Information Theory and Bayesian Modeling of prognosis in Lassa Fever patients

Elizabeth Chin[1,2,3], Andres Colubri[2,3], Pardis Sabeti[2,3]

[1] University of California, Los Angeles [2] FAS Center for Systems Biology, Harvard University [3] Broad Institute of MIT and Harvard

INTRODUCTION

With the global rise in human infectious disease outbreaks[1], faster and more throughput methods must be developed to improve patient diagnosis and treatment. Machine learning is often used as a rigorous approach to predict prognosis, yet most of these techniques are restricted to large datasets with complete information, posing a problem during an epidemic, since patient profiles often contain missing values. We developed a robust pipeline that can effectively model patient prognosis, even in the cases where the data base is small and/or incomplete by incorporating both information theory and machine learning techniques.

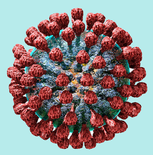
We applied our method to analyze clinical data from approximately 200 patients who suffered from Lassa Fever, an acute viral hemorrhagic disease endemic in parts of west Africa. Despite causing thousands of deaths per year, relatively little is known of the clinical course or predictors of disease outcome. We created data-driven predictive models with an accuracy of over 90% to help health care workers more accurately determine severity and better assess the needs of patients.

CASE STUDY: LASSA FEVER

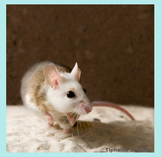
Lassa Virus:

- Negative-strand RNA virus carried and spread by the multimammate rat, the virus' main resevoir[2].
- 100,000 to 300,000 people are estimated to be infected, with approximately 5,000-10,000 deaths per year[2].
- No vaccine is currently available, though Lassa Fever can be treated with the anti-viral drug Ribavirin[2].

Risk map of Lassa Fever in Africa. ISTH one of the few sites with capacity to conduct on-site diagnostics and treatment for Lassa Fever patients[3].



RNA virus. The Lassa virus is spread by the rodent *Mastomys natalensis*[4,5]. Lassa fever infections can be difficult to distinguish from other hemorrhagic fevers such as Ebola and Marburg[3].



Data:

- Demographic, symptomatic, laboratory, and SNP data collected from 203 confirmed Lassa Fever patients between 2012-2013.
- Data collected from the Irrua Specialist Teaching Hospital (ISTH) in Nigeria.

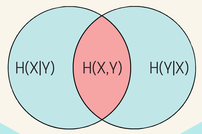
METHODOLOGY

Variable Selection

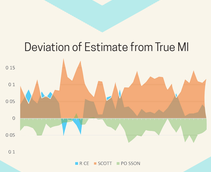
We developed Mirador, a data visualization and exploratory analysis tool used to identify and rank highly correlated variables.

mirador

Estimated relative dependence of variables using **Mutual Information (MI)**.



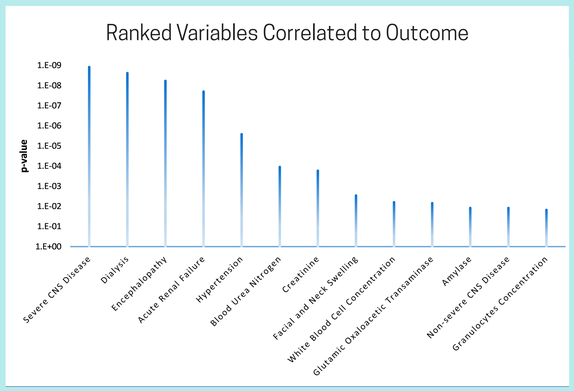
$$I(X,Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$



The distribution of the MI statistic for independent variables can be approximated with a Gamma Distribution[6].

$$\hat{I}(X,Y|Z=z)\ \Gamma(\frac{1}{2}(\chi-1)(Y-1,\frac{1}{N_z})$$

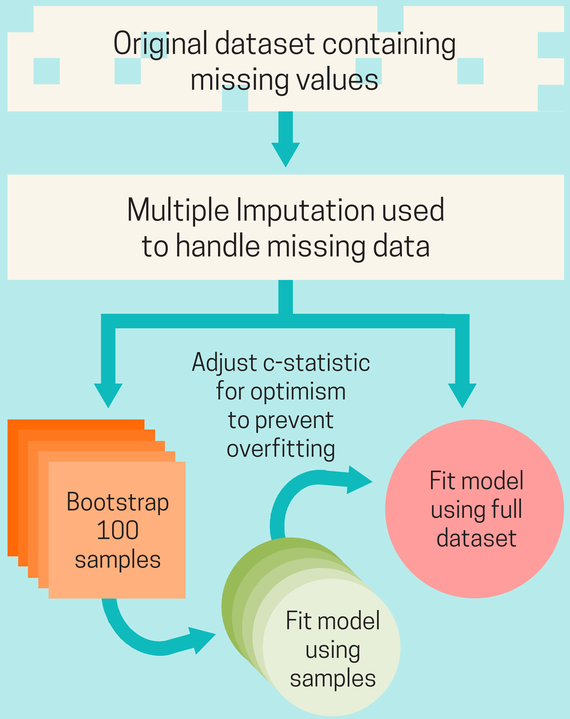
Variables are compared against outcome and sorted by p-value. Variables within a 10% False Discovery Rate are selected for futher processing to control for type I errors (false positives) when conducting multiple comparisons.



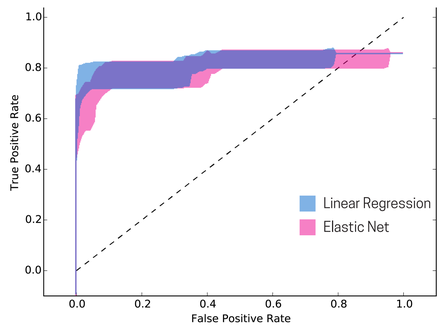
Correlation Ranking. Variables were sorted by p-value and those within 10% FDR were considered for further analysis.

Method Selection

We compared the performance of nine different linear regression methods, and found that the Elastic Net model gave the highest c-statistic. We favored selection reduction methods since more simplistic models are easier for health care workers to implement. The Elastic Net model selected 9 of the top 13 variables in its model.



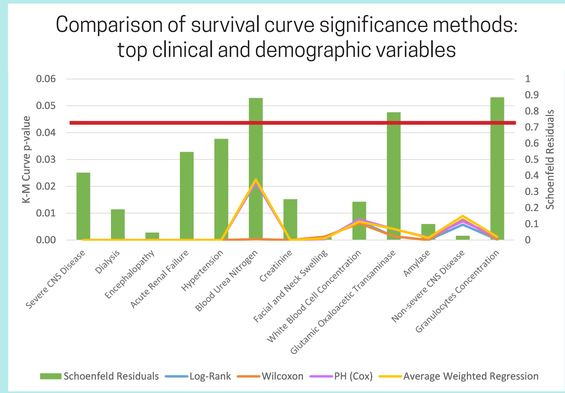
C-Statistic area of Elastic Net vs linear regression models



C-Statistic. The elastic net method had the highest c-statistic, which is the area under the receiver operator characteristic curve, indicating better predictive performance.

Survival Analysis

We analyzed survival using Kaplan-Meier (K-M) curves to better understand the effects of highly correlated variables on survival. Schoenfeld residuals were used to test the non-proportional hazards assumption. Significance between two groups was measured using an average weighted regression model to adjust for non-proportional hazards[7]. All variables had p-values below 0.05 for every survival curve significance test.



Comparison of Survival Significance. A non-significant relationship between residuals and time support the proportional hazards assumption, so Average Weighted Regression test was used when Schoenfeld Residuals were below 0.05.

CONCLUSIONS

The detection of highly correlated variables through mutual information is an exploratory tool for a data-driven hypothesis-generating approach. Our univariate analysis and model selection pipeline was successful in predicting outcome for both Ebola and Lassa Fever. This pipeline has been used to determine dependencies in a diverse range of datasets. Software can be found at http://fathom.info/mirador/.

Acknowledgements

I would like to thank my mentor, Dr. Andres Colubri, for his guidance throughout the project, my adviser, Prof. Pardis Sabeti for her advice and mentorship, and the rest of the Sabeti Lab for their insight. I am grateful for the generous support from the Gwill York and Paul Maeder Research Award for Systems Biology.

References

1. K. F. Smith, M. Goldberg, S. Rosenthal, L. Carlson, J. Chen, C. Chen, S. Ramachandran, Global rise in human infectious disease outbreaks. J. R. Soc. Interface 11, 20140950 (2014).
2. Cunha, Burke A. Infectious Diseases in Critical Care Medicine. New York: CRC Press, 2009. Print.
3. Lassa Risk Map. Adapted from The Mentor Initiative--Reducing deaths and suffering from malaria and other vector borne diseases in humanitarian crises. Lassa Libby - Burch.
4. Kightley, Russell. Lassa Virus. Digital image. Fine Art America. N.p., n.d. Web.
5. Zanutto, Tiphaine. La Mastomys. Digital image. Pavillon Rongeurs. N.p., n.d. Web.
6. Goebel, B., Dawy, Z., Hagenauer, J., Mueller, J.C. (2004), "An approximation to the distribution of finite sample size mutual information estimates," ICC 2005.
7. Schemper, M., Wakounig, S., Heinze, G. (2009): "The estimation of average hazard ratios by weighted Cox regression", Statistics in Medicine 28, 2473 - 2489.

This research was generously supported by the Harvard FAS Center for Systems Biology and the Gwill York and Paul Maeder Research Award for Systems Biology.