LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Anne-Laure Boulesteix*, Silke Janitza*, Roman Hornung, Philipp Probst, Hannah Busen, Alexander Hapfelmeier (* equal contribution)

Making Complex Prediction Rules Applicable for Readers: Current Practice in Random Forest Literature and Recommendations

Technical Report Number 199, 2016
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Making Complex Prediction Rules Applicable for Readers: Current Practice in Random Forest Literature and Recommendations

Anne-Laure Boulesteix[1,*]

Silke Janitza[1,*]

Roman Hornung[1]

Philipp Probst[1]

Hannah Busen[1]

Alexander Hapfelmeier[2]

[*] These two authors contributed equally to this work.

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich (LMU), Marchioninistr. 15, D-81377 Munich, Germany. boulesteix@ibe.med.uni-muenchen.de

[2] Department of Medical Statistics and Epidemiology, Ismaningerstr. 22, D-81375 Munich, Germany

Abstract

Ideally, prediction rules (including classifiers as a special case) should be published in such a way that readers may apply them, for example to make predictions for their own data. While this is straightforward for simple prediction rules, such as those based on the logistic regression model, this is much more difficult for complex prediction rules derived by machine learning tools. We conducted a survey of articles reporting prediction rules that were constructed using the random forest algorithm and published in PLOS ONE in 2014-2015 with the aim to identify issues related to their applicability. The presented prediction rules were applicable in only 2 of 30 identified papers, while for further 8 prediction rules it was possible to obtain the necessary information by contacting the authors. Various problems, such as non-response of the authors, hampered the applicability of prediction rules in the other cases. Based on our experiences from the survey, we formulate a set of recommendations for authors publishing complex prediction rules to ensure their applicability for readers.

1 Introduction

In various scientific fields and in life science and medicine in particular, researchers develop prediction models that aim at predicting a condition or outcome of interest based on features often denoted "predictors". The resulting prediction rule, if applied to a new instance, hopefully yields an accurate and useful prediction. For example, predicting the response of a patient to a given therapy is useful because if this patient is unlikely to respond it may be preferable to treat him/her differently in order to avoid side-effects and costs. In the case of a binary outcome, termed Y in this paper, the most popular and very straightforward statistical approach to build such a prediction rule is to assume a logistic regression model

$$P(Y = 1|x_1, x_2, \ldots, x_p) = \frac{exp\left(\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p\right)}{1 + exp\left(\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p\right)} \quad (1)$$

linking the probability that $Y = 1$ to the predictors $x_1, \ldots, x_p$ and to estimate the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ by maximizing the likelihood based on the available training data. The probability that $Y = 1$ is then estimated by replacing the β's by their estimated counterparts and the x's by their realizations for the considered new instance in the above formula. An observation is assigned to class $Y = 1$ if this probability is $> c$, where c is a fixed threshold, and to class $Y = 0$ otherwise. A prediction model that makes use of a cutoff value of $c = 0.5$ is called a Bayes classifier. Any other dichotomization of the predicted values is possible, of course, and corresponding decisions can be guided by receiver operating curve (ROC) analysis and requirements on the sensitivity and specificity of the prediction rule Pepe (2004). Of note, researchers who want to apply the prediction rule to their data only need to know the values of the fitted coefficients—including the intercept.

The very simple logistic regression model may be adapted to take, say, interactions between predictors or non-linear effects into account. However, such a model may not be able to fully capture complex association structures. Furthermore, in cases where the number of predictors exceeds the sample size — as usual in "omics" applications — maximum-likelihood estimation cannot be performed and a penalized variant, such as Lasso or ridge regression, has to be used instead. In principle, the advantage that readers only need the fitted coefficients to apply the rule remains, no matter how these coefficients are estimated. However, in our experience, penalized regression is unfamiliar to (and not easily understood by) most scientists without statistical background.

For all these and further reasons, model-free methods developed by the machine learning community become more and more popular in life science and medicine. In particular the random forest (RF) algorithm Breiman (2001) by Leo Breiman has gained increasing attention in the last years. It is based on the attractive principle of recursive partitioning underlying regression and classification trees Breiman (1998). In contrast to logistic regression, however, prediction rules derived using such algorithms cannot be simply reported in form of coefficient values Boulesteix and Schmid (2014). For RF, which is

considered in this paper, the partition of the predictor space implied by the prediction rule is a very complex one. It is in practice not feasible to report it without providing some kind of software. This may be a strong limitation in practice. A prediction rule should therefore ideally fulfill the following basic criteria.

Availability: While a prediction rule based on the logistic regression model can simply be made available through listing of the regression coefficients of the predictors, a complex prediction rule like RF is not as easy to make available. A software object (option A) may be made available, for example the output of the function 'randomForest' if the R package of the same name is used. Alternatively, the code and data allowing to produce this software object without human intervention (option B) can be made available. For both options, the corresponding files have to be stored somewhere. If they are provided as supplementary materials published together with the paper, they will be made permanently available by the publisher as part of the publication. However, if these files are provided through other channels, they may not be available permanently. This is, for example, the case for authors' personal websites (when authors change job or when the institute's website is restructured), for materials available "on request" (not all email addresses are life-time addresses), or when resorting to public repositories that are not stable, say, without redirection to the new address if it changes. Also, data sharing policies may hinder or prohibit access to data. In such a case, even if code is available, the prediction rule is not available since it cannot be derived without the data. Option A is then the only possible option to make the prediction rule available. See Section 4 for more details on these issues.

Sustainability: Availability should ideally not be limited in time. While a prediction rule based, say, on the logistic regression model will still be applicable in 50 years, no matter which softwares will then be in use, a software object (such as the output of the function 'randomForest' in the R language) or a code to generate this object using the data may work for a particular version of the statistical software and/or package, but not with the future ones that a potential reader will use in several years.

Ease of use: Users need to be able to apply a prediction rule at manageable time, costs and efforts. This criterion has different implications depending on the target user. Statisticians or machine learning scientists are often familiar to both methods and softwares; they prefer software solutions allowing for the application of the prediction rule automatically to large amounts of data. In contrast, medical doctors could have difficulties making a statistical software object run let alone understanding and manipulating code written in any arbitrary programming language. Also, clinical practice often requires speedy decisions and actions, for example in emergency units. Fast solutions that do not require the handling of code and data and that allow the entering of patient profiles by hand one at a time may be preferable in such cases. A specific component of the ease of use is interpretability. Many software solutions produce output, which is only well understood by experts. A prediction rule should ideally generate the output in such a way that the target group of users can make sense of it.

For a prediction rule to be applicable by users in the long term in practice, it ideally has to fulfill these three criteria (availability, sustainability, ease of use). With this understanding of applicability, we conjecture that, because of the difficulties outlined above, many complex prediction rules currently published in the literature can de facto not be applied to the reader's own data. The goal of this paper is two-fold: (i) providing an up-to-date picture of scientific practice with respect to the applicability of prediction rules constructed by RF through a literature survey in PLOS ONE; in other words, answering the question whether RF-based prediction rules described in papers can be applied to the readers' own data at all/with reasonable time and effort; (ii) formulating recommendations based on and beyond the results of the survey discussing potential solutions including recent technical developments.

This paper is structured as follows. In Section 2 we describe the design and methodology of our survey of papers published from 2014/01/01 to 2015/12/31 in PLOS ONE

in the field "Medical and Health Science" that include the phrase "random forest" in the abstract. Results of this survey are described in Section 3, also including the results of an additional study on the availability of estimated regression coefficients in papers presenting prediction rules based on logistic regression. Section 4 gives recommendations to authors presenting complex prediction rules in their papers and discusses potential solutions to address the difficulties encountered when making complex prediction rules available.

2 Methods

2.1 Search strategy

Using the "advanced search" tool of the journal PLOS ONE, we searched for papers satisfying the following criteria:

- Publication date between $2014/01/01$ and $2015/12/31$
- Field "medical and health science"
- Phrase "random forest" in the abstract
- Article type "research paper"

The rationale behind these criteria was as follows. We decided to focus on recent publication dates (i) to give an up-to-date picture of scientific practice, and (ii) to increase the chance that the authors can be contacted under the addresses given in the paper. Note that (ii) is controversial: we might have, instead, decided to consider older papers as well, because in practice papers are not read only in their first 2 years. Using this date restriction, we expect to obtain optimistic results in the sense that it will be on average easier to contact the authors (they are less likely to have changed job since publication) and to apply/produce the prediction rules from a technical point of view (software is less likely to be obsolete) than without it. We decided to focus on the field "medical and health

science" to keep the study feasible and because it is our area of expertise. To increase the rate of relevant papers within the screened papers, we also focused on research papers and papers including the phrase "random forest" in the abstract since, according to our experience, papers mentioning random forest in the text but not in the abstract often do not present applications of RF but mention the method, say, as a potential alternative algorithm to be used in future research. Among the papers screened according to these criteria, we eliminated those:

- that focus on the computational method rather than on the substantive question addressed by the constructed prediction rule(s);

- that use RF to assess/select variables via the so-called variable importance measures (VIM) rather than to fit a prediction rule.

2.2 Collecting information

Each of the papers satisfying these criteria were read by two independent statisticians (Anne-Laure Boulesteix and (Silke Janitza or Roman Hornung or Philipp Probst), from now on denoted as ALB, SJ, RH and PP) with expertise on RF who collected the following information:

- type of data (e.g., clinical, omics, imaging),

- validation scheme (e.g., cross-validation, independent validation),

- performance measure (e.g., accuracy, area under curve),

- whether RF or a non-standard variant of it was used,

- whether other prediction rules were obtained with other methods as well,

- the software (e.g., R, Weka, Matlab) and relevant package (e.g., 'randomForest', 'party') used to construct the RF,

- RF parameter values that were used,

- whether complex data preprocessing has to be performed before applying RF,

- availability of data used to produce the RF (supplementary files, external link, not available),

- availability of codes used to produce the RF (supplementary files, external link, not available).

2.3 Contacting authors

After the process of collecting information, it was assessed whether the RF prediction rule was available from either supplementary files or an external link, for example in form of a software object (option A), or data and code to produce the RF prediction rule (option B). If this was the case a statistician (PP or RH or SJ) who had read the paper and extracted the information mentioned in Subsection 2.2, tried to apply and – if code and data was available – produce the RF prediction rule using the materials provided by the authors. The corresponding author of the article was only contacted if (i) there was no RF software object/prediction tool publicly available and (ii) the data or the complete code or both were not made publicly available. The project leader (ALB) wrote an e-mail (shown in the Supporting Information) and asked for the RF software object or the necessary file(s) to reproduce the RF prediction rule. If the author did not respond after 8 weeks, he/she was sent the same e-mail again and asked for the relevant material, and another 8 weeks were waited in order to declare non-response. The statistician (PP or RH or SJ) tried to produce (if code and data was available) or apply (if a RF software object or RF-based online tool was available) the RF prediction rule with the materials sent by the author. If the author sent any material that was incomplete or unclear, he/she were once again contacted per e-mail by the statistician.

3 Results

3.1 Description of the research paper collection

There were 51 research papers in the field "medical and health science" that were published in PLOS ONE between 2014/01/01 and 2015/31/12 and contained the phrase "random forest" in the abstract. These were screened by the first independent statistician ALB and 17 papers were excluded that put a focus on computational aspects of RF and/or used RF for variable selection through its variable importance measures. A total of 34 research papers were left that were read carefully by ALB and an additional statistician: PP, RH or SJ (Fig 1). Four papers were excluded in this second screening stage because after more careful examination we felt that the authors used RF for another purpose rather than for deriving a prediction rule. Out of these four papers, one paper was excluded because RF was used in a data preprocessing step only. Two papers were excluded because they focussed on the selection of relevant variables that are then included in a logistic regression model or on the selection of biomarkers, respectively, with the help of RF rather than using RF as final prediction method. Another paper was excluded because it had a rather methodological motivation and only applied RF to data for illustrative purposes. The remaining 30 papers (16 "biomolecular", 7 "imaging", 6 "clinical", 1 "accelerometrics") were considered in our study.

In 15 of 30 papers other prediction methods (besides RF) were also used. In the majority of the papers (25 of 30), a complex preprocessing of data was necessary before deriving the prediction rule. In 16 cases the statistical software R was used together with the packages 'randomForest' ($n = 13$), 'randomSurvivalForest'/now 'randomForestSRC' ($n = 1$) or 'party' ($n = 1$). In one paper the used R package was not specified. Matlab was used in 5 papers, Weka in 3 papers and Java and RapidMiner, respectively, in only one paper. There were three papers which did not report the software and for another paper it is unclear whether R or Stata was used for deriving RF.

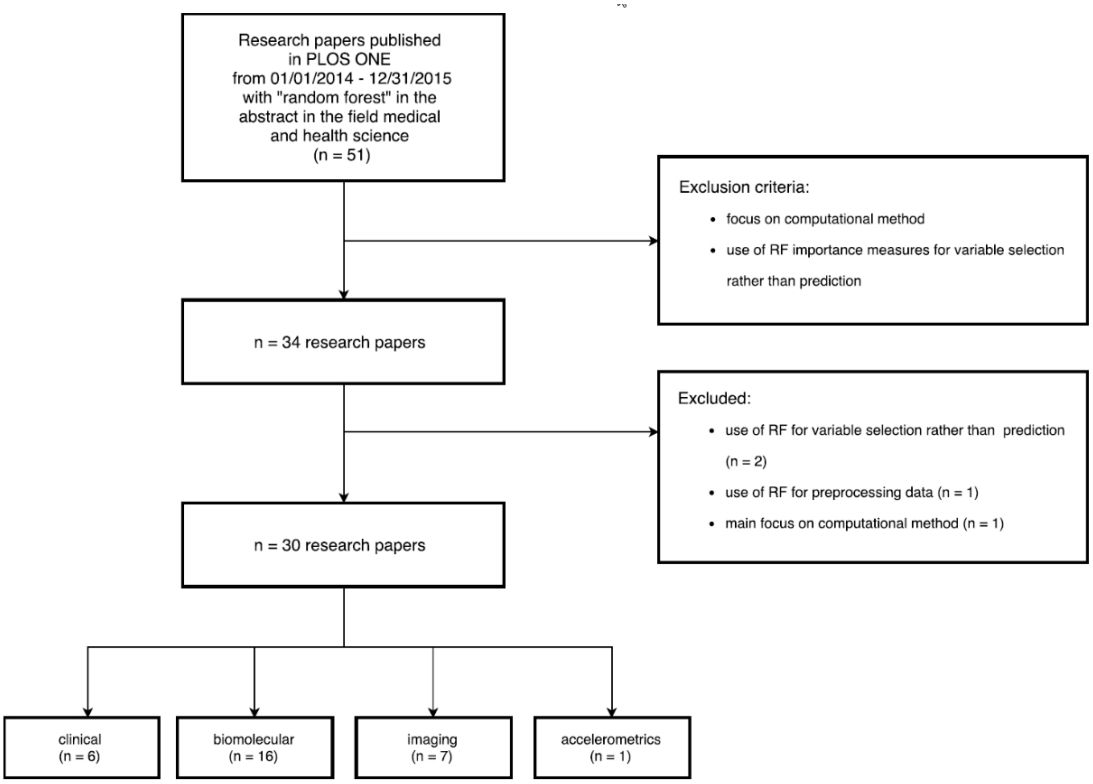

Figure 1: **Flowchart describing the selection of research papers for inclusion in the survey**

3.2 Materials available online

Fig 2 (left stacked barplot) shows the number of papers that published the RF prediction rule, data and code, only the code, only the data or nothing (i.e. neither a prediction tool nor data and code). In only 2 of 30 articles the complete materials for producing or applying, respectively, the RF prediction rule was made publicly available by the authors. One of these two papers reported a link to an RF-based online tool, while the other paper provided both data and code as supplementary files. For the remaining 28 papers neither the RF prediction rule nor the complete materials necessary to produce the RF prediction rule (i.e., data *and* code) were publicly available. The majority of the papers (19 of 30; 63%) did not publish any material of this kind. In 9 papers, only data or only code were published but not both data *and* code. In only one paper the complete codes were published (under a link) but no data. In 8 papers only the data was published (7 as supplementary material, 1 through a link) but no codes. There were also two further papers that gave a

link to the data which did not work.

The red line in Fig 2 crossing the left barplot separates the papers which make the prediction rule or the complete materials used to produce the RF prediction rule publicly available from the papers for which the prediction rule *cannot* be reconstructed *without* having to contact the authors to ask for the necessary materials.

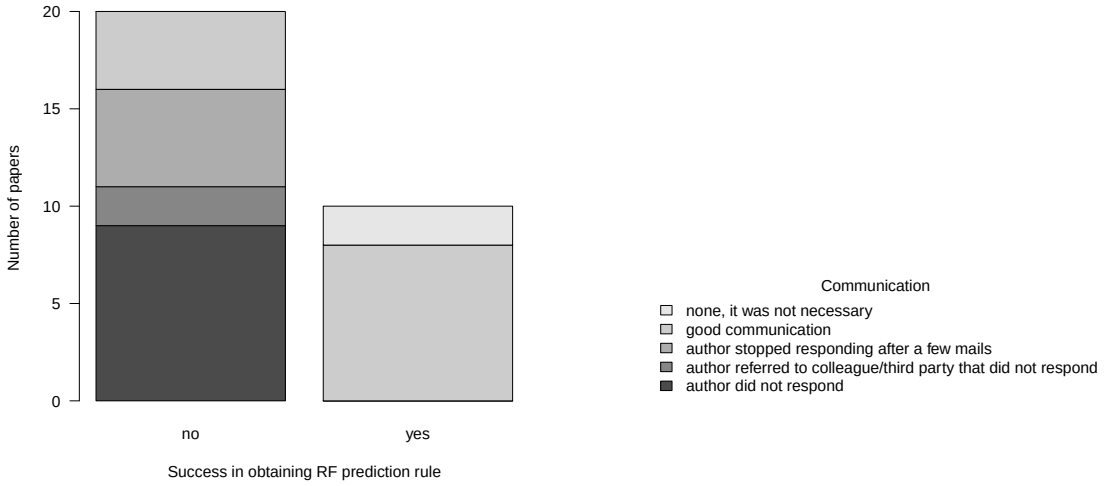

Figure 2: **Number of papers for which the RF prediction rule, both data and code, only the data, only the code or neither data, code nor the RF prediction rule are publicly available (left stacked barplot) or were available after contacting the authors, i.e., publicly available or sent by e-mail (right stacked barplot). The red lines indicate the numbers of papers for which the complete materials used to apply/produce the RF prediction rule were available.**

3.3 Response rate and willingness to share materials

The corresponding authors of nearly all papers (28 of 30) were contacted because the RF prediction rule was not publicly available or could not be produced with the materials provided with the article. For the remaining two papers it was not necessary to contact the authors, because the prediction rule itself or both data and code to produce it were available.

Fig 3 shows the results on the communication process. Among the 28 authors we contacted, 9 authors (32%) did neither respond to our first e-mail nor to the follow-up

e-mail, and 2 authors (7%) referred to a colleague or third party who did not respond. The remaining 17 authors responded to our e-mail. In 8 of 17 cases we obtained the complete materials for producing the RF prediction rule. In the remaining 9 cases the authors responded to our first e-mail, but they did not send the prediction rule or the complete code and data needed to produce it. The reasons for not sending the materials are mostly unknown because in 5 of 9 cases the authors did not respond to our subsequent e-mails. One author did not send the codes and responded that there is a detailed description in the paper that can be used to produce the RF prediction rule. Another author wrote that he is not allowed to make the RF prediction rule available to others, and in another case it was required to write a proposal to get access to the code which we decided not to do. Finally, a further author did not share any material with us because he did not intend the RF prediction rule to be used by practitioners. He said that the prediction accuracy obtained through RF should rather be regarded as an upper limit which can be reached, but that RF is not suitable as a medical prediction tool to be applied in practice due to its complex nature, and that different methods that enable a better interpretation should be used instead.

On the whole, we rated the communication with 12 of the 17 authors who responded as good, meaning that all of our questions were adequately addressed by the authors. As already mentioned, 5 authors stopped the e-mail contact without responding to our question on the availability of data and/or code or—in the case the authors stated that they will not send us both code and data— to the question on the availability of the RF prediction rule.

Besides the information on the number of papers in which material was made publicly available, Fig 2 also shows the number of papers for which material was available after having contacted the authors, that is, the materials were either publicly available or sent us per e-mail by the authors after having contacted them (right stacked barplot). This also includes articles for which, say, the data was publicly available and the codes were sent us per e-mail as well as the two articles for which the complete materials were publicly

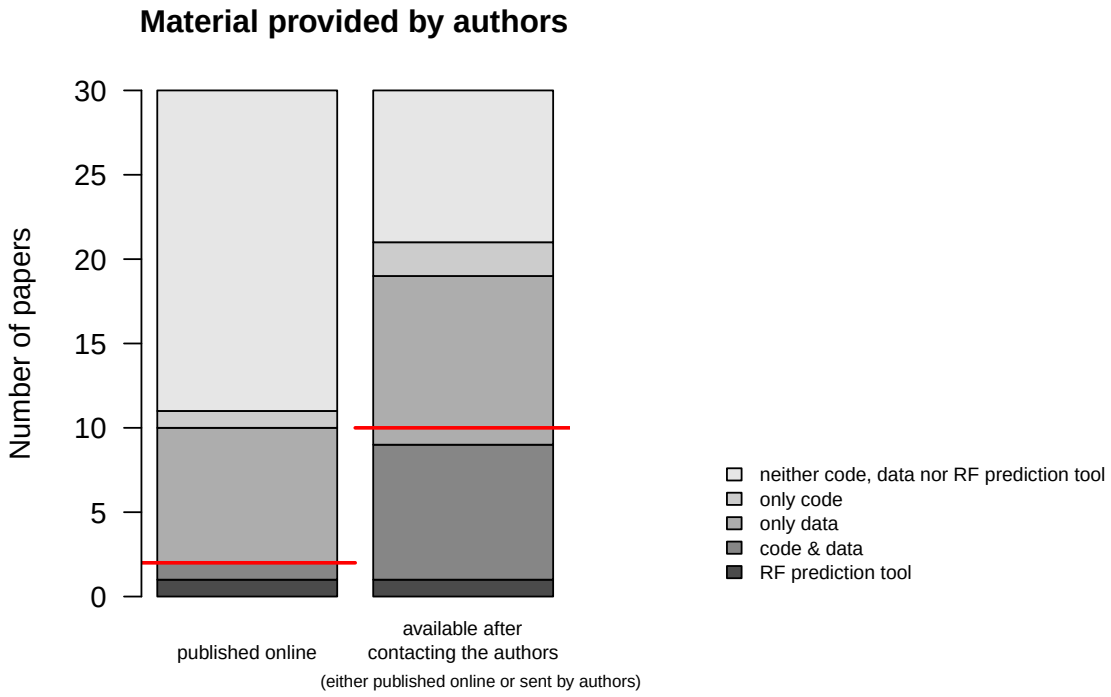

Figure 3: **Rating of the communication with the authors**

available.

For one paper the RF prediction rule was available and for 8 papers we obtained both data and code that are needed to produce a RF prediction rule – note that both code and data were provided online for only one paper, for the remaining 7 papers we obtained the necessary files on request. One further paper used Weka and thus no computer code was necessary. In addition, the authors shared the data and reported the concrete values for RF parameters and the random seed which enables users to produce the RF prediction rule. Thus, in total the authors of 10 of the 30 papers offered materials which are sufficient to make the RF prediction rule available to others. For 9 papers neither data nor code was shared by the authors and for the remaining 11 papers either the data or the code — but not both — was provided which is not sufficient to make a RF prediction rule available.

3.4 Reconstructing RF prediction rules

For all the 10 papers for which we obtained the complete materials after having contacted the authors, we were successful in producing (or applying in the case of the online tool, respectively) the RF prediction rule. This makes up 1/3 of the papers which are included in our survey. The prediction rules we obtained might in principle be applied to future data – ignoring, however, the issue of data preprocessing which is addressed later in this paper. The time effort for producing or applying the prediction rule was very different for the papers. For 4 papers we were able to produce (or apply, respectively) the RF prediction rule in less than 1h. For 4 papers it took between 1h and 4h and for the remaining 2 papers we needed more than 4h to produce the RF prediction rule suggesting that for some papers specific software and subject-matter knowledge (e.g., related to data preprocessing steps and complex data structures) is needed to obtain RF prediction rules. The codes had to be adapted in 3 of the 8 cases for which both data and code was available. Only (very) small changes in the code needed to be performed such as changing paths or names of datasets. This fact illustrates that the reproduction as well as the application of a prediction rule might be difficult for applied researchers who are not familiar with statistical software and requires technical staff that is experienced in the respective software.

We also aimed to assess if it is possible to obtain the same prediction rule that is used by the authors, i.e. if their prediction rule is *reproducible*. This is best accessed through a direct comparison of the two prediction rules. In R, for example, the function 'all.equal' can be used to check if two R-objects are the same. However, we did not have access to the original prediction rule for the 9 papers whose authors shared the complete materials used to produce the RF prediction rule. Therefore, we assessed the reproducibility by comparing the predictions or prediction errors reported in the paper with the results obtained when applying the produced prediction rules to the data offered by the authors. We were able to reproduce the prediction errors reported in 4 papers, which suggests (but does not prove!) that the prediction rule we obtained was the same as the prediction rule of the authors. There was one paper that provided the RF prediction rule as an online tool.

We rated this prediction rule as perfectly reproducible because we can be sure that every user applies exactly the same prediction rule.

To conclude, the prediction rules presented in 5 of the 10 papers for which we were successful in obtaining a prediction rule, are likely reproducible, see also Fig 4. Almost the same prediction rule was obtained for another 4 papers. For these papers the results reported in the paper (e.g. the error rate) were slightly different than those that we obtained. For the last paper we were not able to say whether the prediction rule is the one described in the article: we obtained data and code such that in principle the RF prediction rule could be produced. However, the data preprocessing required very specific expert knowledge. It was thus impossible for us to perform the data preprocessing step to check reproducibility.

3.5 Comparison with logistic regression

The results of our survey point out that applying complex prediction rules derived by RF and presented in research articles most often poses a challenge for the readers—even for articles published in a journal such as PLOS ONE advocating data sharing. While the focus of our study is definitely on RF and we do not intend to run a systematic comparison with other methods, we conducted a simple survey of papers presenting prediction rules based on logistic regression to illustrate that applicability of prediction rules is much less of an issue in this context. One has to keep in mind, however, that our results for logistic regression and for RF are not strictly comparable and should be interpreted very cautiously, since important confounders such as the complexity of the dataset (e.g., whether low- or high-dimensional) or the research field (e.g., medicine or bioinformatics) may explain a large part of the observed differences. Our survey on logistic regression is thus meant as a companion study to give a raw order of magnitude of the applicability of prediction rules derived by logistic regression—without contacting authors since it would have gone beyond the scope of this paper on RF and would not have lead to comparable results anyway for the reasons mentioned above.

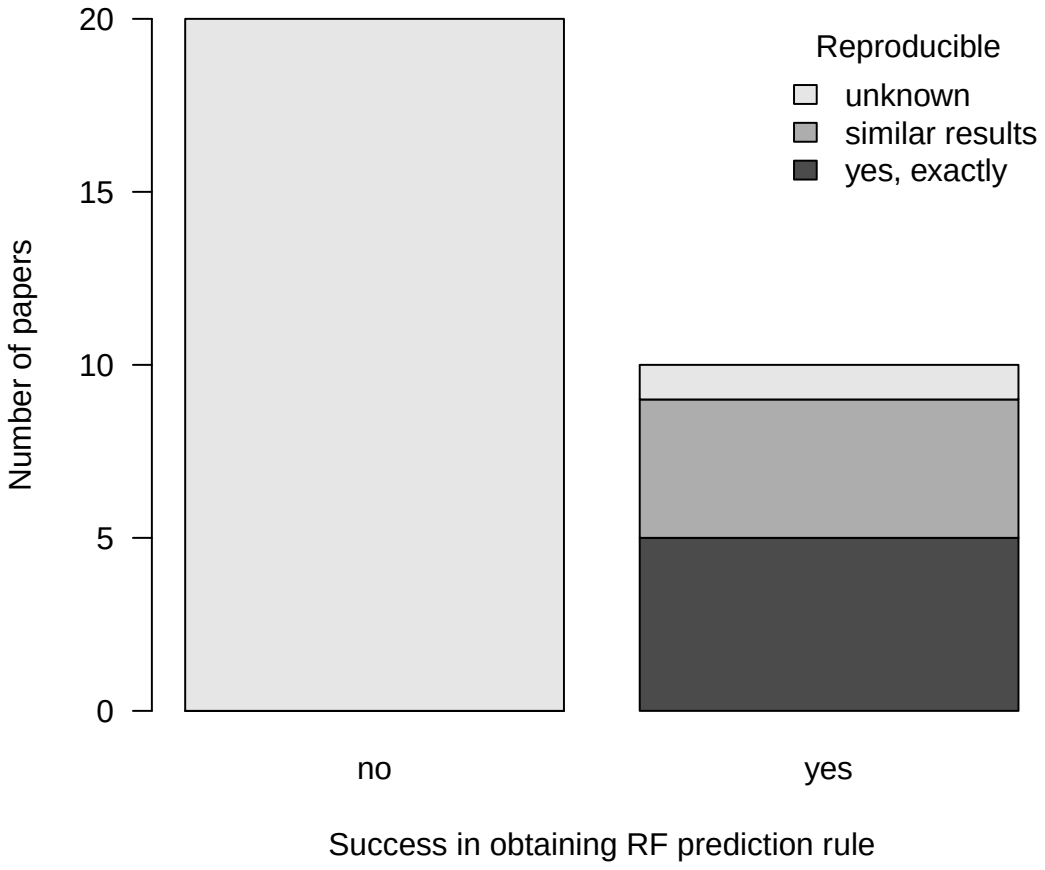

Figure 4: **Number of papers for which we obtained a RF prediction rule and number of papers for which we (likely) obtained exactly the same prediction rule (category "yes, exactly"), almost the same prediction rule ("similar results") or have no information whether we obtained the same prediction rule described in the paper or whether it would be possible to obtain the same prediction rule, respectively ("unknown").**

We searched for articles indexed in Pubmed of type "clinical trial, journal article", published from 2014/01/01 to 2015/12/31, with free full text and with "logistic regression AND (prediction model OR classification model)" in the abstract or title. With these criteria we extracted 130 articles, 8 of which were eliminated because they did not suggest any prediction rule based on logistic regression. From the remaining 122 articles, 53 articles (43%) provided the fitted beta coefficients and intercept of the logistic regression model. Additionally, 2 papers provided the intercept and odds ratios, which can be back-

transformed into coefficients. Therefore 45% of the prediction rules can be applied by the readers without even contacting the authors of the paper. Further, 15 papers and 35 papers omitted the intercept but reported beta coefficients or odds ratios for all predictors included in the model, respectively. Only 17 papers reported neither odds ratios nor beta coefficients. We conjecture that some of these authors would have sent us this information if we had contacted them, perhaps even more often than in the RF survey, because sending information on regression coefficients is much easier than sending the complete codes and data that are needed to reproduce the RF prediction rule.

On the whole, our survey on logistic regression suggests that (i) logistic regression models are often presented in such a way that the resulting prediction rules are applicable by the readers without effort and without contacting the authors; (ii) there is still room for improvement, since this is by far not the case for all articles, although being very easy to implement for logistic regression; (iii) applicability of prediction rules for readers is not a specific problem of RF, even if it is certainly more of an issue for RF than for more simple methods like logistic regression.

4 Recommendations

In this section we formulate recommendations in the form of four possible strategies—denoted as options A, B, C and D—to make complex prediction rules applicable by readers. We also address additional issues related to all four options.

4.1 Making a software object available (option A)

An option to make a prediction rule as obtained using RF applicable for other researchers is to make a software object available that takes new data as input and returns the prediction yielded by the prediction rule. For example, if the R package 'randomForest' is used, one can make the software object returned by the function 'randomForest' available. Predictions can then be obtained by other researchers by passing this object as well as the

new data (for which predictions have to be made) as inputs to the function 'predict'. See the additional materials of Dolch et al. (2016) as an example. This option has the advantage that the prediction rule is applicable by other researchers without making the dataset publicly available. This is an advantage in the case of a confidential dataset, for example medical records. Another benefit is that the user does not have to run (resource intensive) analyses on his/her own to obtain the prediction rule—as opposed to option B presented in the next subsection. This is especially advantageous in the case of high-dimensional datasets.

For the prediction rule to be applicable by other researchers in practice, however, one has to perfectly document the variables' signification, their codings, types and names. Making a toy dataset of the correct format available may be helpful in this respect. Finally, a major disadvantage of option A is that it is completely impossible for the user to modify the prediction rule, for example to make it applicable to a dataset where some variables are missing or coded differently. Furthermore, the prediction rule may become obsolete and do not work anymore, without any possibility to reconstruct it with the new version of the software. These problems are addressed by option B which is described in the next section.

4.2 Making the data and code available (option B)

Alternatively, or in addition to the software object, one can also consider making both the data and code available to potential users. See, e.g., Dolch et al. (2016) and Wang et al. (2014) for examples where both data and code are available as supplementary materials. Data and code should be provided in such a form that one exactly obtains the considered prediction rule automatically. This includes setting random seeds in the code if the considered method producing the prediction rule, such as RF, includes any random component. In principle, all of this should be possible without any other human intervention of the user than a mouse click to run the code (which also calls the data). This principle is the basis of the concept of reproducible research. See for example Hofner et al.

(2016) for a review of issues related to reproducibility and guidelines for publication in the *Biometrical Journal*. This renowned journal now requires authors to submit data and codes implementing their analyses for a systematic check by the reproducibility editor. These guidelines can, in principle, be followed by any researcher who wants to make an RF prediction rule applicable for other researchers. Readers interested in issues related to reproducibility may refer to a special issue of the journal *Science* on this topic Peng (2011).

Besides allowing researchers to adapt the prediction rule to their data, making the computer codes and data available has the advantage that interested readers can find out about details not at all or only briefly mentioned in the paper. For example, readers might want to know how exactly the prediction error rate was estimated (brier score, error rate etc. or cross-validation, independent test data etc.). Such information might easily be obtained by inspecting the computer codes. Note that it is then important to specify the version of the software used for the analysis, since different versions may give (noticeably) different outputs.

Specifying the software and parameters used to construct a complex prediction rule like RF without providing any code should in principle be sufficient to make the prediction rule applicable to other researchers. In practice, however, specifying all parameters is a tedious work and the result of this work will be no better (and often much worse) than making code available. Many journals have a limited page number or word count which further complicates the reporting of prediction rules. Moreover, we experienced that for many medical journals a more detailed description of the statistical methods including parameters used in the models is considered too technical and increases the risk of being rejected.

Please also note that, similarly to option A, code and data should always be carefully documented in order to enable other researchers to use it. Finally, let us point out that applicability as considered in this paper and reproducibility are two related but distinct concepts. Reproducibility (as defined in the previous paragraphs) is only one of several

options to achieve applicability. Conversely, an analysis may be reproducible but fail to be applicable due to problems related to data preprocessing which are common to all options; see Section 4.5.

4.3 Online tools as an option for practical use by lay-persons (option C)

Applied researchers who are typically less familiar with statistical programs would not have to struggle with computer code and data if there is already a user-friendly software solution available that can be applied to their data. The software solution has to suit the needs of the targeted users, who should ideally be involved in the development process. It should be easy to apply and to interpret, both in manageable time and with manageable costs and efforts. A practical software solution which fulfills these criteria is for example provided by Schneider et al. Schneider et al. (2016) who developed a RF risk prediction model for in-hospital mortality of patients with acute cholangitis. It was implemented using the shiny Chang et al. (2015) web application framework for R and is made available online on http://www2.imse.med.tum.de:3838/. A close joint work of researchers from different fields (physicians and statisticians) was supposed to ensure that methodological as well as clinical demands were fulfilled. This lead to a solution that makes a RF prediction model available to physicians who just need to fill in an electronic case report form and are returned a risk estimate with recommendations of treatment actions as a result. Fast applicability of the tool was crucial in this life threatening course of disease.

Another example is the risk prediction model presented in one of the papers included in our survey Gurm et al. (2014) available at https://bmc2.org/calculators/transfusion which predicts the risk of blood transfusion receipt in patients undergoing contemporary Percutaneous Coronary Intervention (PCI) and serves to identify patients who are most likely to receive transfusion after PCI. Note that the calculator available from this link has been updated since the publication of the article, as stated in an introductory note: the webpage is designed in such a way that interested readers have access to both the original

calculator and the updated one.

However, implementing prediction rules as online tools has the same disadvantages as providing software objects: for example, researchers do not have the possibility to adapt the prediction rules to their data.

4.4 Interchange formats (option D)

Beyond the issues of sustainability and data preprocessing, which will be discussed in the next subsections, two additional problems may complicate the application of complex prediction rules presented in the literature. Firstly, it may require much time and effort for a scientist to get familiar with the software that was used to derive the rule (for options A and B). Secondly, in the long term, software objects and code may become obsolete and do not work anymore with current versions of the software (note that old versions of R itself and of R packages are, however, permanently available from CRAN). Interchange formats such as the Predictive Model Markup Language (PMML; see Guazzelli et al. (2009) for an introduction in relation to R and references therein) principally yield solutions to both problems and to the preprocessing issues described in Section 4.6. PMML is an XML-based language and has become the de-facto standard to represent not only predictive models, but also data pre-and post-processing steps. For example, PMML representations of RF can currently be generated in R using the R packages 'pmml' Williams et al. (2016) or its extension 'r2pmml'. In principle, PMML representation of complex prediction rules can be seen as a software-independent "option D", which however requires the corresponding knowhow from potential readers.

4.5 Take care of sustainability when making materials available

No matter which option/options is/are chosen (A - D), the authors should try to maximize the chance that the prediction rule will be applicable by other researchers in the long term. In particular, one should be aware that a good paper's lifetime is usually much longer than the time of its authors at their present institution. Thus, providing the relevant

files on institutional webpages – that may move or may even be completely removed – is not recommended. The same applies to "on request" statements since the authors' contact information may also change. In our study including articles recently published (in years 2014 or 2015), we encountered in two papers that the links to the data did not work. We expect that this problem is even much more prevalent in articles that were published a long time ago. The fact that we observed this problem in two of 30 papers published only 1-2 years ago shows the severity of this problem and illustrates the need for a solution which guarantees permanent access. In general, journal websites are expected to be much more sustainable than personal websites/email-addresses. If hosting of code and data is not allowed by the considered journal, the expected sustainability of the chosen option should be carefully evaluated; for example, a stable repository or the webpage of a senior faculty member may be more sustainable than the webpage of a post-graduate student.

4.6 Report data preprocessing steps carefully

The application of a prediction rule often involves data preprocessing steps such as, depending on the context, normalization, sequence alignment or transformation/recoding of variables. In general, by data preprocessing steps we mean any steps performed before constructing the prediction rule. The data preprocessing step has to be performed prior to applying the prediction rule itself, and depending on the specific context might involve complex operations. This issue seems widely overseen in practice and leads to the inapplicability of many prediction rules, because it is often impossible to perform the preprocessing for independent observations based on the information provided by the authors. Therefore, for all four options (A - D) we recommend authors presenting a prediction rule to systematically describe in detail how to operate to obtain a prediction for a new instance based on unpreprocessed data. Simple examples of such descriptions of prediction rules (in which no data preprocessing steps are involved) would be "give the vector of predictor values of the patient as an input to the function 'predictwithourrandomforest' found in the electronic appendix to obtain a prediction" or "type in the predictor values of a new

patient on our Shiny-website www.ourrandomforestonshiny.com to obtain a prediction." A more sophisticated example (which does involve data preprocessing) would be: "Go to the website www.alignmenttool.com and upload the raw predictor values in the xy-data format in order to obtain the aligned sequence; then, subset the sequence to feature only the spots 1, 3, 10 and 20; then go to our Shiny-website www.ourrandomforestonshiny.com and type in the sequence values at 1, 3, 10 and 20 to obtain a prediction." Apart from an easier applicability of the prediction rule, such an accurate description of the prediction rule would also help assessing whether the prediction rule is applicable to one's own data with reasonable time and efforts. Another recommendation in this context is to automatize all steps that do not require the user's interaction. Note that data preprocessing steps should be performed based on the training dataset only—except if there is good evidence that violating this rule would not lead to over-optimistic prediction error estimation as assessed using an adequate quantitative criterion Hornung et al. (2015). The test dataset has then to be prepared using a so-called *addon procedure*—a term originally used in the context of the normalization of microarray data which is however generalizable to other procedures Hornung et al. (2016).

5 Discussion and concluding remarks

5.1 Summary of the survey

In order to assess how often complex prediction rules presented in the literature are made available we conducted a survey focusing on RF prediction rules and on papers recently published in PLOS ONE. After excluding papers that put a focus on computational aspects of RF and/or used RF for variable selection, our survey comprised 30 articles. Only two of the 30 articles (7%) made the RF prediction rule applicable for readers by providing the necessary materials in supplementary files or through external links (i.e. without having to contact the authors of the paper). This illustrates that in current practice, a minority of the articles reporting RF prediction rules also make these publicly available. In our survey

on RF, we also contacted the authors to assess the chance of obtaining the prediction rule (or code and data to reproduce it) from the authors. We obtained the necessary files from 8 of 28 contacted authors. In all 8 cases the materials were sufficient to produce a RF prediction rule. Accordingly there were 10 articles (out of 30) for which the RF prediction rule could be obtained based on the materials published with the article or sent by the authors after contacting them. Although the majority of the papers did not make the prediction tool available, details on the RF parameters, such as the number of trees for example, were often specified.

5.2 Summary of the recommendations

We mentioned four strategies (options A - D) which might be implemented by authors who want to make their prediction rules available to researchers. The context in which the prediction rule is to be used may guide the decision for or against an option. If the rule is intended to be used by medical doctors for fast decision making, implementing an online tool (option C) might be a good solution. In contrast, if the rule is rather a proof-of-principle and future research has to be conducted to make it usable to practitioners, providing code and data (option B) to allow for further developments might be a better approach. Note that these options are not mutually exclusive. Implementing several options (as Dolch et al. Dolch et al. (2016) who implemented both options A and B) may increase the chance that the prediction rule will be applied by other researchers, especially if it is not clear in which context or by whom (e.g., bioinformatician or medical doctor) the decision rule is to be used. Finally, let us point out that reporting the RF parameters in a paper together with information on the software used is a good starting point but does most often not allow to derive the same prediction rule as described by the authors.

5.3 Limitations of this study

Certain limitations of our study probably lead to over-estimation of the applicability of published random forest prediction rules. Firstly, we are very familiar with RF and RF

software. For applied scientists, for example medical doctors, it might be more difficult and time-consuming to use RF than it was for us. These researchers may finally give up reproducing the prediction rule due to lack of computational expertise. Secondly, our study is limited to the journal PLOS ONE, which has specific strict policies in particular regarding data sharing and transparency. The problem of missing applicability may be (much) more pronounced in other journals. In particular, we expect the rate of data availability to be higher in PLOS ONE than in other journals due to the data policy of PLOS ONE that explicitly encourages authors to publish their data. Thirdly, our study is also limited to recent years: the survey was conducted in 2016 and relates to articles published in 2014 and 2015. The chance to successfully reconstruct a prediction rule obviously drops rapidly during the years after publication due to changes in the authors' addresses (in case it is necessary to contact them) and software obsolescence (in all cases). We thus conjecture that we would have obtained appropriate materials for less papers if we had included older papers that were published, for example, 10 years ago. For these three reasons, our results on the applicability rate of RF prediction rules should be seen as optimistic. Finally, our survey including only 30 papers is intended to identify issues related to applicability of RF rules in a descriptive way but not to provide precise estimates of applicability rates nor to allow making inferences or drawing definitive conclusions.

5.4 Perspectives

Although we put a focus on RF in this paper, the given recommendations are not specific to RF but essentially applicable to any other prediction method, such as support vector machines, neural networks or nearest-neighbors. Exceptions are methods such as logistic regression, which can be easily made available by reporting, for example, regression coefficients (at least when the number of non-zero coefficients is moderate). Note that even in this simple case we found that authors do not systematically make their prediction rules applicable although it seems to be more common than for RF. In our additional survey on logistic regression only 45% of the papers reported the complete list of regression

coefficients including the intercept.

Quite generally, we thus identified two main problems that are relevant to users and developers of prediction methods, respectively, and should be addressed in the future: (i) the lack of awareness and commitment of the scientific community regarding applicability of prediction rules—our paper being a contribution to improve this situation; (ii) the technical difficulties encountered by authors willing to make their prediction rules available. Regarding the latter point, we formulated simple general recommendations, but constant efforts from the statistical and computer science community will be needed to ensure technical feasibility, reliability and user-friendliness of the different options.

Supporting Information

S1_File: **E-mail text sent to the corresponding author**

The text of the standard email sent to the authors of the papers included in our survey. The text was slightly adapted in the cases where materials were already partly publicly available.

Dear [XXX],

I am a professor of biostatistics at the University of Munich, Germany, and currently working on a project on practical applications of random forest methodology. The goal of the project is to investigate in which form scientists make their prediction rules available to readers. I read your article entitled

"[XXX]"

published in PLOS ONE with much interest.

Could you please send me and my colleague [SJ/RH/PP] a software object or some code and data allowing to reproduce your random forest and potentially apply it to data to make predictions?

Many thanks in advance!

Best regards

Anne-Laure Boulesteix

Acknowledgments

We thank all contacted authors who responded to our inquiry. We thank Sarah Tegenfeldt for helping us to prepare the manuscript and Bernd Bischl for helpful comments.

References

Boulesteix, A.-L., Schmid, M., 2014. Machine learning versus statistical modeling. Biometrical Journal 56 (4), 588–593.

Breiman, L., 1998. Classification and regression trees, repr Edition. Chapman & Hall, Boca Raton.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., 2015. shiny: Web Application Framework for R. R package version 0.11.1.
URL http://CRAN.R-project.org/package=shiny

Dolch, M. E., Janitza, S., Boulesteix, A.-L., Grassmann-Lichtenauer, C., Praun, S., Denzer, W., Schelling, G., Schubert, S., 2016. Gram-negative and -positive bacteria differentiation in blood culture samples by headspace volatile compound analysis. Journal of biological research (Thessalonike, Greece) 23, 3.

Guazzelli, A., Zeller, M., Lin, W.-C., Williams, G., 2009. PMML: An open standard for sharing models. The R Journal 1, 60–65.

Gurm, H. S., Kooiman, J., LaLonde, T., Grines, C., Share, D., Seth, M., 2014. A random forest based risk model for reliable and accurate prediction of receipt of transfusion in patients undergoing percutaneous coronary intervention. PLOS ONE 9 (5), e96385.

Hofner, B., Schmid, M., Edler, L., 2016. Reproducible research in statistics: A review and guidelines for the biometrical journal. Biometrical Journal 58 (2), 416–427.

Hornung, R., Bernau, C., Truntzer, C., Wilson, R., Stadler, T., Boulesteix, A.-L., 2015. A measure of the impact of cv incompleteness on prediction error estimation with application to pca and normalization. BMC Medical Research Methodology 15, 95.

Hornung, R., Causeur, D., Bernau, C., Boulesteix, A.-L., 2016. Improving cross-study prediction through addon batch effect adjustment and addon normalization. Bioinformatics.

Peng, R. D., 2011. Reproducible research in computational science. Science 334 (6060), 1226–1227.

Pepe, M. S., 2004. The statistical evaluation of medical tests for classification and prediction. Vol. 31 of Oxford statistical science series. Oxford University Press, Oxford.
URL http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=667806

Schneider, J., Hapfelmeier, A., Thöres, S., Obermeier, A., Schulz, C., Pförringer, D., Nennstiel, S., Spinner, C., Schmid, R. M., Algül, H., et al., 2016. Mortality risk for acute cholangitis (mac): a risk prediction model for in-hospital mortality in patients with acute cholangitis. BMC Gastroenterology 16, 15.

Wang, H., Liu, X., Lv, B., Yang, F., Hong, Y., 2014. Reliable multi-label learning via conformal predictor and random forest for syndrome differentiation of chronic fatigue in traditional chinese medicine. PLOS ONE 9 (6), e99565.

Williams, G., Jena, T., Lin, W. C., Hahsler, M., Inc, Z., Ishwaran, H., Kogalur, U. B., Guha, R., Bolotov, D., 2016. pmml: Generate PMML for Various Models. R package version 1.5.1.
URL https://CRAN.R-project.org/package=pmml