Authors' Reply to "Comments on 'Researcher Bias: The Use of Machine Learning in Software Defect Prediction' "

Martin Shepperd, Tracy Hall, and David Bowes

Abstract—In 2014 we published a meta-analysis of software defect prediction studies [1]. This suggested that the most important factor in determining results was Research Group i.e., who conducts the experiment is more important than the classifier algorithms being investigated. A recent re-analysis [2] sought to argue that the effect is less strong than originally claimed since there is a relationship between Research Group and Dataset. In this response we show (i) the re-analysis is based on a small (21%) subset of our original data, (ii) using the same re-analysis approach with a larger subset shows that Research Group is more important than type of Classifier and (iii) however the data are analysed there is compelling evidence that who conducts the research has an effect on the results. This means that the problem of researcher bias remains. Addressing it should be seen as a matter of priority amongst those of us who conduct and publish experiments comparing the performance of competing software defect prediction systems.

Index Terms—Software quality assurance, defect prediction, researcher bias.

1 Introduction

We thank Tantithamthavorn, McIntosh, Hassan and Matsumoto (TMHM) [2] for their interest in our meta-analysis research published in TSE in 2014 by Shepperd, Bowes and Hall (SBH) [1]. Replication, reanalysis and reinterpretation are vital components to healthy science which is why we were happy to make our research materials open to all. That said we do not fully agree with the analysis TMHM. In particular we consider their use of a small subset (21%) of our data to be a serious flaw which weakens their reanalysis. Second, the usual concerns regarding collinearity apply somewhat differently to categorical explanatory variables.

The remainder of our response provides some context by briefly summarising our original paper SBH and then the results of the TMHM analysis. We then outline our difficulties with the TMHM analysis and conclude by restating what we believe are the major implications for software defect prediction research.

2 Context

Back in 2014 we published a meta-analysis of results derived from computational experiments that compared the performance of competing software defection prediction systems [1]. The motivation for this work was, despite the growing number of such experiments being published, there was little sign of consensus amongst researchers [3]. Our meta-analysis used *all* available studies (42 primary studies reporting 600 prediction results) that satisfied our inclusion criteria and provided sufficient details to enable the meaningful comparisons to be made.

TABLE 1
Eta-Squared Values from the 4-way ANOVA Model (Matthews Correlation Coefficient (MCC) = Response Variable) reproduced from [1, Table 14]

Factor	Sum Sq	% of total variance	F value	Pr(>F)
ResearchGroup	6.38	31.0	16.47	0.000
Dataset	2.31	11.2	6.55	0.000
ResearchGroup: Classifier	1.36	6.6	2.34	0.000
Metric	1.07	5.2	12.19	0.000
Classifier	0.26	1.3	2.12	0.040
ResearchGroup: Dataset	0.22	1.0	6.11	0.002
Residuals	8.98	43.6		

Our main findings were that despite the focus of the experiments being upon the choice of prediction system (specifically classifier) the dominant factors in influencing predictive performance were in decreasing order of effect: (i) Research Group (ii) Data set (iii) Metrics and finally (iv) Classifier. In addition, we found that there are important interactions between Research group and Classifier and also Research Group and Dataset. These results are given in Table 1 in decreasing order of importance.

In short, for our model of predictive performance of software defect classifiers, 31% of the variance in performance is associated with who does the research i.e., which Research Group. However, a further 6.6% is associated with

- *Martin Shepperd and Tracy Hall are with the Department of Computer Science, Brunel University London, UK. E-mail: martin.shepperd@brunel.ac.uk*
- *David Bowes is with the University of Hertfordshire, UK.*

Manuscript received February 13, 2017; accepted July 20, 2017.

the interaction between Research Group and their choice of Classifier and 1% associated with their choice of Dataset. By contrast, the choice of prediction system, i.e., Classifier is only associated with about 1% of the variability in results. We considered this a worrying result.

We have made our raw data available and TMHM have conducted a re-analysis. Their main conclusions are:

- there are strong associations between choice of data set, metrics and research group
- that once this collinearity is addressed then the relationship between research group and experimental results are much reduced when they analyse a subset (21%) derived from using a single data set.

3 Discussion

The main argument of TMHM is therefore that the SBH analysis is weakened due to the lack of independence between factors in our models. They therefore "mitigate for the collinearity between the explanatory variables" (p1093), however we are concerned that their mitigation involves controlling for data set and is therefore based on a subset of seven research groups who use the same data set i.e., Eclipse. The consequence is the remaining 79% of the observations are excluded.

One would not expect each factor to be independent as, for example, it might be reasonable to expect different research groups to focus upon particular types of classifier or repeatedly use the same subset of data sets for reasons of convenience or access. So the question is to what extent does this impact the SBH analysis?

The effect of substantial multicollinearity (i.e., correlation between the factors) in a multivariate linear model such as is used in regression and ANOVA (as per SBH) is potentially twofold. First, it can impact how we interpret the partial correlations of the model and second the predictive stability [4, pp419–430]. But the reader needs to be aware that our models do not use continuous independent variables; that categorical variable with a sparse matrix will almost invariably have some levels that are a linear combination of others. In such cases the corresponding coefficients are not estimated, but the remaining estimates are not affected unlike in the case of collinearity for continuous predictors.

Moreover, in our analysis we are attempting to explain the observed variance in the responses and, in general, the collinearity would affect the variance of the estimated coefficients but cannot, by definition of collinearity, improve the quality of the model because collinear variables span the same subspace.

TABLE 2
Partial Eta-Squared Values for the ANOVA Model (MCC = Response Variable) from [1, Table 13]

Factor	Partial η^2	Significance
Research Group	31.01%	$p < 0.0001$
Dataset	31.00%	$p < 0.0001$
Metrics	12.44%	$p < 0.0001$
Classifier	8.23%	$p < 0.0001$

A related argument from TMHM is that the way the variance is allocated is in some sense arbitrary[1]. Essentially—because not all terms are independent—how the 'shared' variance is allocated amongst the different factors of the model depends upon the order it is specified. This is true although we followed the usual practice by specifying the model in decreasing order of importance of the factors. Nevertheless we can study the effect of each factor *alone* and this was given as the partial eta squared values [1, Table 13] which we reproduce for the reader's convenience as Table 2. Even ignoring other factors and constructing the simplest model imaginable we see Classifier is the least important factor and matters approximately 4x less than Research Group. Note this analysis is conducted using *all* the data.

We also note that TMHM only use the Eclipse subset of data (21% or 126/600 of the cases in our analysis, see Table 3). Even more seriously this reduces the number of distinct research groups from 23 to 7. This alone might explain some of the differences in conclusions. We are not persuaded that it's advantageous to disregard the majority of our evidence particularly since our meta-analysis was based on *all* available studies that satisfied our inclusion criteria. Nor is it clear why the Eclipse subset is chosen since one can extract a larger subset based on usage of the NASA family of data sets.

TABLE 3
Comparison of the Eclipse and NASA dataset subsets

Data set	Groups	Instances	% of SBH Analysis
Eclipse	7	126	21%
NASA	14	351	58%
Original	23	600	100%

Motivated by the existence of another larger subset we repeat the analysis for the NASA data set and find results more in line with our original results and in disagreement with TMHM. The idea is to control for the collinearity by only selecting experimental results from the same data set. Note that Metric is excluded from the model because all observations for the NASA data sets have the same value (the results are given in Table 4 and can be contrasted with a similar 3-way model for Eclipse. In both cases Research Group accounts for more variance in the experimental results than the type of Classifier under investigation, although the effect is more muted for Eclipse.

TABLE 4
Eta-Squared Values for ANOVA Models (MCC = Response Variable) from Eclipse Only (n=126) [2] and NASA only

Subset	Factor	η^2	Significance
Eclipse	Research Group	3.6%	$p < 0.0001$
	Metrics	17.8%	$p < 0.0001$
	Classifier	2.4%	$p = 0.4571$
NASA	Research Group	21.2%	$p < 0.0001$
	Classifier	4.9%	$p < 0.002$
	Group:Classifier	7.8%	$p = 0.0134$

1. Also a minor technicality is that our R implementation uses a sequential method [5] and not hierarchical as suggested by TMHM (p1092).

It is clear that with the much larger NASA subset stronger results are obtained; far more in line with our initial analysis. However the contrasts can be better shown graphically. We simply present the various experimental results as boxplots grouped first by Research Group (our argument) and then by classifier type (TMHM's argument). This may be a simplistic approach but it does reveal the substantive relationships.

Fig. 1. Eclipse: Boxplot of Experimental Result (MCC) grouped by Research Group

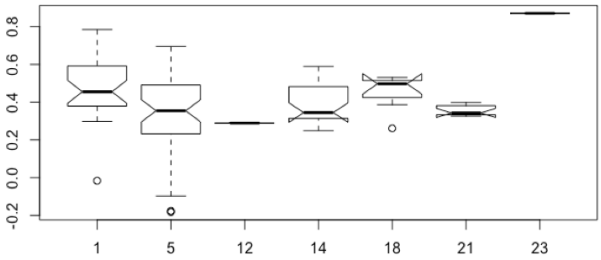
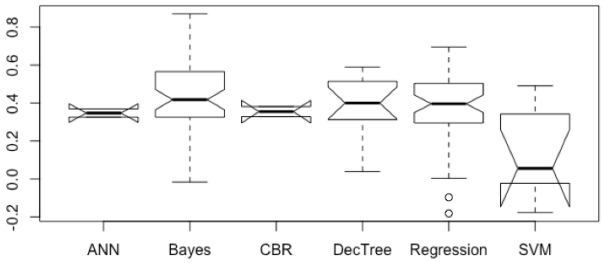


Fig. 2. Eclipse: Boxplot of Experimental Result (MCC) grouped by Classifier Type



The boxplots in Figs. 1 and 2 are based on the Eclipse subset of results showing the spread of predictive performances in terms of MCC, the thick bars show the central tendency as a median and the notches the estimated 95% confidence limits for the medians. Note that in the research group analysis, Group 23 has only a single observation. We see far more variability when the data are grouped by Research Group than when by Classifier with the exception of SVMs where the distribution is skewed by some extremely poor results (negative correlations). Overall this suggests that Research Group is associated with experimental result which is unhelpful if we're seeking repeatable research and conclusion stability.

Repeating the previous analysis for the NASA subset of results (see Figs. 3 and 4) we again observe a similar pattern with greater variability in prediction result associated with Group than Classifier. Note there is only a single observation for Benchmark which in any case one might expect (hope?) to perform less well than the alternative techniques.

Fig. 3. NASA: Boxplot of Experimental Result (MCC) grouped by Research Group

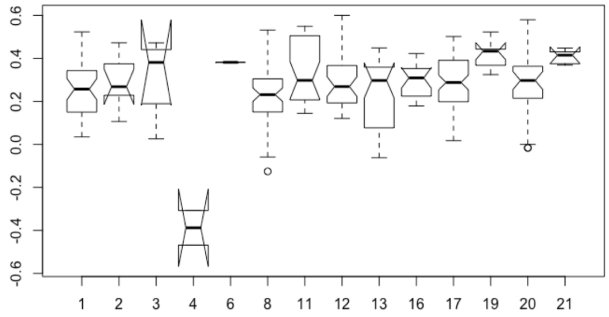
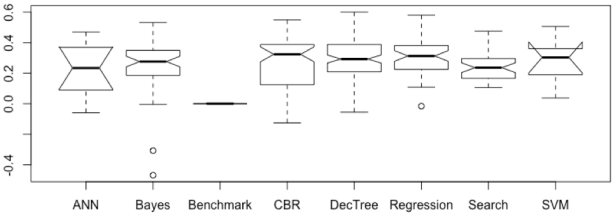


Fig. 4. NASA: Boxplot of Experimental Result (MCC) grouped by Classifier Type



So we see that even when 'controlling' for Dataset there remains a consistent pattern of seeing a stronger association between research group and result than with classifier technique and results. Different approaches to analysing what is a complex meta-analysis yield differences in the degree of the effect but *not* in its existence nor that it dominates the purpose of the experiment i.e., what is being manipulated namely the type of prediction system or classifier algorithm.

4 Conclusion

Of course association does not imply causality and—as TMHM suggest—Research Group is likely to be a proxy for many other factors such as expertise, preferred analysis technique etc. However, it can naturally be interpreted as a collection of factors that confound our research experiments and might reasonably be seen as under the control of the researchers and hence should be of some concern.

So what does this mean for the software engineering community? Clearly the ability to effectively predict defect-prone software components is important so it is unsurprising that this has triggered a good deal of research. Unfortunately there seems to be little consistency in the results which undermines our ability to provide guidance to practitioners.

We have suggested a number of positive steps that might reduce this unwelcome source of variance in experimental results. First, we need reporting protocols to improve

the reproducibility of our experiments. Second, more joint studies could help overcome the problems of comparing the expert application of Technique A with the inexpert application of Technique B. Third, blinding and in particular blind analysis, should become routine practice. Finally, we agree with the suggestion from TMHM that "researchers experiment with a broader selection of datasets and metrics to combat any potential bias in their results".

ACKNOWLEDGMENTS

We would like to thank Chakkrit Tantithamthavorn, Shane McIntosh, Ahmed Hassan and Kenichi Matsumoto for their interest in our study and comments on a previous version of this response. We are particularly indebted to Audris Mockus for his advice on a number of statistical technicalities.

REFERENCES

[1] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," *IEEE Transactions on Software Engineering*, vol. 40, no. 6, pp. 603–616, 2014.

[2] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "Comments on "researcher bias: The use of machine learning in software defect prediction"," *IEEE Transactions on Software Engineering*, vol. 42, no. 11, pp. 1092–1094, 2016.

[3] T. Menzies and M. Shepperd, "Editorial: Special issue on repeatable results in software engineering prediction," *Empirical Software Engineering*, vol. 17, no. 1–2, pp. 1–17, 2012.

[4] J. Cohen, P. Cohen, S. West, and L. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd ed. New York: Routledge, 2003.

[5] R. Kabacoff, *R in Action: Data Analysis and Graphics With R*. Greenwich, CT, USA: Manning Publications, 2013.