_1(t)$  to obtain the estimate of cpp for every hypothesis.

## Appendix

### Mathematical formula of the compound posterior probability (cpp)

#### Step 1: Formula of the prior probability that a hypothesis is alternative based on the information in an EDS

The goal of this section is to derive the equation of prior probability (Theorem 6) and some properties of the contributions. Throughout this section we assume that we have a single external data set.

**Claim 4** *Suppose we have only one external data set. Denote the number of hypotheses in the EDS as  $m^{eds}$ . Furthermore,  $m_1^{eds}$  is the number of hypotheses that are alternative in the EDS,  $m_1^*$  is the number of alternative hypotheses that are in the EDS,  $m_1^{overlap}$  is the number of alternative hypotheses that are also alternative in the EDS, and let  $m_0^{eds} = m^{eds} - m_1^{eds}$ . Then we have that*

$$\gamma_i^* = \pi\gamma(r_i) + \nu(1 - \gamma(r_i)) = \frac{m_1^{overlap}}{m_1^{eds}}\gamma(r_i) + \frac{m_1^* - m_1^{overlap}}{m_0^{eds}}(1 - \gamma(r_i)), \quad (10)$$

where  $r_i$  is the rank of hypothesis  $i$  in the external data, and

$$\begin{aligned} \pi &\stackrel{def}{=} \Pr(H_i = 1 \mid H_i^{(EDS)} = 1) = \frac{m_1^{overlap}}{m_1^{eds}} \\ \nu &\stackrel{def}{=} \Pr(H_i = 1 \mid H_i^{(EDS)} = 0) = \frac{m_1^* - m_1^{overlap}}{m_0^{eds}}. \end{aligned}$$

**Proof.** By definition

$$\gamma_i^* \stackrel{def}{=} \Pr(H_i = 1 \mid S = s),$$

where  $H_i = 1$  or  $0$  if hypothesis  $i$  is alternative or null, respectively, and  $S = s$  represents the information we have from the external data set. Let  $H_i^{(EDS)} = 1$  or  $0$  if hypothesis  $i$  is alternative or null in the external data set, respectively. Then we have

$$\begin{aligned} \gamma_i^* &= \Pr(H_i = 1 \mid H_i^{(EDS)} = 1, S = s) \Pr(H_i^{(EDS)} = 1 \mid S = s) + \\ &\Pr(H_i = 1 \mid H_i^{(EDS)} = 0, S = s) \Pr(H_i^{(EDS)} = 0 \mid S = s) \stackrel{*}{=} \\ &\Pr(H_i = 1 \mid H_i^{(EDS)} = 1) \Pr(H_i^{(EDS)} = 1 \mid S = s) + \end{aligned}$$

$$\Pr\left(H_i = 1 \mid H_i^{(\text{EDS})} = 0\right) \Pr\left(H_i^{(\text{EDS})} = 0 \mid S = s\right) =$$

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

because  $\Pr\left(H_i = 1 \mid H_i^{(\text{EDS})} = 1\right) = \frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}$  and  $\Pr\left(H_i = 1 \mid H_i^{(\text{EDS})} = 0\right) = \frac{m_1^* - m_1^{\text{overlap}}}{m_0^{\text{eds}}}$ .

For  $\overset{*}{=}$  we used that  $\Pr\left(H_i = 1 \mid H_i^{(\text{EDS})} = j, S = s\right) = \Pr\left(H_i = 1 \mid H_i^{(\text{EDS})} = j\right)$  for  $j = 0, 1$ , which is a consequence of that the external data set is independent of the primary test statistics. ■

**Lemma 5** *For hypothesis  $i$  in the external data set we have that*

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

where  $co(i)$  is the contribution of the EDS to hypothesis  $i$  defined in (2),  $m_1^{\text{eds}}$  is the number of hypotheses that are alternative in the EDS, and  $m_1^*$  is the number of alternative hypotheses that are in the EDS.

**Proof.** We start from (10), and for brevity we will denote  $\gamma(r_i)$  as  $\gamma$  in the proof.

$$\begin{aligned} \gamma_i^* &= \frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}\gamma + \frac{m_1^* - m_1^{\text{overlap}}}{m_0^{\text{eds}}}(1 - \gamma) = \\ &= \left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}} - \frac{m_1^* - m_1^{\text{overlap}}}{m_0^{\text{eds}}}\right)\gamma + \frac{m_1^* - m_1^{\text{overlap}}}{m_0^{\text{eds}}} = \\ &= \frac{1}{m_0^{\text{eds}}}\left[\left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}m^{\text{eds}} - m_1^*\right)\gamma + m_1^* - m_1^{\text{overlap}}\right] = \\ &= \frac{1}{m_0^{\text{eds}}}\left[\left(\frac{m_1^{\text{overlap}}}{m_1}m^{\text{eds}} - m_1^*\right)\left(\gamma - \frac{m_1^{\text{eds}}}{m^{\text{eds}}}\right) + \left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}m^{\text{eds}} - m_1^*\right)\frac{m_1^{\text{eds}}}{m^{\text{eds}}} + m_1^* - m_1^{\text{overlap}}\right] = \\ &= \frac{1}{m_0^{\text{eds}}}\left[\left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}m^{\text{eds}} - m_1^*\right)\left(\gamma - \frac{m_1^{\text{eds}}}{m^{\text{eds}}}\right) + m_1^{\text{overlap}} - m_1^*\frac{m_1^{\text{eds}}}{m^{\text{eds}}} + m_1^* - m_1^{\text{overlap}}\right] = \\ &= \frac{1}{m_0^{\text{eds}}}\left[\left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}m^{\text{eds}} - m_1^*\right)\left(\gamma - \frac{m_1^{\text{eds}}}{m^{\text{eds}}}\right) + m_1^*\frac{m_0^{\text{eds}}}{m^{\text{eds}}}\right] = \\ &= \frac{1}{m_0^{\text{eds}}}\left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}}m^{\text{eds}} - m_1^*\right)\left(\gamma - \frac{m_1^{\text{eds}}}{m^{\text{eds}}}\right) + \frac{m_1^*}{m^{\text{eds}}} = \\ &= \frac{1}{m_0^{\text{eds}}}\left(\frac{m_1^{\text{overlap}}}{m_1^{\text{eds}}} - \frac{m_1^*}{m^{\text{eds}}}\right)(m^{\text{eds}}\gamma - m_1^{\text{eds}}) + \frac{m_1^*}{m^{\text{eds}}} = co(i) + \frac{m_1^*}{m^{\text{eds}}}. \end{aligned}$$

■

**Theorem 6** For hypothesis  $i$  we have that

$$\gamma_i^* = \begin{cases} co(i) + \frac{m_1^*}{m^{eds}} & \text{if hypothesis } i \text{ is in the external data set} \\ \frac{m_1 - m_1^*}{m - m^{eds}} & \text{if hypothesis } i \text{ is NOT in the external data set,} \end{cases}$$

where  $m$  is the number of hypotheses.

**Proof.** The statement of the theorem directly follows from Lemma 5 for hypotheses in the external data set. The number of hypotheses that are not in the external data set is  $m - m^{eds}$ ,  $m_1 - m_1^*$  of which is alternative. Consequently,  $\gamma_i^* = \frac{m_1 - m_1^*}{m - m^{eds}}$  for hypothesis  $i$  that is not in the external data set, because we have no prior information for this hypothesis from the external data set. ■

**Corollary 7** Under the reasonable assumption that the proportion of the alternative hypotheses is the same inside and outside the external data set, i.e.  $\frac{m_1^*}{m^{eds}} = \frac{m_1}{m}$ , we have that

$$\gamma_i^* = \begin{cases} co(i) + \frac{m_1}{m} & \text{if hypothesis } i \text{ is in the external data set} \\ \frac{m_1}{m} & \text{if hypothesis } i \text{ is NOT in the external data set.} \end{cases}$$

**Proof.** The statement of the corollary follows from Theorem 6 and from that  $\frac{m_1^*}{m^{eds}} = \frac{m_1}{m}$  implies  $\frac{m_1 - m_1^*}{m - m^{eds}} = \frac{m_1^*}{m^{eds}} = \frac{m_1}{m}$ . ■

**Claim 8** The contribution  $co(\cdot)$  is a decreasing function of the rank of hypotheses in an EDS, i.e. for hypotheses  $i$  and  $j$  in the EDS  $co(i) > co(j)$  if and only if  $r_i < r_j$ , where  $r_i$  and  $r_j$  are the ranks of hypotheses  $i$  and  $j$  in the EDS, respectively.

**Proof.** The claim readily follows from Theorem 6 as the prior probability  $\gamma_i^*$  is clearly a decreasing function of the rank of hypothesis  $i$  in the EDS, i.e. for hypotheses  $i$  and  $j$  in the EDS  $\gamma_i^* > \gamma_j^*$  if and only if  $r_i < r_j$ . ■

**Claim 9** For the ranks of the hypotheses in an EDS we have that

$$\sum_{r=1}^{m^{eds}} \gamma(r) = m_1^{eds}, \quad (11)$$

where  $\gamma(r)$  is the probability that a hypothesis ranked  $r$  in the EDS is alternative in the EDS,  $m^{eds}$  is the number of hypotheses in the EDS,  $m_1^{eds}$  is the number of hypotheses that are alternative in the EDS.

**Proof.** We define random variables  $V_r$ ,  $r = 1, \dots, m^{eds}$ , as  $V_r = 1$  if and only if hypothesis  $r$  is alternative in the EDS, and  $V_r = 0$  otherwise. Clearly,  $\sum_{r=1}^{m^{eds}} V_r = m_1^{eds}$  holds, hence

$$m_1^{eds} = E\left(\sum_{r=1}^{m^{eds}} V_r\right) = \sum_{r=1}^{m^{eds}} E(V_r) = \sum_{r=1}^{m^{eds}} \gamma(r).$$

■

**Lemma 10** *For an EDS we have that*

$$\sum_{i \in EDS^{CO}} (i) = 0,$$

where the summation is on the hypotheses that are in the EDS, and  $co(\cdot)$  is defined in (2).

**Proof.**

$$\begin{aligned} \sum_{i \in EDS^{CO}} (i) &= \sum_{i \in EDS} \frac{\kappa}{m_0^{eds}} (m^{eds} \gamma(r_i) - m_1^{eds}) = \frac{\kappa}{m_0^{eds}} \sum_{r=1}^{m^{eds}} (m^{eds} \gamma(r) - m_1^{eds}) = \\ &= \frac{\kappa}{m_0^{eds}} m^{eds} \left( \left[ \sum_{r=1}^{m^{eds}} \gamma(r) \right] - m_1^{eds} \right) = \frac{\kappa}{m_0^{eds}} m^{eds} (m_1^{eds} - m_1^{eds}) = 0, \end{aligned}$$

where we used (11). ■

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

First we derive the general formula of the combined prior probabilities (Theorem 12). As it is unknown in practice how the set of alternative hypotheses and the sets of alternative hypotheses that are alternative in the EDSs overlap, the general formula is difficult to use for practical purpose. Therefore, we derived a special case of the general formula for an overlap structure of the alternative hypotheses whose probabilities can be given as a function of two parameters (Theorem 13). This is an important special case, because it comprises two extreme scenarios, for both of which we obtain the same formula of the combined prior probabilities (Corollary 16).

**Definition 11** *Suppose we have  $k$  external data sets with (ranks of) test statistics  $S_1^i, \dots, S_m^i$ . The combined prior probability that a hypothesis is alternative based on the information in the external data sets (EDSs) is formally defined as*

$$\gamma_j^{(combined\ prior)} \stackrel{def}{=} \Pr(H_j = 1 \mid S_j^i = s_j^i, i = 1, \dots, k).$$

**Theorem 12** *Suppose we have  $k$  external data sets, and  $S_1^i, \dots, S_m^i$  are the test statistics or the ranks of the test statistics in the  $i$ th external data set,  $i = 1, \dots, k$ , where some  $S_j^i$  may be missing. Denote the number of alternative hypotheses in the  $i$ th external data set as  $m_1^i$ . Then we have that*

$$\gamma_j^{(combined\ prior)} = \frac{\sum_{(\delta_1, \dots, \delta_k) \in \{0,1\}^k} \left[ \prod_{i=1}^k \Pr(S_j^i = s_j^i \mid H_j^{(i)} = \delta_i) \right] \Pr(H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k)}{\sum_{(\tau_1, \dots, \tau_k) \in \{0,1\}^k} \left[ \prod_{i=1}^k \Pr(S_j^i = s_j^i \mid H_j^{(i)} = \tau_i) \right] \Pr(H_j^{(i)} = \tau_i, i = 1, \dots, k)} \quad (12)$$

where  $m_1$  and  $m_0$  is the number of alternative and null hypotheses, respectively, and  $\{0, 1\}^k$  denotes the set of all the 0-1 vectors of length  $k$ . Moreover,

$$\beta_j^{(\text{combined prior})} = \frac{\sum_{(\delta_1, \dots, \delta_k) \in \{0, 1\}^k} \left[ \prod_{i=1}^k \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = \delta_i \right) \right] \Pr \left( H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k \right)}{\sum_{(\tau_1, \dots, \tau_k) \in \{0, 1\}^k} \left[ \prod_{i=1}^k \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = \tau_i \right) \right] \Pr \left( H_j = 0, H_j^{(i)} = \tau_i, i = 1, \dots, k \right)}, \quad (13)$$

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

**Proof.** We have that

$$\begin{aligned} & \gamma_j^{(\text{combined prior})} \stackrel{\text{def}}{=} \Pr \left( H_j = 1 \mid S_j^i = s_j^i, i = 1, \dots, k \right)^* \\ & \sum_{(\delta_1, \dots, \delta_k) \in \{0, 1\}^k} \Pr \left( H_j = 1 \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right) \Pr \left( H_j^{(i)} = \delta_i, i = 1, \dots, k \mid S_j^i = s_j^i, i = 1, \dots, k \right) \\ & \stackrel{\text{Bayes' theo}}{=} \sum_{(\delta_1, \dots, \delta_k) \in \{0, 1\}^k} \left\{ \Pr \left( H_j = 1 \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right) \right. \\ & \left. \frac{\Pr \left( S_j^i = s_j^i, i = 1, \dots, k \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right) \Pr \left( H_j^{(i)} = \delta_i, i = 1, \dots, k \right)}{\sum_{(\tau_1, \dots, \tau_k) \in \{0, 1\}^k} \Pr \left( S_j^i = s_j^i, i = 1, \dots, k \mid H_j^{(i)} = \tau_i, i = 1, \dots, k \right) \Pr \left( H_j^{(i)} = \tau_i, i = 1, \dots, k \right)} \right\}^{**} \\ & \sum_{(\delta_1, \dots, \delta_k) \in \{0, 1\}^k} \frac{\Pr \left( S_j^i = s_j^i, i = 1, \dots, k \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right) \Pr \left( H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k \right)}{\sum_{(\tau_1, \dots, \tau_k) \in \{0, 1\}^k} \Pr \left( S_j^i = s_j^i, i = 1, \dots, k \mid H_j^{(i)} = \tau_i, i = 1, \dots, k \right) \Pr \left( H_j^{(i)} = \tau_i, i = 1, \dots, k \right)} \stackrel{***}{=} \\ & \frac{\sum_{(\delta_1, \dots, \delta_k) \in \{0, 1\}^k} \left[ \prod_{i=1}^k \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = \delta_i \right) \right] \Pr \left( H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k \right)}{\sum_{(\tau_1, \dots, \tau_k) \in \{0, 1\}^k} \left[ \prod_{i=1}^k \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = \tau_i \right) \right] \Pr \left( H_j^{(i)} = \tau_i, i = 1, \dots, k \right)}. \end{aligned}$$

- For equality  $*$  we used that the total probability theorem in the form  $\Pr(B \mid C) = \sum_i \Pr(B \mid A_i, C) \Pr(A_i \mid C)$ , where  $A_i, i = 1, 2, \dots$  is a partition of the probability space; moreover, we used that  $\Pr \left( H_j = 1 \mid H_j^{(i)} = \delta_i, S_j^i = s_j^i, i = 1, \dots, k \right) = \Pr \left( H_j = 1 \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right)$ , which follows from that the external data sets are independent of the primary test statistics.
- For equality  $**$  we used  $\Pr \left( H_j = 1 \mid H_j^{(i)} = \delta_i, i = 1, \dots, k \right) \Pr \left( H_j^{(i)} = \delta_i, i = 1, \dots, k \right) = \Pr \left( H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k \right)$ .

- For equality  $\stackrel{***}{=}$  we used that the external data sets are independent of each other.

■

As it is unknown in practice how the set of alternative hypotheses and the sets of alternative hypotheses that are alternative in the EDSs overlap, the probabilities  $\Pr\left(H_j^{(i)} = \tau_i, i = 1, \dots, k\right)$  and  $\Pr\left(H_j = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k\right)$  in (12) 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 external data sets.

**Theorem 13** *Suppose we have  $k$  external data sets, and suppose that*

$$\begin{aligned} \Pr\left(H_j^{(i)} = \delta_i, i = 1, \dots, k \mid H_j = 1\right) &= \prod_{i=1}^k \Pr\left(H_j^{(i)} = \delta_i \mid H_j = 1\right) a^{\sum_{t=1}^k (1-\delta_t)} \\ \Pr\left(H_j^{(i)} = \delta_i, i = 1, \dots, k \mid H_j = 0\right) &= \prod_{i=1}^k \Pr\left(H_j^{(i)} = \delta_i \mid H_j = 0\right) b^{\sum_{t=1}^k \delta_t}, \end{aligned} \quad (14)$$

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

$$\begin{aligned} \beta_j^{(\text{combined prior})} &= \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} a}{\gamma_j^{(i)} (1 - \pi^{(i)}) b + (1 - \gamma_j^{(i)}) (1 - \nu^{(i)})} = \\ &= \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i} - (1 - a) (1 - \gamma_j^{(i)}) \nu^{(i)}}{1 - \gamma_j^{*i} - \gamma_j^{(i)} (1 - b) (1 - \pi^{(i)})}, \end{aligned} \quad (15)$$

where  $m_1$  is the number of alternative hypotheses as usual,  $m_0 = m - m_1$ ,

$$\begin{aligned} \pi^{(i)} &= \Pr\left(H_j = 1 \mid H_j^{(i)} = 1\right) \\ \nu^{(i)} &= \Pr\left(H_j = 1 \mid H_j^{(i)} = 0\right) \end{aligned}$$

for  $i = 1, \dots, k$ ,

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

is the prior probability that hypothesis  $j$  is alternative based on the information in the  $i$ th EDS (see (10)) only, and  $\gamma_j^{(i)}$  is the probability that hypothesis  $j$  is alternative in the  $i$ th EDS, i.e.

$$\begin{aligned} \gamma_j^{(i)} &= \Pr\left(H_j^{(i)} = 1 \mid S_j^i = s_j^i\right) = \\ &= \frac{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right) \Pr\left(H_j^{(i)} = 1\right)}{\Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 0\right) \Pr\left(H_j^{(i)} = 0\right) + \Pr\left(S_j^i = s_j^i \mid H_j^{(i)} = 1\right) \Pr\left(H_j^{(i)} = 1\right)}, \end{aligned} \quad (16)$$

where  $S_1^i, \dots, S_m^i$  are the test statistic values or the ranks of the test statistic values in the  $i$ th external data set,  $i = 1, \dots, k$ .

**Proof.** Applying criterion in (14) for the formula in (13) we obtain that

$$\begin{aligned}
\beta_j^{(\text{combined prior})} &= \frac{\sum_{(\delta_1, \dots, \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 = 1, H_j^{(i)} = \delta_i, i = 1, \dots, k \right)}{\sum_{(\tau_1, \dots, \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 = 0, H_j^{(i)} = \tau_i, i = 1, \dots, k \right)} = \\
&= \frac{\sum_{(\delta_1, \dots, \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, \dots, k \mid H_j = 1 \right) \Pr(H_j = 1)}{\sum_{(\tau_1, \dots, \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, \dots, k \mid H_j = 0 \right) \Pr(H_j = 0)} = \\
&= \frac{\Pr(H_j = 1) \sum_{(\delta_1, \dots, \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 = 1 \right) a^{\sum_{t=1}^k (1-\delta_t)}}{\Pr(H_j = 0) \sum_{(\tau_1, \dots, \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 = 0 \right) b^{\sum_{t=1}^k \tau_t}} \\
&= \frac{\Pr(H_j = 1)}{\Pr(H_j = 0)} \\
&= \frac{\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 = 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 = 1 \right) a}{\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 = 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 = 0 \right)} \\
&= \left[ \frac{\Pr(H_j = 0)}{\Pr(H_j = 1)} \right]^{k-1} \\
&= \frac{\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 = 1 \right) + \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = 0 \right) \Pr \left( H_j^{(i)} = 0, H_j = 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 = 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 = 0 \right)} = \\
&= \frac{\left[ \frac{\Pr(H_j = 0)}{\Pr(H_j = 1)} \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 = 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 = 0 \mid H_j^{(i)} = 1 \right) b +} \\
&\quad + \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 = 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 = 0 \mid H_j^{(i)} = 0 \right)} = \\
&= \frac{\left[ \frac{\Pr(H_j = 0)}{\Pr(H_j = 1)} \right]^{k-1}}{\left[ \frac{\Pr(H_j = 1)}{\Pr(H_j = 0)} \right]} \\
&= \frac{\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) (1 - \pi^{(i)}) b + \Pr \left( S_j^i = s_j^i \mid H_j^{(i)} = 0 \right) \Pr \left( H_j^{(i)} = 0 \right) (1 - \nu^{(i)})}}{=} \\
&= \frac{* \left[ \frac{\Pr(H_j = 0)}{\Pr(H_j = 1)} \right]^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} a}{\gamma_j^{(i)} (1 - \pi^{(i)}) b + (1 - \gamma_j^{(i)}) (1 - \nu^{(i)})}}{=}
\end{aligned}$$

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

which proves the first equality in (15). For  $\stackrel{*}{=}$  we used (16).

Moreover, we have that

$$\begin{aligned} & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} a}{\gamma_j^{(i)} (1 - \pi^{(i)}) b + (1 - \gamma_j^{(i)}) (1 - \nu^{(i)})} = \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} a}{1 - \left\{ \gamma_j^{(i)} - \gamma_j^{(i)} b + \left( \gamma_j^{(i)} \pi^{(i)} b + (1 - \gamma_j^{(i)}) \nu^{(i)} \right) \right\}} = \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} a}{1 - \left\{ \gamma_j^{(i)} (1 - b + \pi^{(i)} b) + (1 - \gamma_j^{(i)}) \nu^{(i)} \right\}} = \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} - (1 - a) (1 - \gamma_j^{(i)}) \nu^{(i)}}{1 - \left\{ \gamma_j^{(i)} (1 - b + \pi^{(i)} b - \pi^{(i)}) + \gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} \right\}} = \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} - (1 - a) (1 - \gamma_j^{(i)}) \nu^{(i)}}{1 - \left\{ \gamma_j^{(i)} (1 - b) (1 - \pi^{(i)}) + \gamma_j^{(i)} \pi^{(i)} + (1 - \gamma_j^{(i)}) \nu^{(i)} \right\}} \stackrel{**}{=} \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i} - (1 - a) (1 - \gamma_j^{(i)}) \nu^{(i)}}{1 - \left\{ \gamma_j^{(i)} (1 - b) (1 - \pi^{(i)}) + \gamma_j^{*i} \right\}} = \\ & \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i} - (1 - a) (1 - \gamma_j^{(i)}) \nu^{(i)}}{1 - \gamma_j^{*i} - \gamma_j^{(i)} (1 - b) (1 - \pi^{(i)})}, \end{aligned}$$

which proves the second equality in (15). For  $\stackrel{**}{=}$  we used  $\gamma_j^{*i} = \pi^{(i)} \gamma_j^{(i)} + \nu^{(i)} (1 - \gamma_j^{(i)})$ . ■

**Corollary 14** *Suppose we have  $k$  external data sets. Suppose that*

$$\begin{aligned} \Pr \left( H_j^{(i)} = \delta_i, i = 1, \dots, k \mid H_j = 1 \right) &= \prod_{i=1}^k \Pr \left( H_j^{(i)} = \delta_i \mid H_j = 1 \right) \\ \Pr \left( H_j^{(i)} = \delta_i, i = 1, \dots, k \mid H_j = 0 \right) &= \prod_{i=1}^k \Pr \left( H_j^{(i)} = \delta_i \mid H_j = 0 \right), \end{aligned} \quad (17)$$

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

$$\beta_j^{(\text{combined prior})} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i}}{1 - \gamma_j^{*i}} = \frac{m_1}{m_0} \prod_{i=1}^k \frac{m_0 \gamma_j^{*i}}{m_1 (1 - \gamma_j^{*i})}, \quad (18)$$

where  $m_1$  is the number of alternative hypotheses as usual,  $m_0 = m - m_1$ , and  $\gamma_j^{*i}$  is the prior probability that hypothesis  $j$  is alternative based on the  $i$ th EDS.

**Proof.** The condition in (17) is equivalent to (14) with  $a = b = 1$ . Therefore, substituting  $a = b = 1$  in (15) we obtain that

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

■

**Corollary 15** *If*

$$H_j = 1 \Leftrightarrow H_j^{(1)} = 1 \Leftrightarrow \dots \Leftrightarrow H_j^{(k)} = 1, \quad (19)$$

then

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

where  $m_1$  is the number of alternative hypotheses as usual,  $m_0 = m - m_1$ , and  $\gamma_j^{*i}$  is the prior probability that hypothesis  $j$  is alternative based on the information in the  $i$ th EDS only.

**Proof.** First we need to see that the condition in (19) is equivalent to (14) for  $a = b = 0$ . Indeed, substituting  $a = b = 0$  in (14) we obtain that

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

and

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

As

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

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

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

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

Also the condition in (19) implies  $\pi^{(i)} = \Pr(H_j = 1 \mid H_j^{(i)} = 1) = 1$  and

$\nu^{(i)} = \Pr(H_j = 1 \mid H_j^{(i)} = 0) = 0$ . Therefore, substituting  $a = b = 0$ ,  $\pi^{(i)} = 1$  and  $\nu^{(i)} = 0$  in (15) we obtain that

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

$$\left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{(i)}}{(1 - \gamma_j^{(i)})} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i}}{1 - \gamma_j^{*i}},$$

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

**Corollary 16** *The conditions in (17) and (19) represent the two extrema of (14), and in both cases the combined odds can be calculated as*

$$\beta_j^{(\text{combined prior})} = \left(\frac{m_0}{m_1}\right)^{k-1} \prod_{i=1}^k \frac{\gamma_j^{*i}}{1 - \gamma_j^{*i}} = \frac{m_1}{m_0} \prod_{i=1}^k \frac{m_0 \gamma_j^{*i}}{m_1 (1 - \gamma_j^{*i})} = \frac{m_1}{m_0} \prod_{i=1}^k \frac{m_0}{m_1} \beta_j^{*i} \quad (20)$$

where  $m_1$  is the number of alternative hypotheses as usual,  $m_0 = m - m_1$ ,  $\gamma_j^{*i}$  is the prior probability that hypothesis  $j$  is alternative based on the information in the  $i$ th EDS only, and the odd  $\beta_j^{*i}$  is defined as  $\beta_j^{*i} = \gamma_j^{*i} / (1 - \gamma_j^{*i})$ .

**Remark 17** *The terms in the product in (15) can be approximated with  $\beta_j^{*i}$  even if (17) and (19) do not hold, suggesting that the formula in (20) is reasonable even for the general case. For the general formula (12) the structure of how the sets of hypotheses alternative in the EDSs and the set of alternative hypotheses overlap each other need to be known. This, however, can be very difficult to estimate in practice.*

## Estimating the contributions

**Theorem 18** *Denote the number of hypotheses in the external data set (EDS) as  $m^{\text{eds}}$ . For a positive integer  $M \leq m^{\text{eds}}$  and real  $d \geq 0$  we have that*

$$E(O_{d,M}) = (F_0(d) - F_1(d)) \sum_{j, r_j \leq M} co(j), \quad (21)$$

where

$$O_{d,M} = |\{j : |t_j| \geq d, r_j \leq M\}| - \frac{M}{m^{eds}} |\{j : |t_j| \geq d\}|$$

where  $|a|$  denotes the number of elements of  $a$  if  $a$  is a set, and  $|a|$  denotes the absolute value of  $a$  if  $a$  is a real, as usual,  $t_j$  is the observed statistic of hypothesis  $j$ ,  $r_j$  is the rank of hypothesis  $j$  in the EDS. .

**Proof.** For the sake of simplicity, throughout the proof we will suppress the superscript  $eds$ , i.e. the number of hypotheses (alternative) in the EDS  $m_1^{eds}$  and  $m^{eds}$  will be denoted as  $m_1$  and  $m$ , respectively. Denote the set of alternative and null hypotheses as  $N_1$  and  $N_0$ , respectively, and denote the set of hypotheses that are alternative in the EDS as  $E_1$  and the set of hypotheses that are null in the EDS as  $E_0$ . Also we define  $Q_{d,M}$  and  $m'(d)$  as

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

We have that

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

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

$$\begin{aligned} E(Q_{d,M} \cap E_0 \cap N_1) &= (m_1^* - m_1^{overlap}) (1 - F_1(d)) \sum_{i=1}^M \frac{1 - \gamma_i}{m_0} = \\ & (m_1^* - m_1^{overlap}) (1 - F_1(d)) \frac{1}{m_0} (M - \sum_{i=1}^M \gamma_i) = (m_1^* - m_1^{overlap}) (1 - F_1(d)) \frac{1}{m_0} (M - M\Gamma_1) = \\ & (m_1^* - m_1^{overlap}) (1 - F_1(d)) \frac{M}{m_0} (1 - \Gamma_1), \end{aligned}$$

$$\begin{aligned} E(Q_{d,M} \cap E_0 \cap N_0) &= (m - m_1 - m_1^* + m_1^{overlap}) (1 - F_0(d)) \sum_{i=1}^M \frac{1 - \gamma_i}{m_0} = \\ & (m - m_1 - m_1^* + m_1^{overlap}) (1 - F_0(d)) \frac{M}{m_0} (1 - \Gamma_1). \end{aligned}$$

Therefore, we have that

$$\begin{aligned} E(Q_{d,M}) &= E(Q_{d,M} \cap E_1 \cap N_1 + Q_{d,M} \cap E_1 \cap N_0 + Q_{d,M} \cap E_0 \cap N_1 + Q_{d,M} \cap E_0 \cap N_0) = \\ & m_1^{overlap} (1 - F_1(d)) \frac{M}{m_1} \Gamma_1 + (m_1 - m_1^{overlap}) (1 - F_0(d)) \frac{M}{m_1} \Gamma_1 + \\ & (m_1^* - m_1^{overlap}) (1 - F_1(d)) \frac{M}{m_0} (1 - \Gamma_1) + (m - m_1 - m_1^* + m_1^{overlap}) (1 - F_0(d)) \frac{M}{m_0} (1 - \Gamma_1). \end{aligned}$$

Moreover, we have that

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

Combining the above two we obtain

$$\begin{aligned} E(O_{d,M}) &= E\left(Q_{d,M} - \frac{M}{m}m'(d)\right) = \\ & m_1^{overlap}(1 - F_1(d))\frac{M}{m_1}\Gamma_1 + (m_1 - m_1^{overlap})(1 - F_0(d))\frac{M}{m_1}\Gamma_1 + \\ & (m_1^* - m_1^{overlap})(1 - F_1(d))\frac{M}{m_0}(1 - \Gamma_1) + (m - m_1 - m_1^* + m_1^{overlap})(1 - F_0(d))\frac{M}{m_0}(1 - \Gamma_1) - \\ & \frac{M}{m}m_1^*(1 - F_1(d)) - \frac{M}{m}(m - m_1^*)(1 - F_0(d)) = \\ & (1 - F_1(d))\left\{m_1^{overlap}\frac{M}{m_1}\Gamma_1 + (m_1^* - m_1^{overlap})\frac{M}{m_0}(1 - \Gamma_1) - \frac{M}{m}m_1^*\right\} + \\ & (1 - F_0(d))\left\{(m_1 - m_1^{overlap})\frac{M}{m_1}\Gamma_1 + (m - m_1 - m_1^* + m_1^{overlap})\frac{M}{m_0}(1 - \Gamma_1) - \frac{M}{m}(m - m_1^*)\right\} = \\ & (1 - F_1(d))M\left\{\left(m_1^{overlap}\frac{1}{m_1} - (m_1^* - m_1^{overlap})\frac{1}{m_0}\right)\Gamma_1 + (m_1^* - m_1^{overlap})\frac{1}{m_0} - \frac{1}{m}m_1^*\right\} + \\ & (1 - F_0(d))M\left\{\left[(m_1 - m_1^{overlap})\frac{1}{m_1} - (m - m_1 - m_1^* + m_1^{overlap})\frac{1}{m_0}\right]\Gamma_1 + \right. \\ & \left. (m - m_1 - m_1^* + m_1^{overlap})\frac{1}{m_0} - \frac{1}{m}(m - m_1^*)\right\} = \\ & (1 - F_1(d))M\left\{\left(m_1^{overlap}\frac{m}{m_1m_0} - \frac{m_1^*}{m_0}\right)\Gamma_1 + \frac{m_1^*m_1}{mm_0} - \frac{m_1^{overlap}}{m_0}\right\} + \\ & (1 - F_0(d))M\left\{\left[(m_1 - m_1^{overlap})\frac{m}{m_1m_0} - \frac{(m - m_1^*)}{m_0}\right]\Gamma_1 + (m - m_1^*)\frac{m_1}{mm_0} - \frac{m_1 - m_1^{overlap}}{m_0}\right\} = \\ & (1 - F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1} - m_1^*\right)\Gamma_1 + \frac{m_1^*m_1}{m} - m_1^{overlap}\right\} + \\ & (1 - F_0(d))\frac{M}{m_0}\left\{\left[(m_1 - m_1^{overlap})\frac{m}{m_1} - (m - m_1^*)\right]\Gamma_1 + (m - m_1^*)\frac{m_1}{m} - m_1 + m_1^{overlap}\right\} = \\ & (1 - F_1(d))\frac{M}{m_0}\left\{\left(m_1^{overlap}\frac{m}{m_1} - m_1^*\right)\Gamma_1 + \frac{m_1^*m_1}{m} - m_1^{overlap}\right\} + \end{aligned}$$

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

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

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

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

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

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

$$(F_0(d) - F_1(d)) (m M \Gamma_1 - M m_1) \left[ \frac{1}{m_0} \left( \frac{m_1^{\text{overlap}}}{m_1} - \frac{m_1^*}{m} \right) \right] = \left\{ \Gamma_1 = \frac{1}{M} \sum_{j=1}^M \gamma(j) \right\}$$

$$(F_0(d) - F_1(d)) \sum_{j=1}^M (m \gamma(j) - m_1) \left[ \frac{1}{m_0} \left( \frac{m_1^{\text{overlap}}}{m_1} - \frac{m_1^*}{m} \right) \right] = (F_0(d) - F_1(d)) \sum_{j=1}^M co(j),$$

which concludes the proof of the theorem. ■