Publication Bias in Methodological Computational Research

Anne-Laure Boulesteix[1], Veronika Stierle[1] and Alexander Hapfelmeier[2]

[1]Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilian University, Munich, Germany. [2]Department of Medical Statistics and Epidemiology, Klinikum rechts der Isar Technical University of Munich, Munich, Germany.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: The problem of publication bias has long been discussed in research fields such as medicine. There is a consensus that publication bias is a reality and that solutions should be found to reduce it. In methodological computational research, including cancer informatics, publication bias may also be at work. The publication of negative research findings is certainly also a relevant issue, but has attracted very little attention to date. The present paper aims at providing a new formal framework to describe the notion of publication bias in the context of methodological computational research, facilitate and stimulate discussions on this topic, and increase awareness in the scientific community. We report an exemplary pilot study that aims at gaining experiences with the collection and analysis of information on unpublished research efforts with respect to publication bias, and we outline the encountered problems. Based on these experiences, we try to formalize the notion of publication bias.

KEYWORDS: epistemology, publication practice, false research findings, overoptimism

SUPPLEMENT: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

CITATION: Boulesteix et al. Publication Bias in Methodological Computational Research. *Cancer Informatics* 2015:14(S5) 11–19 doi: 10.4137/CIN.S30747.

TYPE: Methodology

RECEIVED: June 17, 2015. **RESUBMITTED:** August 04, 2015. **ACCEPTED FOR PUBLICATION:** August 12, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 855 words, excluding any confidential comments to the academic editor.

FUNDING: ALB is supported by the German Research Foundation (project BO3139/2). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: boulesteix@ibe.med.uni-muenchen.de

Introduction

The concept of "publication bias" is well known in various scientific areas, in particular medical research and social sciences, see the study by Sterling[1] for an early reference and the study by Easterbrook et al.[2] for an empirical assessment of the publication bias for a cohort of medical research projects conducted at Oxford University. The publication bias is defined by Dickersin as the "tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings."[3] Different, mutually interacting problems may favor publication bias. The expected higher impact of studies with significant results leads journals to adopt an editorial policy favoring such studies, while Goldacre[4] argues, on the basis of several investigations,[5–8] that there is no distinct evidence of this higher impact. Considering these editorial policies and the expected poorer impact of their research, or simply because they are disappointed that their hypothesis is not confirmed by the conducted study, authors may be reluctant to invest time in the publication of negative results.[4,9–11] Medical statisticians often see their clinical partners' motivation for a project suddenly drop after they show them a few insignificant *P*-values – even though many of these partners are probably uncertain about the correct interpretation of *P*-values.[12] While it is unclear to what extent this motivation drop is a consequence or one of the causes of the publication bias, it is indisputable that the publication bias is a relevant issue affecting the interpretation of published research results.

Recently, much attention has also been devoted to the problem of false research findings in medical literature, for example, in the deliberately provocative pioneering essay by Ioannidis entitled "Why most published research findings are false."[13] False research findings are most often false positive findings, ie, findings indicating, say, relevant effects of therapies or relevant association between risk factors and outcomes, whereas, in reality, there are no such effects or associations. Such issues are also addressed in the recent series, "Increasing value, reducing waste" recently published in *The Lancet*[14] and in Ioannidis's new essay on "How to make more published research true."[15] The publication bias is likely to increase the proportion of false positive research findings within published results – not only by keeping away true negative results from publication but also, more subtly and more importantly, by virtually forcing authors into data dredging and fishing for significance, thus making true negative results appear as (false) positive results in publications.

The publication bias in biomedical research has been widely investigated and is well known to all scientists – including statisticians – working in this field. Statistical methods have been proposed to detect it and to correct for its impact in meta-analyses; see, for instance, the studies by Sutton et al.[16] and Jin et al.[17] for reviews. In contrast, the publication bias in biostatistics/medical statistics methodological research or, more generally, in data analysis sciences (including statistics, machine learning, and cancer informatics) has been widely ignored by the community so far. By methodological computational research, we mean research that aims at developing new data analysis methods or algorithms. These algorithms are intended to produce results that are in some sense closer to the truth than results of currently used algorithms or, more generally, to have some other advantage over existing algorithms, such as computational efficiency or better interpretability. The interest is not in the results of the algorithms for a particular cancer dataset but in the general performance of the algorithm across datasets. Note that methodological computational journals may also include, for example, comparison studies or papers presenting properties of existing algorithms. In our paper, we deliberately ignore such studies and focus on studies that introduce new algorithms.

The publication bias has been widely discussed in the context of medical research. Specialized journals, such as the *Journal of Negative Results in Biomedicine* or *Journal of Pharmaceutical Negative Results*, explicitly welcome the publication of negative research findings; however, these problems have, to our knowledge, never been addressed in the context of methodological computational research. Methodological journals require superiority of the new algorithm as a perquisite to publication, which implies some sort of publication bias in and of itself. Our aim is to provide the first definitions and discussions of the publication bias in this context.

To illustrate our point more precisely, let us imagine that 10 research teams in the world have similar ideas to develop a new algorithm addressing an interesting challenging research question, for example, supervised classification for predicting response to therapy of cancer patients based on a particular type of complex data. Eight teams obtain disappointing results when implementing their idea in practice, ie, the new classification algorithm does not perform well. They give up the idea and do not report their failed attempts. Two teams obtain satisfying results and publish them. One may argue that if these teams have found out clever tricks to make the new idea – which was disappointing to the other teams – work fine, it is okay that they publish their results and it is good that the other teams do not overcrowd scientific literature by reporting their failed attempts. After all, a failed attempt is not proof that the idea will never have a chance to work. In a similar way, insignificant P-values should never be used to disprove a research question. The "absence of evidence is not evidence of absence."[18] Roughly speaking, this argumentation suggests that there is no point in investigating the issue of publication bias in the context of methodological computational research.

Taking the opposite view, we argue that the scenario with the 10 teams sketched above is not satisfactory. The two apparently successful teams may not be as successful as thought at first glance. For example, the apparently successful teams may have obtained these good results in very specific settings, for example, after preparing the datasets in an unusual way. They might have been only partially successful (eg, for datasets of a particular type), but formulate their papers in a way that they give hope to readers. They may also have consciously or subconsciously "fished for significance" to obtain these good results (for instance "fished for datasets" yielding advantageous results for the new algorithms), see our previous work[19] on how this can be done. Likewise, overfitting of algorithms could have led to overly optimistic results. Last but not the least, the good result may be a "false positive," ie, from the perspective of statistical testing, a type I error.[1] Moreover, it may be a problem that the research activities related to the eight unsuccessful attempts remain unpublished. A possible consequence is that further teams, not knowing about them, may start investing time and funds in the same dead-end ideas with little chance of success. This is particularly true in the even more extreme scenario where none of the teams are successful, and all research efforts related to the considered research question remain unpublished. Moreover, the new algorithm may start establishing itself as a standard well-performing algorithm based on the positive results of the two teams. Such a trend may be difficult to reverse in the future.

Clearly, there is something like a publication bias at work in methodological computational research in the sense that the literature publishes only (or mostly) successful attempts and that ideas that turn out to not work well remain unpublished. But this topic has, to our knowledge, never been discussed in the literature so far. The present paper aims at filling this gap. Note that the concept of "publication bias" considered in our paper has to be contrasted from the general meaning of this term in clinical or epidemiological meta-analyses, where it refers to actual bias in the estimation of effect measures. In contrast, our paper uses it to denote the distortion between results obtained by research teams and results ultimately published, but does not refer to an effect in the classical sense.

The present paper aims at providing the first formal framework to describe the notion of publication bias in this context, facilitate and stimulate discussions on this topic, and increase awareness in the scientific community. It is structured as follows: in the Exemplary Pilot Study section, we report an exemplary pilot study whose aim was at gaining first experiences with the collection and analysis of information on unpublished research efforts with respect to publication bias and outlining the encountered problems. Based on these experiences, we try to formalize the notion of publication bias in the context of methodological computational research in the Formal Framework for Defining the Publication Bias section. The Conclusion and Future Work section outlines limitations of the framework and directions for future discussions.

Exemplary Pilot Study

Historical background and investigated topic. In one of our areas of expertise, namely, the random forest (RF) algorithm[20] for supervised classification and regression – now widely used in genetic and cancer research – we identified a topic that was possibly subject to publication bias: the identification of pairs of variables with interaction effects based on the RF output. The terminology commonly used in literature on RF is somewhat confusing, mixing up different notions.[21] What we mean by "identification of interacting variables" here is neither the good accuracy of RF in the case of data with interacting variables nor the use of interaction effects to improve prediction accuracy of prediction rules, but rather the identification of the pairs of variables, yielding conclusions like "variable X3 and variable X9 have an interaction effect," whereby the term interaction effect may be defined in various ways, for instance, in terms of deviation from the linear additive model within the linear regression framework.[21] Another example would be in the context of prediction of response to therapy that there may be an interaction between a genetic marker and treatment in the sense that the benefit of a new treatment may be more pronounced for patients with a particular genetic pattern.

Through informal discussions with other experts, we felt that the identification of interacting variables was often mentioned in connection with RF, suggesting that many researchers conducted (preliminary) studies on this topic. In the literature, we found a moderate number of articles showing evidence of the ability of RF to reliably recover pairs of interacting variables, in contrast to the high number of articles mentioning the connection between RF and interactions in some way.

Let us simplistically label studies as "published" or "unpublished" ($P = 1$ or $P = 0$) and "successful" or "unsuccessful" ($S = 1$ or $S = 0$). Here, $S = 1$ means that the authors found that the new algorithm performs well (most often, well means better than other algorithms). Note that S does not necessarily reflect whether the new algorithm is truly better than existing algorithms or not. In other words, we do not always have $B = S$, where $B = 1$ for a new algorithm that is better than existing algorithms and $B = 0$ otherwise. The term "false positive result," commonly used in the literature, refers to a study with $S = 1$ and $B = 0$. We now forget B until the end of the Formal Framework for Defining the Publication Bias section and focus on S and P. All notations (S, P, B, and others) are summarized in Table 1.

Let us consider the corresponding contingency table for P and S displayed as Table 2. On the other hand, our informal discussions with colleagues and the fact that RFs are often mentioned in connection with interactions suggests that many studies are run underground, ie, $\Pr(P = 0)$ is large. Successful studies are often ultimately published, meaning that $\Pr(P = 0, S = 1)$ is rather small, while $\Pr(P = 0, S = 0)$ is rather large. As far as published studies are concerned, in most fields, we probably have $\Pr(P = 1, S = 1) > \Pr(P = 1, S = 0)$: reporting of unsuccessful studies (ie, new algorithms with disappointing performance) is very uncommon in methodological computational literature, see the study by Boulesteix et al.[22] for a survey on this topic. Altogether, there is obviously an association between these two binary variables or, in other words, the odds ratio

$$\theta = (\Pr(P = 1, S = 1) \times \Pr(P = 0, S = 0))/(\Pr(P = 1, S = 0) \times \Pr(P = 0, S = 1))$$

is larger than 1. This formulation of the problem suggests a naïve definition of the publication bias. We will see, however, that (i) the underlying concepts "published" and "successful" are ambiguous (Interpretation of the Naive Definition in the Context of Methodological Research section) and (ii) it makes little sense to simply define the publication bias as $\theta > 1$, because the problem is in fact more elaborate (Further Aspects to be Taken into Account when Defining Publication Bias section).

Motivation and design of the pilot study. In an attempt to gain first experiences around the topic of publication bias in methodological computational research beyond these

Table 1. Definitions.

	DEFINITION	TYPE
$P = 1$ $P = 0$	Study is published Study is not published	Observed
$S = 1$ $S = 0$	Study is successful (ie, suggests that new method is better) Study is not successful	Observed, potentially subject to interpretation problems
$B = 1$ $B = 0$	New method is truly better New method is not truly better	Unobserved, what everyone want to know
$M = 1$ $M = 0$	New method makes sense New method does not make sense	Unobserved, varies in time, subjective
$D = 1$ D = 0	Study is well-designed Study is not well-designed	Unobserved, varies in time, subjective

Table 2. Contigency table for the naïve definition of publication bias.

	Unpublished ($P = 0$)	**Published ($P = 1$)**	
Unsuccessful ($S = 0$)	Pr($P = 0$, $S = 0$)	Pr($P = 1$, $S = 0$)	Pr($S = 1$)
Successful ($S = 1$)	Pr($P = 0$, $S = 1$)	Pr($P = 1$, $S = 1$)	Pr($S = 1$)
	Pr($P = 0$)	Pr($P = 1$)	

subjective considerations, we conducted a pilot study focused on the specific topic "identification of pairs of interacting variables based on the output of RF." The goal of the pilot study was to obtain information on both published and unpublished research efforts and to examine the results in an explorative way with respect to publication bias mechanisms. Here, we deliberately focus our brief report on the methodological aspects of the pilot study in the perspective of publication bias rather than on technical aspects of RF.

Our pilot study consisted of two distinct parts: (A) a thorough general literature search on RFs in relation to interactions between predictor variables in order to identify published studies on this topic and (B) the collection of information on unpublished studies on this topic by directly contacting researchers by email who might have investigated this research question – without publishing their results. Part A was relatively standard from a methodological point of view (such literature searches are routinely performed as a preliminary step of any research project, using search engines, databases, and reference lists of already identified articles), while Part B was a challenge.

For Part B, we defined target groups of researchers having possibly performed attempts to identify interacting variables using RF methodology: the corresponding authors of papers including "random forest" in the title published in nine selected methodological computational journals, the corresponding authors of the 10 most cited papers (according to Web of Science) with "random forest" in their title, the corresponding authors of papers identified in our literature search in Part A, and researchers known by us as having thought of the considered research question. After eliminating duplicates from this list of researchers, we sent them a questionnaire by email containing questions on their attempts to develop an algorithm for identifying interacting variables from RFs, how successful they were, whether they gave up, and if yes, then when, whether they believe that success will ever be possible, and whether they communicated their results in any form to the outside world.

Main results of the pilot study. *Part A: Literature search to identify published studies.* We identified different types of papers on RF mentioning interactions between predictors: papers describing successful algorithms that can extract pairs of interacting variables from the output of RF, papers describing two-stage algorithms (RF applied in the first stage to select promising candidate variables, then another algorithm applied in the second stage to identify interactions between the filtered variables), papers demonstrating the good prediction performance of RF in the presence of interactions, papers simply mentioning RF in connection with interactions without showing any own analysis, and papers dealing primarily with other algorithms but including negative or reserved statements on the ability of RF to identify pairs of interacting variables. Interestingly, we also found one paper that reported its main result as "RF variable importance measures fail to detect interaction effects in high-dimensional data in the absence of a strong marginal component."[23] Such negative results are rather uncommon in the literature.

It would go beyond the scope of this paper to describe details of the identified papers. Instead, we report in the Identified Difficulties section important difficulties highlighted directly or indirectly by the results of our search and related to the publication bias. Some of these problems make it difficult to define the publication bias at all, while other problems make it difficult to assess the publication bias in practice (whatever definition is adopted).

Part B: Email contact to identify unpublished studies. We contacted 67 researchers and obtained 28 responses, 24 of them with answers to at least one of our questions. Many of them had interesting ideas on the topic "RF and interactions," what was more or less successful, and whether they published or not. We will not go into detail here, both because it is not the subject of the present paper and because these ideas were presented in private communication only.

Identified difficulties. The results of Parts A and B suggested directly or indirectly that (i) the naïve definition given in the Motivation and Design of the Pilot Study section involves several notions that are themselves not clearly defined, (ii) other aspects of the considered studies (beyond the aspects "published or not" and "successful or not" considered in the naïve definition) should be taken into account when defining publication bias, and (iii) it is challenging to assess the publication bias in practice – whatever definition is considered. These three aspects are treated in the Interpretation of the Naive Definition in the Context of Methodological Research section, the Further Aspects to be Taken into Account when Defining Publication Bias section, and the Difficulties Related to the Assessment of Publication Bias section, respectively. Whenever appropriate, we draw parallels to the issue of publication bias in biomedical research, in particular in the context of meta-analysis.

Interpretation of the naive definition in the context of methodological research. It is difficult to summarize the results of a study in the form of one or a few summary indices: While in biomedical research this is typically done by considering objective measures such as the mean outcome difference between two treatment groups (in a clinical trial) or the risk ratio between two exposure groups (in an epidemiological study), in methodological research it is extremely difficult, or often even does not make sense, to synthesize the performance of a new algorithm, which essentially includes several features, through a low-/one-dimensional representation. For instance, in our example, an algorithm may be able to, say, detect interacting variables better if they are binary than if they are continuous. As far as supervised learning algorithms are concerned, accuracy may be measured in different ways (eg, error rate, area under the curve, Matthews correlation coefficient, etc) and other criteria, such as robustness, computational efficiency, or sensitivity to tuning parameters, can be relevant as well.

It is difficult to define a "success": In the simplistic view given in Table 2, we have implicitly assumed that a study can always be classified as "successful" or "unsuccessful" or, in other words (which will be used in the Formal Framework for Defining the Publication Bias section), that it is clear whether it is better than existing algorithms or not. The definition of the words "better" and "success", however, is delicate in the context of methodological computational research. First, success essentially has several dimensions as outlined in the earlier paragraph. Second, even if we had a single summary index, it would probably not be binary but rather measured on an ordinal or continuous scale, just as the treatment effect in clinical trials or the odds ratio in epidemiological studies. In this context, the "success" of a study can be subject to controversy. Moreover, we see at least two additional issues: (1) researchers might declare their research outcome a "success" by setting a low threshold on this ordinal or continuous scale and (2) the success may have been generated artificially, either wittingly or unwittingly, for example, by selecting the settings, the datasets, or the competing algorithms that make their algorithm look best. See our previous studies for discussions and illustrations of these issues.[19,22] In this manner, a false positive finding may be presented as a "successful" attempt, while "correct negative" findings are discarded as "unsuccessful." In a nutshell, grouping studies into "successful" and "unsuccessful" studies is obviously too simplistic.

It is difficult to define a publication: The definition of "publication" is also ambiguous. What is recognized as publication? A paper published in a journal indexed in Web of Science (or any other prespecified database)? A paper published in a journal or a proceedings volume indexed in Web of Science? A paper published in any referred journal (or referred proceedings volume)? A paper published in any journal or proceedings volume (no matter whether referred or not)? A paper made publicly available in any way, through publication as aforementioned or also simply available from a public repository or from the authors' webpage? A software package available from a public repository? While software repositories and publication on the authors' webpage are not used to disseminate biomedical research results, they play a non-negligible role in methodological computational research. While publications in conference proceedings are not recognized as full-fledged publications in many fields such as statistics, they are of major importance in informatics. There are several ways to disseminate methodological results – beyond traditional journal publication, so "publication" is not binary. Beside the type of publication, the topic of the publication is an important aspect. For example, let us consider a study on "RF and interactions" having yielded negative results. Do we consider it as published if these results are included in a compact form as a small part of a paper presenting an extensive comparison study of existing algorithms? Or should the article be devoted mainly to the considered study in order to be counted as publication? In summary, the term "publication" has to be defined precisely and several options are conceivable.

Further aspects to be taken into account when defining publication bias. The studies are heterogeneous with respect to the quality of the idea: Some of the ideas underlying the proposed algorithms appeared particularly clever, while others did not convince us as much. Of course this aspect is highly subjective. To better outline this aspect, let us consider a virtual example of an idea that would be of particularly low quality. Suppose someone suggests declaring the pairs as "interacting" if they were formed by variables having consecutive ranks in the list of variables ranked according to their univariate variable importance measure, it would make very little sense. There is no reason why this algorithm should allow for the identifying of pairs of interacting variables! In contrast, algorithms based on the joint occurrence of variables in the branches of trees appear more promising. This example is of course exaggerated, but in practice, the algorithms investigated by researchers in their studies may differ in their quality in this sense.

The studies are heterogeneous with respect to soundness: For example, in some studies, it was claimed that the proposed algorithm can extract pairs of interacting variables, but there was no simulation examining whether the algorithm can differentiate between variables having an interaction effect and variables both having main effects but no interaction effect.[21] In the same vein, one may imagine that some studies suffer from programing errors (although we found no hint for this in our pilot study). These problems may be paralleled to problems potentially affecting biomedical studies, such as inadequate data management or errors in the measurement of the outcome.

Difficulties related to the assessment of publication bias. Suppose that we have somehow defined the publication bias, for example, through the naïve definition considered in the Motivation and Design of the Pilot Study section or through the more elaborate definition that will be proposed in the

Formal Framework for Defining the Publication Bias section. Our pilot study suggests that substantial difficulties will make it challenging to assess the extent of publication bias in practice.

It is difficult to define the research question or research area of interest: For defining the publication bias, we do not need to refer to a particular research question or research area. When assessing publication bias in practice, however, one concentrates on a specific area or question, which can be difficult to define. This issue, which can be related to the eligibility criteria for biomedical studies to be included in a meta-analysis, is very problematic for methodological computational research since research questions are usually less clearly delimited.

Sometimes the studies are not independent of each other: Even if we had clearly defined a research area or research question of interest, we would have the problem that, within this area/question, the studies might be strongly related to each other. For example, two papers may be authored by partly overlapping groups of authors and handle different variants of an algorithm. Or authors may first briefly introduce an idea in a paper and systematically investigate its performance in another one. Note that overlapping studies also occasionally occur in the biomedical context, but in a completely different sense, hence calling for different solutions, which cannot be of any use here.

It is difficult to find all unpublished studies: While a careful literature search allows for the identification of most relevant publications in general, it is much more difficult to identify unpublished studies, which are by definition not disseminated publicly via usual information channels. Asking "all researchers in the world" is simply impossible from a practical point of view. Thus, feasible strategies have to be adopted to reach "as many target researchers" as possible (by target researcher we mean a researcher having performed an unpublished study on the topic of interest). In our pilot study, we decided to primarily contact researchers published on RF in indexed journals in recent years. This is obviously a restriction. To keep the study manageable, and because the addresses of coauthors are not always mentioned on published papers, we also restricted ourselves to the corresponding authors of these papers, thereby potentially missing some target researchers. Finally, even if one were able to identify and contact all target researchers, the low response rate would still be an issue. Some of the reasons being: the high workload of academic researchers, the amount of spam emails they receive daily, and the lack of incentive for participation in the survey. We suspect, without being able to provide evidence, that the response rate may be higher for the target researchers than for researchers who have not performed any unpublished study on the topic. A counterargument against this conjecture is that researchers currently working on such a project might be unwilling to reveal hot information on their study before publication. No matter how these factors affect the response rate, the response rate is very unlikely to equal 1.

Soundness, quality of the idea, success, and independence of the studies are most often difficult to assess: Even if the notions of "soundness," "quality of the idea," "success," and "independence of the studies" are carefully defined, and even if the considered research area/question is clearly stated, it is in many cases difficult to assess them in concrete studies and this assessment is obviously highly subjective. This problem particularly affects unpublished studies. Whatever communication channel (email, phone calls, meetings in person, etc) is used to collect information on unpublished studies, less details will be available than for published studies, thus complicating the assessment. For example, since for unpublished studies there is no official list of authors, it may be difficult to determine whether a study is connected to another study from a cooperation partner or not.

Unpublished studies may have been stopped prematurely: As far as unpublished studies are concerned, the definition of the term "study" is unclear. The researchers may have stopped their study prematurely after it yielded a few disappointing results, so the results of the study are not completely available. Should such a stopped study be considered as a study and how should we handle the missing results?

Unpublished studies may ultimately get published: When defining the publication bias, it is acceptable to consider "publication" as a binary event, implying that the time point of publication is unimportant. However, the fact that completed studies are usually not published immediately also complicates the assessment of publication bias. Should a study, which is under consideration for publication, be considered as published? After how much time should a study be considered as "unpublished"? In a way, this problem is related to the question of whether it makes sense to dichotomize a censored time to event, with publication being considered as the event of interest. In methodological computational science where the time between article submission and final publication are very long (often several years!), this dichotomization may be a problem, especially in rapidly evolving fields like cancer informatics where the most relevant studies have been conducted in the last few years.

We conclude from all the problems outlined in this section that – whatever definition of the publication bias is adopted – several challenges have to be addressed before the publication can be assessed in practice.

Formal Framework for Defining the Publication Bias

Naïve definition. Following the naïve definition, we consider the two random variables P, taking values $P = 0$ (unpublished study) or $P = 1$ (published study) and S, which reflects the apparent success of the study, ie, whether the new algorithm performs well as suggested by the results of the study (note that this does not necessarily correspond to the truth reflected by variable B). Without going into the details of probability theory, let us say that the set Ω of all possible outcomes ω is here the set of all possible studies, and that the random variables P and S are functions from Ω to $\{0,1\}$.

As discussed in the Interpretation of the Naive Definition in the Context of Methodological Research section, P should be defined carefully, and different definitions are conceivable. Similarly, S should also be defined carefully as it may be problematic to summarize the "good performance of an algorithm" in the form of a single variable. In the present definitions of publication bias, we ignore these problems by simply assuming that sensible definitions of "publication" and "success" have been adopted in the considered context. For the sake of simplicity, we will also assume that S is binary, ie, a study is either successful or unsuccessful. For example, one might think of S as the output of a statistical test comparing the performances of the new algorithm and its competitors. $S = 1$ means that the null hypothesis that their performances are equal can be rejected in favor of the alternative hypothesis that new algorithm performs better.

The naïve definition from the Motivation and Design of the Pilot Study section says that there is a publication bias if $\theta > 1$, ie, if

$$\Pr(P = 1|S = 1) > \Pr(P = 1|S = 0).$$

Refined definition. Obviously, this definition is too simplistic, because it does not take into account whether the considered studies investigate algorithms that make sense and whether the evaluation of the algorithms is based on a sound design (including aspects such as adequacy of simulation design, choice of example datasets, and correctness of computer programs). We define the random variable M as $M = 1$ if the investigated (new) algorithm appears to make sense (before the study is run) and $M = 0$ otherwise. At this stage, two important remarks can be made. First, whether an algorithm makes sense or not is obviously highly subjective. Hence, it is helpful to think of M as an unobserved latent class. We can merely observe the (differing) opinions of some individual researchers. Second, M essentially varies in time as the state-of-the-art evolves. In our definition, we consider the time of publication as the reference. For example, in the case of RF and interactions considered in the Exemplary Pilot Study section, M would equal 0 for a study on the algorithm consisting to declare the pairs as "interacting" if they were formed by variables having consecutive ranks in the list of variables ranked according to their univariate variable importance measure.

Similarly, we define the random variable D as $D = 1$ if the design for the evaluation of the algorithms is sound and $D = 0$ otherwise. Similar to M, D can be viewed as a latent class and is subject to variation in time since, for example, knowledge on study designs makes progress and computing performance improves.

The random variables P, S, D, and M share the joint probability $\Pr(P, S, D, M)$. The relation of S, M, and D to P is best described by the conditional probability $\Pr(\mathrm{P}|S, M, D)$, which will be considered into our refined definition of the publication bias.

Obviously, studies with unsound design ($D = 0$) and/or on nonsensible algorithms ($M = 0$) are less publication worthy than sound studies. The naïve definition from the Naïve Definition section is too simplistic because it does not take these aspects into account. Instead, we suggest defining the publication bias as follows.

There is no publication bias if

$$\Pr(P|S, M, D) = \Pr(P|M, D),$$

meaning that the probability of a manuscript to be published only depends on the soundness of the design and whether the new algorithm makes sense but not on its success. In contrast, if

$$\Pr(P|S = 1, M, D) > \Pr(P|S = 0, M, D),$$

publication bias is at work.

Toward the definition of an ideal editorial strategy? In an ideal world, we would wish editorial strategies and review processes to be such that:

- $\Pr(P = 1|M = 0, D = 0)$ equals 0 since studies on nonsensible algorithms with unsound design should not be published.
- $\Pr(P = 1|M = 1, D = 1)$ equals 1, no matter whether $S = 0$ or $S = 1$.
- $\Pr(P = 1|M = 1, D = 0)$ is low; scientific publications must be sound. One might, however, argue that a not completely sound study on a pioneering promising idea might be publication worthy, for instance in the hope that this pioneering idea will be further investigated by other teams in a sounder way. This is a delicate question and the answer strongly depends on the exact definition of soundness considered. But, in general, unsound studies are certainly not publication worthy even if the underlying idea is sensible. In such a case, editors and reviewers may also suggest improvements of the study design during the review process to turn D from 0 to 1.
- $\Pr(P = 1|S = 0, M = 0, D = 1)$ equals 0, since an unsuccessful study on a nonsensible algorithm is not publication worthy. One could, on the contrary, argue that any sound study is publication worthy. We take the view, however, that a study on a nonsensible algorithm, which turns out to be unsuccessful, is not publication worthy. It is just the confirmation of what everyone suspected and publication would be useless.

As far as $\Pr(P = 1|S = 1, M = 0, D = 1)$ is concerned, one might argue that it may make sense to publish a soundly designed study on an algorithm that appeared nonsensible before the study was run but which yielded successful results. Indeed, important scientific discoveries are sometimes "good surprises" in the sense that nobody would have expected the idea to work. Giving a chance to algorithms, which a priori make poor sense but turn out to yield good results, is important to avoid conformism and to promote nonmainstream ideas. Note that it would imply a publication bias according to our definition, since we would then have $\Pr(P = 1|S = 1, M = 0, D = 1) > \Pr(P = 1|S = 0, M = 0, D = 1)$.

To better outline this point, it is conceptually helpful to think again of the variable B introduced in the Exemplary Pilot Study section (whether the new algorithm is truly better or not). B is essentially unknown. The association between B and M reflects whether the truth corresponds to what researchers expect before running their studies. It may be good that $\Pr(P = 1|S = 1, M = 0, D = 1)$ is not 0 to account for the fact that B and M are not perfectly equal.

Important remarks on the joint distribution *Pr(B, S, D)*.

- In medical research, some voices suggested running the review process based on the description of the data and methods only, ie, without seeing the results, in order to avoid publication bias. Would this principle make sense in methodological computational research? The task of referees would be to assess M and D, but S would not play any role.
- In our nonideal world, the variables S and D are probably negatively associated, ie,

$$\Pr(S = 1|D = 0) > \Pr(S = 1|D = 1),$$

which corresponds to fishing for significance or overoptimism due to overfitting. If the design is flawed, for instance, if the researcher reports only the results for the datasets that make the new algorithm look best or omits the best competitor, then the study is more likely to be apparently successful. Considering S as the main criterion for publication would mean that we indirectly favor studies with unsound design.

- For a soundly designed study ($D = 1$), the discrepancy between S and B can be interpreted in terms of type I and type II errors: $\Pr(S = 1|B = 0)$ is a false positive result to be paralleled to type I error and $\Pr(S = 0|B = 1)$ is a false negative result to be paralleled to type II error. To better outline this idea, let us consider the following simplified example. In a comparison study based on a number of real datasets from a repository, such as GEO, ArrayExpress, or The Cancer Genome Atlas, the authors compute a goodness criterion for the considered algorithm and for a competing algorithm. For example, the considered goodness criterion may be a fivefold cross-validation error in the case of supervised classification methods. They subsequently perform a paired t-test to compare the goodness criterion for the two algorithms, considering the datasets as units.[24] Doing so, they can commit a type I error or a type II error. In this context, it is worth noting that considering a large number of datasets decreases the type II error.[24,25]
- The discrepancy between S and B is expected to be larger for unsound studies ($D = 0$). We expect $\Pr(S = 1|B = 0, D = 0) > \Pr(S = 1|B = 0, D = 1)$ for different reasons. For example, more false positives may be obtained through fishing for significance, which is a reason why a study may be unsound. In contrast, fishing for significance is not so useful for algorithms that are truly better than existing algorithms, and unsound studies may fail to detect the superiority of a new algorithm due to, say, inadequate design or programing errors, so we also expect $\Pr(S = 0|B = 1, D = 0) > \Pr(S = 0|B = 1, D = 1)$. For example, the discrepancy between B and S increases if a suboptimal goodness criterion (such as the highly variable leave-one-out cross-validation error in the context of supervised classification) is used, or if a simulation study does not include enough iterations to achieve reliable results. See our previously published checklist[25] for more advice on the design of a sound study.

Review processes and editorial policies. Most methodological computational journals tend to consider the (unobserved) variables B and D as unique publication criteria. Roughly speaking, journals expect referees to make a guess on the unobserved variable B based on the apparent success S and their assessment of the study design D. Based on several referee reports, they hope to be able to reliably determine whether the study satisfies $B = 1$ and $D = 1$.

We claim that it might be better to consider M and D as publication criteria (and perhaps, in the case where $M = 0$ and $D = 1$, also B as reflected by S). Note that, in any case, publication criteria are defined based on unobserved variables, thus implying much subjectivity and inter-rater variability. Our criteria would have the advantage of not driving authors into fishing for significance and overoptimistic reporting,[25] thus improving the probability $\Pr(B = S)$ of agreement between the unobserved variable B and the observed variable S.

Furthermore, we would like to stress the importance of articles devoted to "neutral comparisons," ie, articles that do not present any new algorithm but focus on the comparison of existing algorithms, see our previous study[22] for definitions and discussions. Such studies may be useful to provide unbiased (or less biased) assessments of published algorithms, and thus to correct for the publication bias and bias due to fishing for significance a posteriori through independent testing by other teams.

Conclusion and Future Work

In summary, we have introduced a formal framework with the aim to define the notion of publication bias in the context of methodological research findings. In our definition, there is a publication bias when a well-designed study on a sensible algorithm has a greater chance to get published if it reports successful results regarding the new algorithm. This definition, however, is a first proposal and should be subject to discussion in the community. Moreover, it raises several important questions. Is publication bias always undesirable for the progress of scientific knowledge? In the cases where publication bias is detrimental, how can it be reduced? More generally, which criteria are appropriate for deciding whether a study is publication worthy or not? How can these criteria be realistically implemented in practice without overcrowding the literature with futile studies? In the meantime, how can researchers realistically get an undistorted picture of all conducted studies – including unpublished studies? These questions should be answered. We hope that our paper will lead to further critical and fruitful discussions about this interesting and important issue in future works.

Acknowledgments

We thank Sarah Tegenfeldt for helpful comments and all participants of the survey conducted as part of the pilot study.

Author Contributions

Conceived and designed the survey: ALB, VS, AH. Conducted and analyzed the data: VS. Wrote the first draft of the manuscript: ALB. Contributed to the writing of the manuscript: ALB, AH. Agreed with manuscript results and conclusions: ALB, VS, AH. Jointly developed the structure and arguments for the paper: ALB, AH. All the authors reviewed and approved the final manuscript.

REFERENCES

1. Sterling TD. Publication decisions and their possible effects on inferences drawn from tests of significance – or vice-versa. *JASA*. 1959;54:30–4.
2. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867–72.
3. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. 1990;263(10):1385–9.
4. Goldacre B. Bad Pharma: How Medicine is Broken, and How We Can Fix It. London: Fourth Estate Ltd; 2013.
5. Olson CM, Rennie D, Cook D, et al. Publication bias in editorial decision making. *JAMA*. 2002;287(21):2825–8.
6. Lee KP, Boyd EA, Holroyd-Leduc JM, Bacchetti P, Bero LA. Predictors of publication: characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Med J Aust*. 2006;184(12):621–6.
7. Lynch JR, Cunningham MR, Warme WJ, Schaad DC, Wolf FM, Leopold SS. Commercially funded and United States-based research is more likely to be published; good-quality studies with negative outcomes are not. *J Bone Joint Surg Am*. 2007;89(5):1010–8.
8. Okike K, Kocher MS, Mehlman CT, Heckman JD, Bhandari M. Publication bias in orthopaedic research: an analysis of scientific factors associated with publication in the Journal of Bone and Joint Surgery (American Volume). *J Bone Joint Surg Am*. 2008;90(3):595–601.
9. Weber EJ, Callaham ML, Wears RL, Barton C, Young G. Unpublished research from a medical specialty meeting: why investigators fail to publish. *JAMA*. 1998;280(3):257–9.
10. Kupfersmid J, Fiala M. A survey of attitudes and behaviors of authors who publish in psychology and education journals. *Am Psychol*. 1991;46(3):249–50.
11. Song F, Parekh S, Hooper L, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*. 2010;14(8):1–193.
12. Haller H, Krauss S. Misinterpretation of significance: a problem students share with their teachers? *Methods Psychol Res Online*. 2002;7(1):1–20.
13. Ioannidis J. Why most published research findings are false. *PLoS Med*. 2005; 2(8):e124.
14. Macleod M, Michie S, Roberts I, et al. Biomedical research: increasing value, reducing waste. *Lancet*. 2014;383(9912):101–4.
15. Ioannidis J. How to make more published research true. *PLoS Med*. 2014;11(10): e1001747.
16. Sutton AJ, Song F, Gilbody SM, Abrams KR. Modelling publication bias in meta-analysis: a review. *Stat Methods Med Res*. 2000;9:421–45.
17. Jin ZC, Zhou XH, He J. Statistical methods for dealing with publication bias in meta-analysis. *Stat Med*. 2015;34(2):343–60.
18. Altman D, Bland J. Statistics notes: absence of evidence is not evidence of absence. *BMJ*. 1995;311(7003):485–485.
19. Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix AL. Overoptimism in bioinformatics: an illustration. *Bioinformatics*. 2010;26(16):1990–8.
20. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
21. Boulesteix AL, Janitza S, Hapfelmeier A, van Steen K, Strobl C. Letter to the editor: on the term 'interaction' and related phrases in the literature on random forests. *Brief Bioinform*. 2015;16(2):338–45.
22. Boulesteix AL, Lauer S, Eugster M. A plea for neutral comparison studies in computational sciences. *PLoS One*. 2013;8(4):e61562.
23. Winham SJ, Colby CL, Freimuth RR, et al. SNP interaction detection with random forests in high-dimensional genetic data. *BMC Bioinform*. 2012;13:164.
24. Boulesteix AL, Hable R, Lauer S, Eugster M. A statistical framework for hypothesis testing in real data comparison studies. *Am Stat*. 2015;69(3):201–12.
25. Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Comput Biol*. 2015;11(4):e1004191.