The effects of sample size on omics study: from the perspective of robustness and diagnostic accuracy

Xiaoyu Zhang¹, a, Xuejiao Cui¹, b and Feng Zhu², c

¹School of Pharmaceutical Sciences, Chongqing University, Chongqing 400044, China; ²School of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China.

a zhangxy12@cqu.edu.cn, b cuixj@cqu.edu.cn, c zhufeng@zju.edu.cn

Abstract. In the design of a specific omics study, sample size is an essential consideration that represents the balance among investigation ability, financial requirements and ethical issues. And it was reported critical to increase robustness and diagnostic accuracy. In this study, two metabolomics datasets were used to conduct analysis to understand the effects of sample size on omics study and strengthen the importance of it. Overlap and AUC were applied to measure the robustness and diagnostic accuracy respectively. The results showed the positive correlation between sample size and robustness, sample size and diagnostic accuracy. Moreover, the effects of sample size vary in datasets and indexes. So, for a specific dataset, sample size should be determined according to its purpose.

1. Introduction

The completion of the Human Genome Project in 2003 [1] and the advancement of high-throughput techniques have triggered the massive growth and deluge of biomedical information, providing an unprecedented opportunity for personalized medicine programs that will significantly improve patient care [2]. The age of big data has arrived [3, 4]. Of the big data in the field of biomedical, omics data is one of the three primary sources, containing a comprehensive catalog of molecular profiles (e.g. genomic, transcriptomic, epigenomic, proteomic, and metabolomic) [5]. Through the application of high-throughput techniques, omics study can measure the expression levels of tens of thousands of variables simultaneously. Therefore, omics is a magic tool to generate lists of markers and construct predictor by identifying differentially expressed genes, proteins or metabolites between different classes.

However, despite considerable successes of omics study in many fields of disease, such as breast cancer, pancreatic cancer, colorectal cancer and so on, there still exist several problems severely hampering the rapid development of omics study, such as the lack of reproducibility of markers [6] and the lack of consistency in generating classifiers [7]. Specifically, the issue of the stability of markers in omics study has been a problem for quite a long time [8-11]. For example, among the two lists of markers identified by two prominent studies about breast cancer [12, 13], only three genes shared on both. Moreover, the discriminatory power of the predictor generated by one of the most acclaimed prognostic gene lists [13], the 70-gene “Amsterdam signature”, dropped to a lower value when validating on another dataset [14].
The problems mentioned above have received great attention and were reported the results of many reasons, such as the heterogeneity in patient cohorts, treatment decisions, technological platforms, approaches used for data analysis, sample size and more. Among these reasons, sample size is an essential consideration that represents the balance among investigation ability, financial requirements and ethical issues[15]. An undersized study means study with insufficient sensitivity that an interesting feature may be missed, resulting in a waste of time and money. Alternatively, when the study is oversized, meaning too many subjects are submitted, the overall costs are increased, especially if it is a potentially harmful or invasive study, there will raise ethical questions [16]. Thus, sample size was suggested critical for improving the predictor accuracy and establishing a reliable list of markers.

In this study, two metabolomic datasets MTBLS28 ESI+ and ESI− [17] from the Metabolites [18] were used to conduct several analyses to understand the effects of sample size in omics study from two aspects, diagnostic accuracy and robustness. The former is the prediction accuracy of a classifier based on marker lists, which could predict the class of a test sample, which is evaluated by receiver operating characteristic (ROC) analysis and measurement of the area under the curve (AUC). The latter is the similarity in content between marker lists generated from various sample sizes, which is measured by overlap [19] between two lists of markers. Different sample sizes are modeled by randomly sampling approach and through the indexes mentioned above, we can see the effects of sample size directly. As a part of omics study, by understating the effects of sample size in metabolomics study, we can extrapolate the results in omics study.

2. Materials and Methods

2.1. Data pre-processing.
Several factors, like unwanted experimental, biological variations and technical errors, can hamper the identification of markers, resulting in missing values and bias of the data [20]. Since several statistical methods based on complete datasets without bias, it is common and useful to conduct imputation and normalization before further study. KNN (k-Nearest Neighbor) algorithm, which was reported as an efficient method for missing value imputation [21], was applied to reduce the sparsity of the data in this study. And MSTUS [22], was applied for data normalization to remove the unwanted sample-to-sample variations. And finally, because of the huge amount of hypothesis tests during the analysis, it is also necessary to conduct a multiple testing adjustment. However, in case of large size of tests and low differential features, the above adjustment will lead to a substantially low power for identifying truly differential features. Data filtering by variance is thus introduced to reduce the size of tests and increase the power [23].

2.2. Statistical analysis.
Prior to statistical analysis, random sampling approach and k-means clustering were applied to construct independent training datasets, testing datasets and sub-datasets of various sample size. About 25% the samples (117 lung cancer patients and 134 healthy individuals) of the data were sampled by k-means clustering as the independent testing data, and the remaining data were divided into 6 sub-datasets with various sample size, which was varied from 50 to 300 by an increment of 50. Theses sub-datasets was constructed by random sampling without replacement, which was repeated 200 times for each sample size.

For each sub-dataset, first, the differentially expressed features were identified by selecting top 100 features with the lowest p-value of Student t-test. Second, based on the Support Vector Machine (SVM), 6 prediction models were constructed from the training samples. The SVM algorithm used in this study is in the R package ‘e1071’ (http://cran.r-project.org/web/packages/e1071).

2.3. Robustness and Diagnostic Accuracy
Robustness is the fraction of shared features that appear on both two lists of markers which is measured by the similarity of two lists of identified markers. For the biomarker discovery from several
datasets, the 100 top ranked features are discovered from the 6 sub-datasets. The value of overlap is then calculated by the part of shared features that appear on both two lists (a and b) of markers. The metric overlap can be calculated using the following equation[19]:

$$\text{Overlap} = 2 \times \frac{\text{interset (Na, Nb)}}{(Na + Nb)}$$  \hspace{1cm} (1)

Diagnostic Accuracy of each model is evaluated by the receiver operating characteristic (ROC) analysis together with the measurement of the area under the curve (AUC) based on the Support Vector Machine (SVM). The testing dataset generated was then used to evaluate the prediction performance based on the markers by R package ‘ROCR’ [24].

3. Results and Discussion

3.1. Increasing sample size improves the robustness of the markers.

As was shown by Fig. 1a and Fig. 1b, the degree of overlap varies in sample size and the larger the sample size, the higher the overlap. For example, the median overlap of MTBS 28 ESI+ and ESI– with a sample size of 100 are 0.21 and 0.20 respectively, while they increase to 0.52 and 0.43 with 300 samples. If you want to increase the stability of the markers identified to improve the robustness, you need to increase the sample size first. The results preliminarily indicate that an increasing sample size can increase the robustness of the markers. And these results are in concordance with others who study the effects of sample size on genomics [10, 25, 26].

To further understand the effects of sample size on robustness, we combine Fig. 1 and Fig. 2 to get Fig. 1c. As illustrated by Fig. 1c, while a good reliability of the markers correlates positively with the sample size, the effects of sample size on robustness differ in datasets. For example, to reach a value of 30%, MTBS 28 ESI+ needs about 150 samples while MTBS 28 ESI– needs about 200 samples. So, if you want to reach a specific high value of overlap, samples needed for a dataset may be different from another.
Fig. 1 Overlap' between Lists of Markers Identified from Two Sub-datasets

(a) MTBLS 28 ESI+  (b) MTBLS 28 ESI−  (c) MTBLS 28 ESI+ and ESI−

3.2. Large sample size guarantees the overall diagnostic accuracy.
The results of how sample size affect the diagnostic accuracy are displayed by Fig.2, which are identical with that of on robustness.

When there are 100 samples of MTBLS 28 ESI+ and ESI−, the AUC values are 0.73 and 0.68 respectively. And when increasing samples to 200, the AUC values increased as well. Therefore, to guarantee a high diagnostic accuracy, we usually need a large sample size.

Fig.2c showed the ROC plot of MTBLS 28 ESI+ and ESI− with a sample size of 300. It suggests that the effect of sample size on diagnostic accuracy also differs in datasets. So, in a specific study, you need to calculate and determine what exactly the sample size is needed to acquire a satisfied result and that is our next work.
3.3. The degree of sample size effects differs in robustness and diagnostic accuracy.

Fig. 1 and Fig. 2 show the positive correlation between sample size and robustness, sample size and diagnostic accuracy respectively. But through the comparison between them, we discover that the degree of sample size effects differs in robustness and diagnostic accuracy and the samples needed to increase overlap and AUC by a specific value are different. As was shown by Fig.1 and Fig.2, less samples are needed to increase overlap than AUC. For example, 100 samples are needed for MTBLS28 ESI+ to increase the value of AUC from 0.66 to 0.76, while only approximate 50 samples are needed to increase the value of overlap by 0.1. The results suggested that for MTBLS ESI+ and ESI−, sample size affects more on robustness than on diagnostic accuracy and thus the sample size determination should depend on the purpose of the study. Further researches are needed to study the differences among these indexes.

4. Conclusion

In this research, we applied two metabolomics datasets to study the effects of the sample size on omics study from the perspective of robustness and diagnostic accuracy. Two indexes, overlap and AUC, were used to measure the robustness and diagnostic accuracy respectively. From the results, we can conclude that the consideration of sample size is essential to whom wants to conduct a study with high robustness and diagnostic accuracy and different sample sizes are needed to reach a specific value of the index for different datasets. Also, the degree of sample size varies in robustness and diagnostic
accuracy. Therefore, sample size determination should be based on the specific purpose of the study. Although, our study focused on metabolomics study, as a part of omics study, we can extrapolate the results to other fields of omics study. Moreover, the results of overlap have been proven by other studies [10, 25, 26]. Our next work is to study the difference among different indexes related with sample size and to build a tool to assess and calculate the sample size for a specific omics study.

References
[1] Collins, F.S., M. Morgan, and A. Patrinos, The Human Genome Project: lessons from large-scale biology. Science, 2003. 300 (5617): p. 286-90.
[2] Costa, F.F., Big data in biomedicine. Drug Discov Today, 2014. 19 (4): p. 433-40.
[3] Chute, C.G., et al., Some experiences and opportunities for big data in translational research. Genet Med, 2013. 15(10): p. 802-9.
[4] Einav, L. and J. Levin, Economics in the age of big data. Science, 2014. 346 (6210): p. 1243089.
[5] Starren, J., M.S. Williams, and E.P. Bottinger, Crossing the omic chasm: a time for omic ancillary systems. JAMA, 2013. 309(12): p. 1237-8.
[6] Shamir, R., et al., Analysis of blood-based gene expression in idiopathic Parkinson disease. Neurology, 2017. 89(16): p. 1676-1683.
[7] Shi, L., et al., The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol, 2010. 28(8): p. 827-38.
[8] Chuang, H.Y., et al., Network-based classification of breast cancer metastasis. Mol Syst Biol, 2007. 3: p. 140.
[9] Kim, K., S.O. Zakharkin, and D.B. Allison, Expectations, validity, and reality in gene expression profiling. J Clin Epidemiol, 2010. 63(9): p. 950-9.
[10] Ein-Dor, L., et al., Outcome signature genes in breast cancer: is there a unique set? Bioinformatics, 2005. 21(2): p. 171-8.
[11] Michiels, S., S. Koscielny, and C. Hill, Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet, 2005. 365(9458): p. 488-92.
[12] Wang, Y., et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet, 2005. 365(9460): p. 671-9.
[13] van ’t Veer, L.J., et al., Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002. 415(6871): p. 530-6.
[14] Buyse, M., et al., Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst, 2006. 98(17): p. 1183-92.
[15] Rutterford, C., A. Copas, and S. Eldridge, Methods for sample size determination in cluster randomized trials. Int J Epidemiol, 2015. 44(3): p. 1051-67.
[16] MckCown, A., et al., Reporting of sample size calculations in analgesic clinical trials: ACTTION systematic review. J Pain, 2015. 16(3): p. 199-206 e1-7.
[17] Mathe, E.A., et al., Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer Res, 2014. 74(12): p. 3259-70.
[18] Haug, K., et al., MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Res, 2013. 41(Database issue): p. D781-6.
[19] Wang, C., et al., The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. Nat Biotechnol, 2014. 32(9): p. 926-32.
[20] Di Guida, R., et al., Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. Metabolomics, 2016. 12: p. 93.
[21] Beretta, L. and A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak, 2016. 16 Suppl 3: p. 74.
[22] Warrack, B.M., et al., Normalization strategies for metabonomic analysis of urine samples. J Chromatogr B Analyt Technol Biomed Life Sci, 2009. 877(5-6): p. 547-52.
[23] Hackstadt, A.J. and A.M. Hess, *Filtering for increased power for microarray data analysis*. BMC Bioinformatics, 2009. 10: p. 11.

[24] Sing, T., et al., *ROCR: visualizing classifier performance in R*. Bioinformatics, 2005. 21(20): p. 3940-1.

[25] van Vliet, M.H., et al., *Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability*. BMC Genomics, 2008. 9: p. 375.

[26] Kim, S.Y., *Effects of sample size on robustness and prediction accuracy of a prognostic gene signature*. BMC Bioinformatics, 2009. 10: p. 147.