**[Name]: A Reproducible Metabolomics Data Processing Pipeline for Supercomputer**

Xinsong Du1, Luran Manfio1, Alexander Kirpich2, William Hogan1, Timothy J. Garrett3, Dominick J. Lemas1

1 Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida

2 Department of Population Health Sciences, School of Public Health, Georgia State University

3 Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida

Keywords: febrile neutropenia; machine learning; mortality; cancer; HCUP

Author to whom correspondence should be addressed:

Dominick J. Lemas, Ph.D

Assistant Professor

Department of Health Outcomes and Biomedical Informatics

University of Florida College of Medicine

2004 Mowry Road- Clinical and Translational Research Building

Gainesville, FL 32608

Ph: 352-294-5971

Email: [djlemas@ufl.edu](mailto:djlemas@ufl.edu)

Abstract Words: 216

Target Journal: GigaScience

Reference limit= No limit

Reference count= 57

Total word limit=3,000

Current word count=2,181

**ABSTRACT**

***Background:***Febrile neutropenia (FN) has been associated with high mortality among adults with cancer. Current systems for early detection of inpatient FN mortality are based on scoring indexes that require intensive physicians’ subjective evaluation.

***Objective****:* In this study, we leveraged machine learning techniques to build a FN mortality risk evaluation tool focused on FN admissions without physicians’ subjective evaluation.

***Methods****:* We used the National Inpatient Sample and Nationwide Inpatient Sample (NIS) that included mortality data among adult inpatients who were diagnosed with FN during a hospital admission. Machine learning techniques that we compared included a linear model (ridge logistic regression) and a non-linear model (gradient boosting tree). The primary outcome for this study was death among individuals with a recorded FN admission. Model comparison was evaluated based on areas under the receiver operating characteristic curve (AUROC) and model performance was estimated using 10 times bootstrap.

***Results****:* Our analysis detected 126,013 adult admissions within the NIS data that were diagnosed with FN, among which 5,856 were declared as deceased (4.6%). Our machine learning results demonstrate linear model (ridge logistic regression) and non-linear model (gradient boosting tree) achieved areas under the receiver operating characteristic (AUROC) above 91% in survival prediction.

***Conclusions****:* We developed machine learning models that do not require physicians’ subjective evaluation for FN mortality risk prediction.

**1.1 INTRODUCTION**

**1.2 METHODS**

**1.3 RESULTS**

**1.4 DISCUSSIONS**

**1.5 CONCLUSIONS:**

**1.5 Acknowledgements**

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases [K01DK115632] and the University of Florida Clinical and Translational Science Institute [UL1TR001427]. The content is solely the responsibility of the authors and does not necessarily represent the official views the University of Florida’s Clinical and Translational Science Institute, or the National Institutes of Health.

**1.6 Authors’ Contributions**

SC, JM, RB, XD designed the initial study protocol. WRH, DJL, SC, JM provided critical suggestions on the clinical part of study design. WRH, DJL, JM, XD provided critical suggestions on the statistical and machine learning related study design. XD wrote codes for all experiments. DJL, JM, XD double checked correctness of codes and experiment results. WRH, DJL, JM, SC, RB, XD wrote and modified the manuscript.

**1.7 Statement on conflict of interest**

The authors have no financial or personal relationships with other people or organizations that could inappropriately influence (bias) their work.

**1.7 SOURCES CITED**

**FIGURES**

**TABLES**

**SUPPLEMENTARY MATERIAL**