Review of *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-throughput Biology*

Grace Yi Chen

This paper discusses the errors associated with poor data processing documentation and irreproducibility using published studies which analyze high-throughput biological assays. With more high-throughput biological data available, we are able to answer more questions about signature genes and prediction models to discover the relationship between disease response and treatments. However, sometimes data analysis procedures are not documented well enough for other people to reproduce the results. This might lead to errors in the analysis process and wrong conclusions for future implementation of the clinical trials. The authors examined five studies in details and found errors when they tried to reproduce the results according to the procedures described in the paper. Some common mistakes are like mixing up gene labels and group labels in gene expression data. The authors called for detailed documentation and publicly available script for publications.

I think the author demonstrates how easy simple errors could be made when analyzing high-throughput biological data, which could lead to wrong conclusions and waste of resource for future clinical trials. Insufficient data analysis documentation could easily hide simple errors but it takes much more efforts to discover these simple mistakes. I find the statistical workflow mentioned in the last part a great way to make the analysis process standardized and reproducible. Nowadays, with R packages like Rmarkdown and tidyverse, it is much easier to tidy data and prepare reproducible statistical analysis reports. I also find having more than one statistician independently doing data analysis and cross checking very helpful to discover mistakes. In one of my projects, we have two statisticians working independently on the same analysis. We cross-checked from time to time to ensure that the results we provide is correct. In the end, we are more confident with our results to present to the collaborators.

Question:

1. In these highly impactful journals, how will the editors and reviewers check the reproducibility of the analysis scripts to ensure that the conclusion is valid?

2. Having a reproducible workflow is great if the projects we encounter are in similar format. However, for smaller projects, I am not sure if the cost of building up a reproducible workflow as the author described is too high. What do you think?