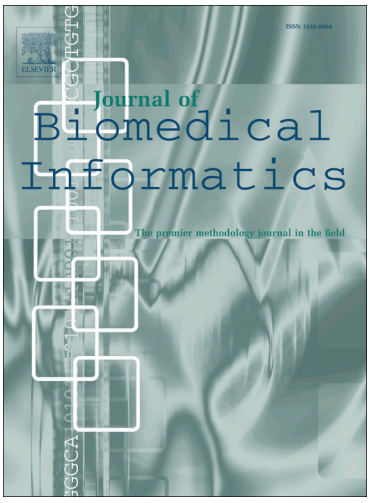

Identification of immune correlates of protection in Shigella infection by application of machine learning

Jorge M. Arevalillo, Marcelo B. Sztein, Karen L. Kotloff, Myron M. Levine, Jakub K. Simon

PII: S1532-0464(17)30184-3
DOI: http://dx.doi.org/10.1016/j.jbi.2017.08.005
Reference: YJBIN 2841

To appear in: *Journal of Biomedical Informatics*

Received Date: 21 February 2017
Revised Date: 10 July 2017
Accepted Date: 8 August 2017

Please cite this article as: Arevalillo, J.M., Sztein, M.B., Kotloff, K.L., Levine, M.M., Simon, J.K., Identification of immune correlates of protection in Shigella infection by application of machine learning, *Journal of Biomedical Informatics* (2017), doi: http://dx.doi.org/10.1016/j.jbi.2017.08.005

Identification of immune correlates of protection in Shigella infection by application of machine learning

Jorge M Arevalillo[a,*], Marcelo B Sztein[b], Karen L Kotloff[b], Myron M Levine[b], Jakub K Simon[c]

[a] *Department of Statistics and Operational Research, University Nacional Educación a Distancia, Paseo Senda del Rey 9, 28040 Madrid, Spain*
[b] *Center for Vaccine Development, Departments of Pediatrics and Medicine, University of Maryland School of Medicine, Baltimore, MD 21201-1509 USA*
[c] *Merck & Co., Inc., Kenilworth, New Jersey USA*

Abstract

Background: Immunologic correlates of protection are important in vaccine development because they give insight into mechanisms of protection, assist in the identification of promising vaccine candidates, and serve as endpoints in bridging clinical vaccine studies. Our goal is the development of a methodology to identify immunologic correlates of protection using the *Shigella* challenge as a model.

Methods: The proposed methodology utilizes the Random Forests (RF) machine learning algorithm as well as Classification and Regression Trees (CART) to detect immune markers that predict protection, identify interactions between variables, and define optimal cutoffs. Logistic regression

*Corresponding author: University Nacional Educación a Distancia. Department of Statistics and Operational Research. Paseo Senda del Rey 9. 28040 Madrid. Spain
Email addresses: jmartin@ccia.uned.es (Jorge M Arevalillo), msztein@medicine.umaryland.edu (Marcelo B Sztein), KKOTLOFF@medicine.umaryland.edu (Karen L Kotloff), MLevine@medicine.umaryland.edu (Myron M Levine), jakub.simon@merck.com (Jakub K Simon)