

PAPER • OPEN ACCESS

## The effect of data integration on LC-MS-based metabolomics data: evaluation on the comparative classification capacities

To cite this article: Xuejiao Cui *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 032166

View the [article online](#) for updates and enhancements.

# The effect of data integration on LC-MS-based metabolomics data: evaluation on the comparative classification capacities

Xuejiao Cui<sup>a</sup>, Xiaoyu Zhang<sup>b</sup>, Feng Zhu<sup>c</sup>

School of Pharmaceutical Sciences, Chongqing University, Chongqing 400044, China

<sup>a</sup>cuixj@cqu.edu.cn, <sup>b</sup>zhangxy12@cqu.edu.cn, <sup>c</sup>zhufeng@zju.edu.cn

**Abstract.** Data from large-scale LC-MS based metabolomics experiments are generally collected over long periods varying from months to years and has to be divided into several batches, which means for such studies data integration is essential to combine them into one large dataset for data-processing and statistical analysis. This study aims to evaluate the performance of the direct data merge strategy by comparing the performance of classification capacity in direct data merge, result integration and single experiments. Classification capacity 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) applied on both training and testing datasets together with the biomarkers identified by Student's *t*-test ( $p$ -value <0.05). Finally, direct data merge was found to outperform both result integration and single experiment as assessed in this study. In sum, this study shows the classifying accuracy of direct data merge in metabolomics profiling, which gives critical information in data integration in current metabolomics studies.

## 1. Introduction

Metabolomics based on liquid chromatography-mass spectrometry (LC-MS) has been applied in pharmaceutical and clinical research to identify therapeutic targets and discover biomarkers for early disease diagnosis, prognosis [1-3] and illuminate the mechanism of action of new drugs [4-6]. Data from large-scale LC-MS based metabolomics experiments are generally collected over long periods varying from months to years and has to be divided into several batches, which means for such studies information is essential to combine for data-processing and statistical analysis [5, 10]. Therefore, several strategies were applied to data integration, among which the most popular one is *results integration*. So far, the integration of the results of multiple experiments in a large-scale metabolomics profiling has been widely adopted to enhance the reliability of the analytical results [11].

However, *results integration* inevitably results in reducing the statistical power and precludes reanalysing the original data due to the lack of availability of quantitative data [11]. It's thus necessary for us to start from quantitative data instead of directly carrying out integration of results. Up till now, *direct data merge* strategy has made great progress in other omics such as genomics, transcriptomics. And this method is achieved by rescaling of the expression values has been proposed to improve the interpretation of data mining outcomes [11, 12], carrying the potential towards higher accuracy. It is still elusive whether *direct data merge* could enhance the performance of metabolic profiling or not [13]. One of the most important factors is the *classification capacity* which reveals the performance evaluation of biomarker discovery in different disease states [13, 14] or the reliability of identified markers [16].



In the study, *direct data merge* method based on the data related to kidney cancer was performed to integrate the data according to m/z values tolerance which evidently expand the sample size [17]. We select one already published and publicly available data sets containing both healthy person and cancer patient and identify robust and stable biomarkers for integrative analysis. Through evaluating the *classification capacity*, we can assess the reliability of *direct data merge* for a particular study from the point of view of classification accuracy. Based on the results, it can be easily found out that data integration based on *direct data merge* strategy will typically increase the value of the accuracy of marker-finding in metabolomics analysis. In summary, this work aimed at evaluating LC-MS-based metabolomics data integration, and providing the assessment result based on classification accuracy, which is a valuable reference to the choose of *direct data merge* method in metabolomics data analysis

## 2. Materials and Methods

### 2.1. Collection of LC-MS Based Benchmark Metabolomic Datasets.

A systematic literature review on the metabolomics and the analysis on the datasets provided in the *Metabolites* database [18] were collectively conducted to find LC-MS based benchmark metabolomics datasets. After searching against by keyword “LC-MS”, “human” in *MetaboLights*, as shown in Table 1, a publicly available metabolomics datasets were selected (MTBLS17 ESI+) [18,19].

### 2.2. Direct Data Merge Methods Used in This Study

The workflow of the *direct data merge* strategy applied in this work was systematically illustrated in Fig. 1(i). K-Nearest Neighbor (*KNN*) algorithm was preformed to reduce the sparsity of the data in this study. MS Total Useful Signal (*MSTUS*) was applied for data normalization to correct bias. After the above preparations, the training and independent test datasets were further constructed based on the random sampling of the merged dataset. For data-integration's influence on classification accuracy of the identified metabolic markers. First, for the training data, the differentially expressed features were identified by selecting features with  $p$ -value  $>0.05$  of Student  $t$ -test [24]. Second, based on the Support Vector Machine (*SVM*) [25], a prediction model was constructed from the training samples. Finally, the evaluation of the *classification capacity* was obtained by predicting the independent dataset.

### 2.3. Results Integration Methods Used in This Study

As shown in Fig. 2(i), each experimental dataset was conducted using *KNN* for missing value imputation and *MSTUS* for data normalization. Then the training and independent test datasets were constructed by random sampling each pretreated experimental dataset. These datasets were prepared for assessing the *classification capacity* of the *results integration* strategy (described in the last section). With the differential peaks identified by the Student's  $t$ -test ( $p$ -value  $<0.05$ ), the classification models constructed based on experimental datasets were integrated for evaluating the *classification capacity* of *results integration*.

### 2.4. Classification Capacity as the criteria used for evaluating the data integration performance

*Classification capacity* 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*) applied on both training and testing datasets together with the biomarkers identified by Student's  $t$ -test ( $p$ -value  $<0.05$ ) [25]. Samples with large area under *ROC* curve and high *AUC* value is recognized as with better discrimination power.

## 3. Results and Discussion

*Classification capacity* has been a popular criterion in metabolomics [27, 28], which is expected to be quite important because it's necessary for features of the targeted classes to be identified through a combination of their natural properties and external metabolic level. The capacities of the constructed classification model were evaluated by various metrics including accuracy (*ACC*), sensitivity (*SEN*),

specificity (*SPE*), Matthews correlation coefficient (*MCC*), receiver operating characteristics (*ROC*), and the area under *ROC* curve [29,30]. As shown in Fig. 1, four different analytical strategies, including two strategies based on datasets collected from single experiment and two additional strategies of *results integration* and *direct data merge* were evaluated by calculating their *ACC*, *SEN*, *SPE*, and *MCC*. The metrics *ACC* and *MCC* were often used in metabolomics study to evaluate validity of prediction models [31]. In Table 2, the *ACC* of *direct data merge* reaches 0.80, which were substantially and consistently higher than that of the other 3 strategies (0.6~0.69). Similar to *ACC*, the *MCC* of *direct data merge* (0.50) were discovered to be robustly higher than that of the other strategies (0.12~0.13). In this study, *classification capacity* of each model is also 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*) [25]. The pipeline of this assessment method in different data integration methods is shown in the Fig. 1(i) (*Direct data merge*) and Fig. 1(ii) (*Result integration*) below. And a visual processing of statistical results on all data groups were provided in Fig. 2. As shown in the Fig. 2, there were significant differences between the four groups from the *ROC* curve in which *direct data merge* was obviously superior to other methods. And the *AUC* value of *direct data merge* even showed an overwhelming classification capacity performance in all 4 datasets whose value reach 0.83, while the *result integration* method and the single datasets underperform whose only achieve 0.66 and 0.57, 0.76 respectively. As shown, the *ROC* of direct merge group model was found to be significantly higher than these remaining three sets of data model, indicating the usefulness of the *direct merging* method in developing biomarker prediction classification. As a result, *direct data merge* has excellent ability in classification.

#### 4. Conclusion

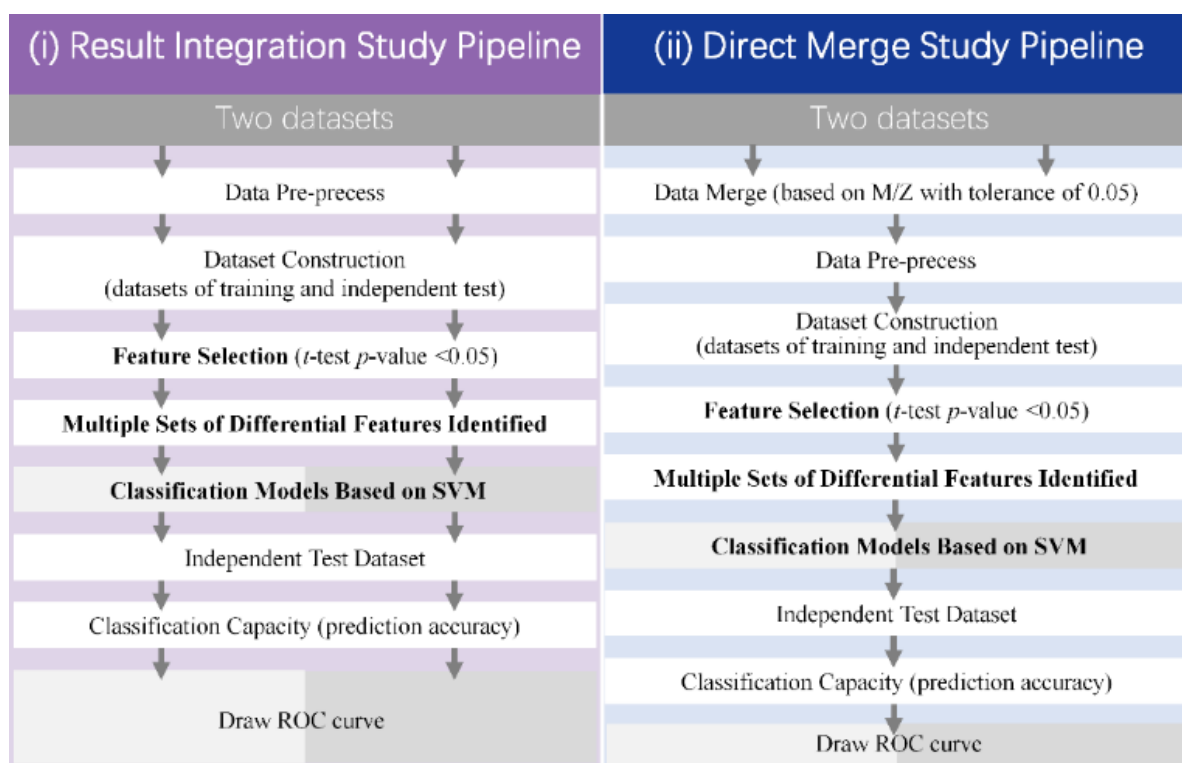
Based on the systematic review of *MetaboLights*, an evaluation on the selected metabolomics data of different analytical strategies was conducted by assessing the its *classification capacity*. As a result, the merging-based strategies (*Direct Data Merge*) performed better than strategies based on *results Integration* and single experiment (Exp1 & Exp2) in term of the *classification capacity*. In conclusion, the findings of this study provided a meaningful guidance to the selection of suitable analytical strategy in a given metabolomics study.

**Table 1.** The number of cases/controls and peaks detected in 4 metabolomics datasets collected in *Metabolites* database

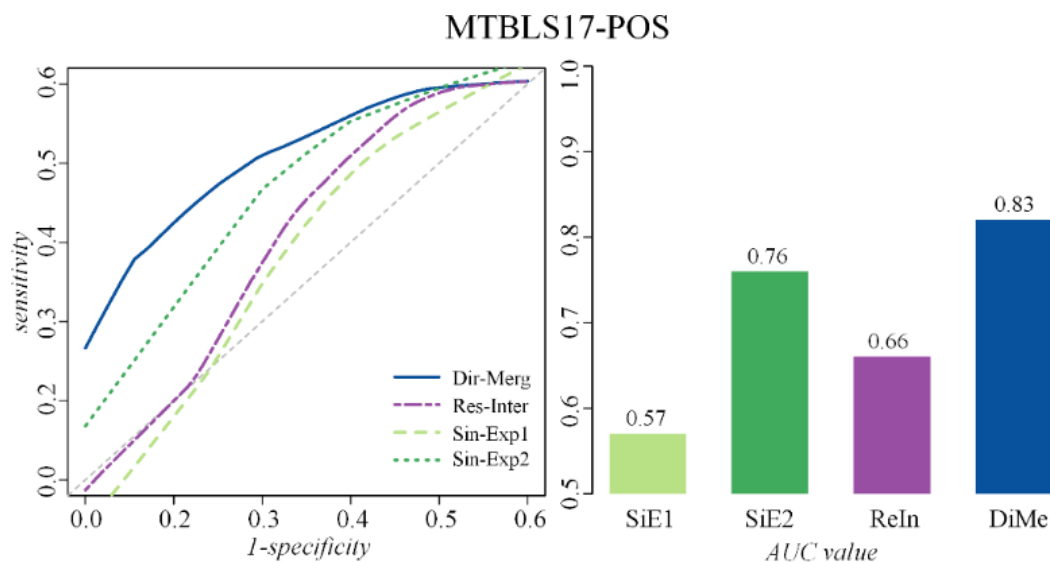
Experiment ID		No. of Cases / Controls	No. of MS Peaks Detected
MTBLS17-P	Single Experiment 1	60/129	1,586
	Single Experiment 2	13/50	3,230
	Result integration	73/179	1,586/3,230
	Direct data merging	73/179	1,144

**Table 2.** Classification capacities of different analytical strategies assessed by accuracy (*ACC*), sensitivity (*SEN*), specificity (*SPE*), Matthews's correlation coefficient (*MCC*) and area under the curve (*AUC*) based on the benchmark datasets.

Experiment ID		ACC	SEN	SPE	MCC	AUC
MTBLS17-POS	Single Experiment 1	0.59	0.58	0.59	0.13	0.57
	Single Experiment 2	0.69	0.33	0.80	0.13	0.76
	Result integration	0.60	0.53	0.62	0.12	0.66
	Direct data merging	0.80	0.53	0.92	0.50	0.83



**Figure 1.** The workflows of the analytical strategies used in this study. (i) the pipeline of direct data merge; (ii) the pipeline of results integration.



**Figure 2.** Classification capacities of different analytical strategies assessed by receiver operating characteristic (ROC) and area under the curve (AUC) based on the benchmark datasets.

## References

- [1] Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov.* 15 (2016) 473 - 484.
- [2] Hu, X., Shen, J., Pu, X. et al. Urinary Time- or Dose-Dependent Metabolic Biomarkers of Aristolochic Acid-Induced Nephrotoxicity in Rats. *Toxicol Sci.* 156 (2017) 123 - 132.
- [3] Jia, H., Shen, X., Guan, Y. et al. Predicting the pathological response to neoadjuvant

- chemoradiation using untargeted metabolomics in locally advanced rectal cancer. *Radiother Oncol.* (2018).
- [4] Zhang, A., Sun, H. & Wang, X. Mass spectrometry-driven drug discovery for development of herbal medicine. *Mass Spectrom Rev.* 37 (2018) 307 - 320.
- [5] Zhou, Z., Tu, J. & Zhu, Z.J. Advancing the large-scale CCS database for metabolomics and lipidomics at the machine-learning era. *Curr Opin Chem Biol.* 42 (2018) 34 - 41.
- [6] Zhang, Q.Q., Huang, W.Q., Gao, Y.Q. et al. Metabolomics Reveals the Efficacy of Caspase Inhibition for Saikosaponin D-Induced Hepatotoxicity. *Front Pharmacol.* 9 (2018) 732.
- [7] Zhao, Y., Hao, Z., Zhao, C. et al. A Novel Strategy for Large-Scale Metabolomics Study by Calibrating Gross and Systematic Errors in Gas Chromatography-Mass Spectrometry. *Anal Chem.* 88 (2016) 2234 - 2242.
- [8] Zhou, Z., Shen, X., Tu, J. et al. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal Chem.* 88 (2016) 11084 - 11091.
- [9] He, J., Wang, K., Zheng, N. et al. Metformin suppressed the proliferation of LoVo cells and induced a time-dependent metabolic and transcriptional alteration. *Sci Rep.* 5 (2015) 17423.
- [10] Brunius, C., Shi, L. & Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics.* 12 (2016) 173.
- [11] Goveia, J., Pircher, A., Conradi, L.C. et al. Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. *EMBO Mol Med.* 8 (2016) 1134 - 1142.
- [12] Larsson, O., Wennmalm, K. & Sandberg, R. Comparative microarray analysis. *OMICS.* 10 (2006) 381 - 397.
- [13] Soto-Iglesias, D., Butakoff, C., Andreu, D. et al. Integration of electro-anatomical and imaging data of the left ventricle: An evaluation framework. *Med Image Anal.* 32 (2016) 131 - 144.
- [14] Date, Y. & Kikuchi, J. Application of a Deep Neural Network to Metabolomics Studies and Its Performance in Determining Important Variables. *Anal Chem.* 90 (2018) 1805 - 1810.
- [15] Maudsley, S., Devanarayan, V., Martin, B. et al. Intelligent and effective informatic deconvolution of "Big Data" and its future impact on the quantitative nature of neurodegenerative disease therapy. *Alzheimers Dement.* 14 (2018) 961 - 975.
- [16] Song, L., Zhuang, P., Lin, M. et al. Urine Metabonomics Reveals Early Biomarkers in Diabetic Cognitive Dysfunction. *J Proteome Res.* 16 (2017) 3180 - 3189.
- [17] Zhang, W., Chang, J., Lei, Z. et al. MET-COFEA: a liquid chromatography/mass spectrometry data processing platform for metabolite compound feature extraction and annotation. *Anal Chem.* 86 (2014) 6245 - 6253.
- [18] Haug, K., Salek, R.M., Conesa, P. et al. MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41 (2013) D781 - 786.
- [19] Resson, H.W., Xiao, J.F., Tuli, L. et al. Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis. *Anal Chim Acta.* 743 (2012) 90 - 100.
- [20] Smith, C.A., Want, E.J., O'Maille, G. et al. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem.* 78 (2006) 779 - 787.
- [21] Di Guida, R., Engel, J., Allwood, J.W. et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics.* 12 (2016) 93.
- [22] Li, B., Tang, J., Yang, Q. et al. Performance Evaluation and Online Realization of Data-driven Normalization Methods Used in LC/MS based Untargeted Metabolomics Analysis. *Sci Rep.* 6 (2016) 38881.
- [23] Beretta, L. & Santaniello, A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inform Decis Mak.* 16 Suppl 3 (2016) 74.

- [24] Sreekumar, A., Poisson, L.M., Rajendiran, T.M. et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*. 457 (2009) 910 - 914.
- [25] Kohl, S.M., Klein, M.S., Hochrein, J. et al. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*. 8 (2012) 146 - 160.
- [26] Xia, J., Sinelnikov, I.V., Han, B. et al. MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic Acids Res.* 43 (2015) W251 - 257.
- [27] De Livera, A.M., Dias, D.A., De Souza, D. et al. Normalizing and integrating metabolomics data. *Anal Chem.* 84 (2012) 10768 - 10776.
- [28] Parsons, H.M., Ludwig, C., Gunther, U.L. et al. Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*. 8 (2007) 234.
- [29] Hart, C.D., Vignoli, A., Tenori, L. et al. Serum Metabolomic Profiles Identify ER-Positive Early Breast Cancer Patients at Increased Risk of Disease Recurrence in a Multicenter Population. *Clin Cancer Res.* 23 (2017) 1422 - 1431.
- [30] Yu, C.Y., Li, X.X., Yang, H. et al. Assessing the Performances of Protein Function Prediction Algorithms from the Perspectives of Identification Accuracy and False Discovery Rate. *Int J Mol Sci.* 19 (2018).
- [31] Alonso, A., Julia, A., Vinaixa, M. et al. Urine metabolome profiling of immune-mediated inflammatory diseases. *BMC Med.* 14 (2016) 133.