CASoS ENGINEERING
http://www.sandia.gov/CasosEngineering

Importance Ranking of Demographic Characteristics that Effect Pandemic Influenza Mitigation Strategies

Gregory J. Lambert, Andrew G. Huff, Patrick D. Finley, Leland B. Evans

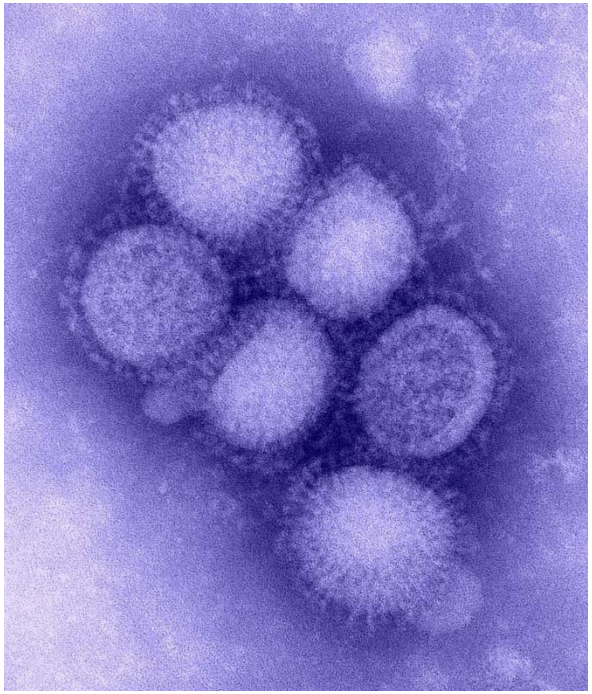
2014 Joint Statistical Meetings
Boston, MA
August 2 - 7, 2014

Sandia National Laboratories

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND2014-16443 PE

Introduction

- 2009 H1N1 influenza pandemic demonstrated that society is inadequately prepared to prevent and mitigate the spread of the disease
- Agent-based model (ABM): used for simulating the actions and interactions of autonomous agents with a goal of assessing their effect on the system
- In this study, a social network agent-based model was used to simulate an influenza pandemic to determine if demographics play a key role in the application of mitigation resources



Electron microscope image of the H1N1 influenza virus

Sandia National Laboratories

Motivation

- Quantify differences in demographic characteristics that result in consistent changes in the recommended mitigation strategies during an influenza pandemic
- Determine changes in demographics that are consistent or inconsistent across different mitigation strategies
- Find a variable importance ranking scheme that can quantify in a meaningful way the loss of predictive power each demographic characteristic has in classifying different mitigation strategies

Sandia National Laboratories

Social network agent-based model

- Simulations begin by creating a community (10,000 individuals) and then seeding the community with a randomly selected infected adults (cases)
- Cases may infect susceptible individuals

Mitigation strategies simulated with 90% compliance rate		
Child Social Distancing	C	Social distancing of children and teenagers. All non-school and non-household contacts with or between children and teenagers reduced by 60% and 90%, household contacts doubled
Household Member Antiviral Prophylaxis	P	Household members given an antiviral with probability (60%, 90%) for 10 days starting immediately after household reference case is diagnosed, reduces infection susceptibility by 30%, reduces probability of clinical illness by 65%, reduces infectivity by 60%
School Closure and Child Social Distancing	S+C	Combination of School Closure and Child Social Distancing

Sandia National Laboratories

Social network agent-based model

- Mitigation strategies are reapplied in additional cycles anytime the number of cases rises above the threshold (3) or until there are no cases in the community. When no cases remain in the community, the simulation ends.

Parameter	Range	Description
Sick at home	0-1	• Scale all non-household contact frequencies by (1-AtHome) • Household contacts are not scaled • Proxy for liberal leave polices in schools and workplace
Babysitting	0-1	If infected person sent home is a child or teen, an adult in their household is nominated as a babysitter. The sitter's household contacts are doubled while all non-household contacts are scaled by (1-Babysitting)
Children	0-3	Number of Children per Household
Teen	0-3	Number of Teens per Household
Adults	1-3	Number of Adults per Household

Sandia National Laboratories

Random forest: permutation variable importance

$$PBVI(X_i) = \frac{1}{T}\sum_{t} errorOOB_{ti} - \ error\overline{OOB}_{ti}$$

- Where $errorOOB_{ti}$ is the error rate of the original variable X_i, $error\overline{OOB}_{ti}$ is the error rate on the permuted variable X_i, and t is an individual tree in the forest of T trees
- Permutation based variable importance (PVI) rankings are difficult to interpret

Sandia National Laboratories

Random forest: K-fold variable importance

$$KFVI_{X_i}: \Delta(errorOOB_{(X_i)^c}, errorOOB_{Full})$$

$$\underbrace{\boxed{X_{i\text{-}1}}\ \boxed{X_{i+1}}\ \ldots\ \boxed{X_n}}_{\textbf{RF without } X_i} - \underbrace{\boxed{X_{i\text{-}1}}\ \boxed{X_i}\ \ldots\ \boxed{X_n}}_{\textbf{RF with } X_i} = \boxed{X_i}$$

- Increase in OOB due to X_i removed
- Can be generalized for *n* independent predictor variables
- Better interpretation of variable importance ranking in terms of loss of predictive power than traditional PBVI method

Simulation study

- The dichotomous response variable was sampled from a multivariate normal distribution

$$X_1, \ldots, X_4 \sim N(0, \Sigma).$$

- Where $\sum$ is the variance-covariance matrix where all four predictors are independent, with $\sigma_{i,i'} = 0$ for all (i, i') .
- Logistic regression coefficients used in the simulation.

X_i	X_1	X_2	X_3	X_4
β_i	3	2	1	0

Sandia National Laboratories

Simulation study results

- Full model correctly predicts category (a) 83% of the time with an OOB error of 17%
- Full model correctly predicts category (b) 84% of the time with an OOB error of 16%
- KFVI allows for more detailed ranking based on different categories in the response variable

X_i removed	Category	OOBE	KFVI
X_1	a	0.391	0.221
X_2	a	0.28	0.110
X_3	a	0.203	0.033
X_4	a	0.167	0.000
X_1	b	0.341	0.181
X_2	b	0.235	0.075
X_3	b	0.185	0.025
X_4	b	0.160	0.000

Comparing variable importance ranking: demographic characteristics

- Full model correctly predicts mitigation (S+C) 90% of the time with an OOB error of 10%
- Full model correctly predicts mitigation (P) 95% of the time with an OOB error of 5%

Demographic removed	Mitigation	OOBE	KFVI
Sick at home	S+C	0.567	0.467
Children	S+C	0.194	0.094
Teen	S+C	0.150	0.050
Babysitting	S+C	0.100	0.000
Adult	S+C	0.100	0.000

Demographic removed	Mitigation	OOBE	KFVI
Sick at home	P	0.398	0.348
Children	P	0.084	0.034
Babysitting	P	0.081	0.031
Adult	P	0.072	0.022
Teen	P	0.048	0.000

Summary and next steps

- KFVI ranking can be used to
  - Quantify the loss of predictive power in main and interaction effects
  - Compare the loss of predictive power across the different categories of a dependent variable
  - Quantify variable importance in terms of probability, allowing for a more natural interpretation in terms of the statistical model
  - Determine changes in demographic characteristics that have consistent or inconsistent effects across different mitigation strategies
- Future challenges
  - Find ways to quantify variable importance of correlated demographic characteristics that effect mitigation strategies that have strong interpretive characteristics