Comments on "Researcher bias: The use of machine learning in software defect prediction"

Shepperd et al. find that the reported performance of a defect prediction model shares a strong relationship with the group of researchers who construct the models. In this paper, we perform an alternative investigation of Shepperd et al.'s data. We observe that (a) research group shares a strong association with other explanatory variables (i.e., the dataset and metric families that are used to build a model); (b) the strong association among these explanatory variables makes it difficult to discern the impact of the research group on model performance; and (c) after mitigating the impact of this strong association, we find that the research group has a smaller impact than the metric family. These observations lead us to conclude that the relationship between the researcher group and the performance of a defect prediction model are more likely due to the tendency of researchers to reuse experimental components (e.g., datasets and metrics). We recommend that researchers experiment with a broader selection of datasets and metrics to combat any potential bias in their results.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1260v2 | CC-BY 4.0 Open Access | rec: 11 Apr 2016, publ: 11 Apr 2016

Comments on "Researcher Bias: The Use of Machine Learning in Software Defect Prediction"

Chakkrit Tantithamthavorn, *Student Member, IEEE,*
Shane McIntosh, *Member, IEEE,*
Ahmed E. Hassan, *Member, IEEE,*
and Kenichi Matsumoto, *Senior Member, IEEE*

***Abstract*—Shepperd *et al.* find that the reported performance of a defect prediction model shares a strong relationship with the group of researchers who construct the models. In this paper, we perform an alternative investigation of Shepperd *et al.*'s data. We observe that (a) research group shares a strong association with other explanatory variables (i.e., the dataset and metric families that are used to build a model); (b) the strong association among these explanatory variables makes it difficult to discern the impact of the research group on model performance; and (c) after mitigating the impact of this strong association, we find that the research group has a smaller impact than the metric family. These observations lead us to conclude that the relationship between the researcher group and the performance of a defect prediction model are more likely due to the tendency of researchers to reuse experimental components (e.g., datasets and metrics). We recommend that researchers experiment with a broader selection of datasets and metrics to combat any potential bias in their results.**

I. Introduction

Recently, Shepperd *et al.* [16] study the extent to which the research group that performs a defect prediction study associates with the reported performance of defect prediction models. Through a meta-analysis of 42 primary studies, they find that the reported performance of a defect prediction model shares a strong relationship with the group of researchers who construct the models. Shepperd *et al.*'s findings raise several concerns about the current state of the defect prediction field. Indeed, their findings suggest that many published defect prediction studies are biased, and calls their validity into question.

In this paper, we perform an alternative investigation of Shepperd *et al.*'s data. More specifically, we set out to investigate (1) the strength of the association among the explanatory variables, e.g., research group and metric family (Section II); (2) the interference that these associations introduce when interpreting the impact that explanatory variables have on the reported performance (Section III); and (3) the impact that the explanatory variables have on the reported performance after we mitigate the strong associations among the explanatory variables (Section IV).

II. The Presence of Collinearity

We suspect that research groups are likely to reuse experimental components (e.g., datasets, metrics, and classifiers) in several studies. This tendency to reuse experimental components would introduce a strong association among the explanatory variables of Shepperd *et al.* [16]. To investigate our suspicion, we measure the strength of the association between each pair of the explanatory variables that are used by Shepperd *et al.* [16], i.e., ResearcherGroup, DatasetFamily, MetricFamily, and ClassifierFamily.

TABLE I: The association among explanatory variables.

Pair	Cramer's V	Magnitude
ResearcherGroup & MetricFamily	0.65^{***}	Strong
ResearcherGroup & DatasetFamily	0.56^{***}	Relatively strong
MetricFamily & DatasetFamily	0.55^{***}	Relatively strong
ResearcherGroup & ClassifierFamily	0.54^{***}	Relatively strong
DatasetFamily & ClassifierFamily	0.34^{***}	Moderate
MetricFamily & ClassifierFamily	0.21^{***}	Moderate

Statistical significance of the Pearson χ^2 test:
$\circ p \geq .05$; * $p < .05$; ** $p < .01$; *** $p < .001$

Approach. Since the explanatory variables are categorical, we first use a Pearson χ^2 test [1] to check whether a statistically significant association exists between each pair of explanatory variables ($\alpha = 0.05$). Then, we compute Cramer's V [4] to quantify the strength of the association between each pair of two categorical variables. The value of Cramer's V ranges between 0 (no association) and 1 (strongest association). We use the convention of Rea *et al.* [14] for describing the magnitude of an association. To compute the Pearson's χ^2 and Cramer's V values, we use the implementation provided by the `assocstats` function of the `vcd` R package [10].

Results. **Research group shares a strong association with the dataset and metrics that are used.** Table I shows the Cramer's V values and the p-value of the Pearson χ^2 test for each pair of explanatory variables. The Cramer's V values indicate that research group shares a strong association with the dataset and metrics that are used. Indeed, we find that 13 of the 23 research groups (57%) only experiment with one dataset family, where 9 of them only use one NASA dataset, which contains only one family of software metrics (i.e., static metrics). Moreover, 39% of researcher groups only use the static metric family of the NASA dataset in several studies. The strong association among research groups, dataset, and metrics confirms our suspicion that researchers often reuse experimental components.

III. The Interference of Collinearity

The strong association among explanatory variables that we observe in Section II may introduce interference when one studies the impact that these explanatory variables have on the outcome [7, 18]. Furthermore, this interference among variables may cause impact analyses, such as ANOVA, to report spurious relationships that are dependent on the ordering of variables in the model formula. Indeed, ANOVA is a hierarchical model that first attributes as much variance as it can to the first variable before attributing residual variance to the second variable in the model formula [12]. If two variables share a strong association, the variable that appear first in the model formula will have the brunt of the variance associated with it. Hence, we set out to investigate the interference that is introduced by the strong association among explanatory variables.

C. Tantithamthavorn and K. Matsumoto are with the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. E-mail: {chakkrit-t,matumoto}@is.naist.jp.

S. McIntosh is with the Department of Electrical and Computer Engineering, McGill University, Canada. E-mail: shanemcintosh@acm.org.

A. E. Hassan is with the School of Computing, Queen's University, Canada. E-mail: ahmed@cs.queensu.ca.

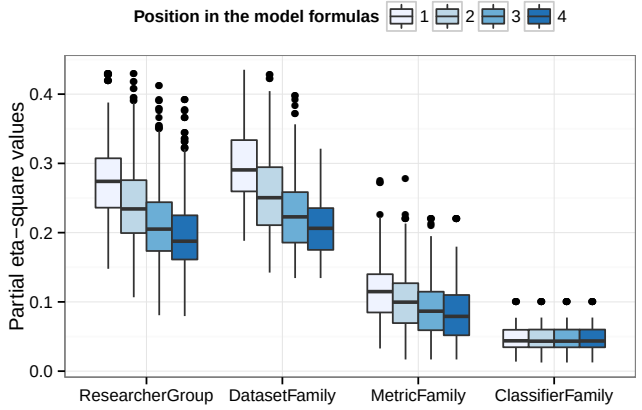

Fig. 1: Distribution of the partial η^2 values for each explanatory variable when it appears at each position in the model formula.

Approach. To investigate this interference, we use a bootstrap analysis approach, which leverages aspects of statistical inference [5]. We first draw a bootstrap sample of size N that is randomly drawn with replacement from an original dataset that is also of size N. We train linear regression models with the data of the bootstrap sample using the implementation provided by the `lm` function of the `stats` R package [13]. For each bootstrap sample, we train multi-way ANOVA models with all of the 24 possible ordering of the explanatory variables (e.g., ANOVA(ResearcherGroup $\times$ DatasetFamily $\times$ MetricFamily $\times$ ClassifierFamily) versus ANOVA(DatasetFamily $\times$ ResearcherGroup $\times$ MetricFamily $\times$ ClassifierFamily)). Following the prior study [16], we compute the partial η^2 values [15], which describe the proportion of the total variance that is attributed to an explanatory variable for each of the 24 models. We use the implementation provided by the `etasq` function of the `heplots` R package [6]. We repeat the experiment 1,000 times for each of the 24 models to produce a distribution of the partial η^2 values for each explanatory variable.

Results. **The strong association among the explanatory variables introduces interference when interpreting the impact that research group has on model performance.** Figure 1 shows the distributions of the partial η^2 values for each explanatory variable when it appears at each position in the model formula. Each boxplot is derived from the models of the 24 possible variable ordering combinations. The results show that there is a decreasing trend in partial eta-squared values when collinearity is not mitigated. Indeed, research group and dataset family tend to have the largest impact when they appear in earlier positions in the model formula, indicating that the impact that explanatory variables have on the outcome depends on the ordering of variables in the model formula. Moreover, we observe that research group and dataset family tend to have comparable partial η^2 values, indicating that the strong association introduces interference. In particular, a model with DatasetFamily in the first position and ResearcherGroup in the second position in the model formula would lead us to believe that DatasetFamily plays a bigger role than the ResearcherGroup. However, once we flip the positions of DatasetFamily and ResearcherGroup, we would reach a different conclusion, i.e., that ResearcherGroup plays a bigger role in the model. On the other hand, the moderate association between the metric and classifier families introduces much less interference.

TABLE II: Partial η^2 values of the multi-way ANOVA analysis with respect to the Eclipse dataset family.

	ResearcherGroup Model	**MetricFamily Model**
Adjusted R^2	0.19	0.36
AIC	-77	-105
	Partial η^2	Partial η^2
Research Group	0.127 (medium)	†
Metric Family	†	0.235 (large)
Classifier Family	0.122 (medium)	0.113 (medium)
Research Group:Classifier Family	0.022 (small)	-
Metric Family:Classifier Family	-	0.162 (large)

†Strong association variables

IV. Mitigating Collinearity

In the study of Shepperd *et al.*, Table 13 is derived from a one-way ANOVA analysis ($y = x_n$) for each of the explanatory variables and Table 14 is derived from a multi-way ANOVA analysis ($y = x_1 \times ... \times x_n$). One of the main assumption of ANOVA analysis is that the explanatory variables must be independent. The prior sections show that such strong association among the explanatory variables (Section II) introduces interference when we interpret the impact that the explanatory variables have on the outcome (Section III). However, Shepperd *et al.*'s multi-way ANOVA analysis did not mitigate for the collinearity between the explanatory variables.

In this section, we set out to investigate the impact that the explanatory variables have on the reported performance after mitigating the interference that is introduced by strongly associated explanatory variables. Thus, our earlier analysis indicates that we cannot include all three of the explanatory variables in the same model. Instead, we should only include two of these three variables in our models. Hence, we opt to focus on one studied dataset in order to control the dataset family metric by holding it constant.

Approach. To mitigate the interference, we first select the Eclipse dataset family, which is the second-most popular dataset family in the studied dataset. We choose the Eclipse dataset family instead of selecting the NASA dataset family because Section II reveals that 39% of the research groups only use the static metric family of the NASA dataset in several studies. Overall, the Eclipse dataset family contains 6 metrics families, which are used by 7 research groups who fit defect prediction model using 6 classifier families. While controlling for the dataset family metric, only research group, metric family, and classifier family will be included in the model formulas. Since the dataset family metric is now a constant, it is excluded from our model formula.

Since Table I shows that research group shares a strong association with the metric family, we build two different linear regression models by removing one of the two strongly

associated variables, i.e., one model uses the research group and classifier family variables, while another model uses the metric family and classifier family variables.

To confirm the absence of collinearity in the models, we perform a redundancy analysis [8] in order to detect redundant variables prior to constructing the models. We use the implementation provided by the `redun` function in the `rms` R package [9].

To assess the fit of our models, we compute the adjusted R^2 and the Akaike Information Criterion (AIC) [2]. The adjusted R^2 measures the amount of variance, while AIC measures the goodness-of-fit based on information entropy. In general, higher adjusted R^2 and lower AIC values correspond to a better fit of the model to the underlying data.

Finally, we perform a multi-way ANOVA analysis and compute the partial η^2 values. As suggested by Richardson *et al.* [15] and Mittas *et al.* [11], we use the convention of Cohen [3] for describing the effect size of the partial η^2 values — values below 0.01, 0.06, and 0.14 describe small, medium, and large effect sizes, respectively.

Results. **When we mitigate the interference of strongly associated explanatory variables, we find that the research group has a smaller impact than the metric family with respect to the Eclipse dataset family.** Table II shows partial η^2 values of the ANOVA analysis with respect to the Eclipse dataset family. The results show that the MetricFamily model, which achieves a higher adjusted R^2 and a lower AIC, tends to represent the underlying data better than the ResearcherGroup model. The redundancy analysis also confirms that there are no redundant variables in the MetricFamily model. Unlike Shepperd *et al.*'s earlier observations, our ANOVA analysis of the MetricFamily model shows that the choice of metrics that are used to build defect prediction models tends to have a large impact on the reported performance with respect to the Eclipse dataset family. Moreover, since the interference has been mitigated, the ANOVA results still hold when the explanatory variables are reordered.

V. Conclusions

The prior work of Shepperd *et al.* [16] suggests that the reported performance of a defect prediction model shares a strong relationship with the research group who conducted the study. This observation raises several concerns about the state of the defect prediction field. In this paper, we investigate (1) the strength of the association among the explanatory variables of Shepperd *et al.*'s study [16]; (2) the interference that these associations introduce when interpreting the impact of the explanatory variables on the reported performance; and (3) the impact that the explanatory variables have on the reported performance after we mitigate the interference introduced by strongly associated explanatory variables. We make the following observations:

- Research group shares a strong association with the dataset and metrics families that are used in building models, suggesting that researchers should experiment with a broader selection of datasets and metrics in order to maximize external validity.
- The strong association among explanatory variables introduces interference when interpreting the impact that research group has on the reported model performance, suggesting that researchers should carefully mitigate collinearity issues prior to analysis in order to maximize internal and construct validity.
- After mitigating the interference, we find that the research group has a smaller impact than metric family with respect to the Eclipse dataset family, suggesting that researchers should carefully examine the choice of metrics when building defect prediction models.

These observations lead us to conclude that the relationship between research groups and the performance of a defect prediction model have more to do with the tendency of researchers to reuse experimental components (e.g., datasets and metrics). Hence, a threat of bias exists if authors fixate on studying the same datasets with the same metrics. We recommend that research groups experiment with different datasets and metrics rather than relying entirely on reusing experimental components.

When adhering to our recommendation, researchers should be mindful of the inherent trade-off between maximizing internal and external validity in empirical research [17]. For example, maximizing external validity by studying a large corpus of datasets may raise threats to the internal validity of the study (i.e., the insights may be difficult to discern due to a broad selection of the studied systems). On the other hand, maximizing internal validity by focusing on a highly controlled experiment may raise threats to the external validity of the study (i.e., the insights may be too specific to the studied systems to generalize to other systems).

Acknowledgments

We greatly appreciate that Shepperd *et al.* [16] shared both their dataset and experimental scripts online, which provided us with the means to conduct this study. In the same spirit, we also provide access to our experimental scripts.[1] This study also would not have been possible without High Performance Computing (HPC) systems provided by the Compute Canada[2] and HPCVL.[3] This work was supported by the JSPS Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers: Interdisciplinary Global Networks for Accelerating Theory and Practice in Software Ecosystem, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] A. Agresti, *An introduction to categorical data analysis*. Wiley New York, 1996, vol. 135.

[2] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[3] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 1988.

[4] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 9.

[1] http://sailhome.cs.queensu.ca/replication/researcher_bias_comments/
[2] https://www.computecanada.ca/
[3] http://www.hpcvl.org/

[5] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boston, MA: Springer US, 1993.

[6] M. Friendly, "heplots: Visualizing hypothesis tests in multivariate linear models." http://CRAN.R-project.org/package=heplots, 2015.

[7] R. Grewal, J. a. Cote, and H. Baumgartner, "Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing," *Marketing Science*, vol. 23, no. 4, pp. 519–529, 2004.

[8] F. E. Harrell Jr., *Regression Modeling Strategies*, 1st ed. Springer, 2002.

[9] ——, "rms: Regression modeling strategies," http://CRAN.R-project.org/package=rms, 2015.

[10] D. Meyer, A. Zeileis, K. Hornik, F. Gerber, and M. Friendly, "vcd: Visualizing categorical data," http://CRAN.R-project.org/package=vcd, 2015.

[11] N. Mittas and L. Angelis, "Ranking and Clustering Software Cost Estimation Models through a Multiple Comparisons Algorithm," *IEEE Transactions on Software Engineering*, vol. 39, no. 4, pp. 537–551, 2013.

[12] B. R. Clifford, "Tests of Hypotheses for Unbalanced Factorial Designs Under Various Regression/Coding Method Combinations," *Educational and Psychological Measurement*, no. 38, pp. 621–631, 1978.

[13] R Core Team, "R: A language and environment for statistical computing," http://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[14] L. M. Rea and R. A. Parker, *Designing and conducting survey research: A comprehensive guide*. John Wiley & Sons, 2014.

[15] J. T. E. Richardson, "Eta squared and partial eta squared as measures of effect size in educational research," *Educational Research Review*, vol. 6, no. 2, pp. 135–147, 2011.

[16] 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.

[17] J. Siegmund, N. Siegmund, and S. Apel, "Views on internal and external validity in empirical software engineering," in *Proceedings of the International Conference on Software Engineering (ICSE)*, 2015, pp. 1276–1304.

[18] Y.-K. Tu, M. Kellett, V. Clerehugh, and M. S. Gilthorpe, "Problems of correlations between explanatory variables in multiple regression analyses in the dental literature." *British dental journal*, vol. 199, no. 7, pp. 457–461, 2005.