To the Editor of Drug Discovery Today

Dear Steve,

We would like to submit our manuscript entitled “Why batch sensitization is important for missing value imputation” for consideration in your journal.

While high-throughput data are important for gene selection in the early phases of drug discovery, missing values are pervasive in such data; and imputing without considering batch factors (batch sensitization) produces confounding effects.

Although many methods have been developed for missing value imputation (MVI) and batch correction respectively, no study has directly considered the confounding impact of MVI on downstream batch correction. This is surprising as missing values are usually resolved prior to batch effect correction, and such MVI approaches usually ignore the batch covariate (i.e., how batch effects are distributed across samples). It is unknown how ignoring batch covariate during MVI impacts proper batch effect inference and removal.

In this manuscript, we explore and evaluate how batch sensitization in MVI impacts batch correction. We examine this problem by modelling 3 imputation strategies: global (M1), self-batch (M2) and cross-batch (M3) using simple matrix simulations, and then demonstrate generalizability on real proteomics and genomics data. Outcome is evaluated based on estimation of remnant batch effects via gPCA delta and imputation accuracy via RMSE.

We demonstrate that considering batch covariates early (M2) on is important as M2 (batch-sensitized) gives the best imputation accuracy after batch correction. Although batch effects do not appear to be problematic following M1 and M3 imputation, this is misleading as M1 and M3 generate noise, with concomitant impact on statistical feature selection. We conclude that performing MVI without considering carefully important co-variates (such as batch effects) can mislead.

We also show that in spite of improved performances given more reasonable imputation assumptions, the performance of imputed matrices does not come close to original data without missing values

The key discussions of this manuscript are:

1. The lack of apparent batch effects in data, does not mean data quality is pristine – M1 and M3 trade batch effect for noise.
2. Imputation methods are no substitute for complete data.

We hope this manuscript is of relevance and interest to your readership. Thank you for your consideration. We hope to hear from you soon.

Warm regards,

Wilson Wen Bin Goh

School of Biological Sciences, Nanyang Technological University, Singapore

Email: wilsongoh@ntu.edu.sg

Limsoon Wong

School of Computing, National University of Singapore, Singapore

Email: wongls@comp.nus.edu.sg