

## **Building a Predictive Model for Influenza Mortality**

**C. Poulin & A. Madsen**

In this initial case study we will demonstrate the building of a predictive model for Influenza Mortality using the Patterns and Predictions™ Bayesian algorithms based tool.

The scope of this predictor currently deals with influenza as defined by the International Classification of Diseases 9th revision (ICD-9) diagnostic codes specific to the various common strains of influenza.

Our technique was to build various Bayes nets given the prior information of the database. Once the models were built, we also provide a preliminary analysis of these models using our analytical tools. Various analysis techniques such as, ‘Inference’, ‘Value of Information’, and scenario based ‘Sensitivity Analysis’ are therefore possible and are included.

### **The Data**

Our experiment is designed to identify patterns in the correlation between various the demographic statistics of ‘zip’ code, ‘year’ of ‘death’, ‘racesex’ (A combined attribute), ‘age’ at death, and the primary prediction target, a classification of the corresponding ‘icdcode’.

Our primary data source is therefore the ‘Compressed Mortality File from 1978-1988’ (CMF) provided by the Office of Analysis and Epidemiology, National Center for Health Statistics, Centers for Disease Control and Prevention. Specifically, this database contains ICD-9 codes associated with a cause of death of influenza. These codes are 487.0, 487.1, and 487.8.

For the CMF file a sample data entry would be: 02215200121448715301

This 20 digit entry can be read by parsing out the relevant data by location. Specifically the key for the locations are: 1-5 ZIP Code : 6-9 Year of Death : 10 Race-sex (1 White Male, 2 White Female, 3 Black Male, 4 Black Female, 5 Other Male, 6 Other Female) : 11-12 Age at Death (01 under a day, 02 1-6 days, 03 7-27 days, 04 28-364 days, 05 1-4 years, 06 5-9 years, 07 10-14 years, 08 15-19 years, 09 20-24 years, 10 25-34 years, 11 35-44 years, 12 45-54 years, 13 55-64 years, 14 65-74 years, 15 75-84 years, 16 85+ years) : 13-16 ICD-9 Code, 17-19 ICD Recode (ICD-8 was used previous to 1978) : 20-23 Total number of deaths

Therefore based on this key, we divide the data as;  
02215 – 2001 – 2 – 14 – 4871 – 530 - 1

Therefore, we can read that this was an individual that was:

Zipcode: Boston : Year of Death: 2001 : Race-Sex: White Female : Age at Death: 65-74  
ICD-9 cause of death: Influenza with pneumonia

For more information please visit: <http://wonder.cdc.gov/wonder/help/mort.html>

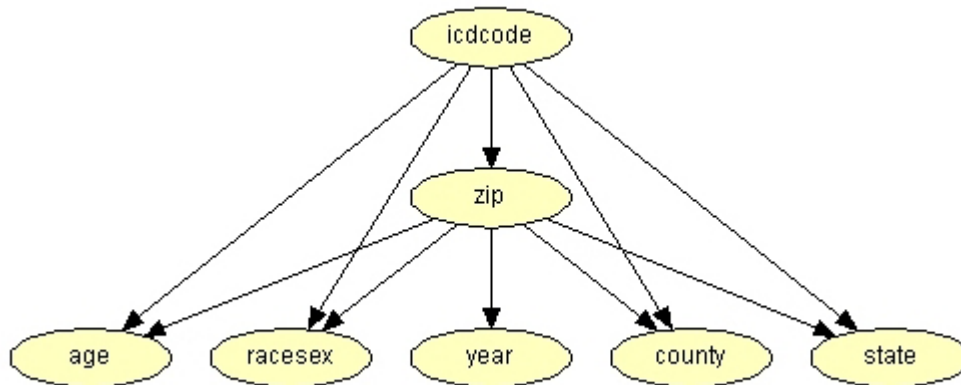
## The Models

The main component of our Bayesian model is a knowledge base that (in this case) depicts the correlation between the *icdcode* and the demographic data within the CMF Mortality database. The three model types that our system allows are Naïve, Tree-Augmented, and Hierarchical Bayesian models. Each has trade-offs in efficiency, and detail on these models will not be discussed in this summary. For more information: <http://www.poulinhugin.com>

Not only does the knowledge base in the model describe possible interdependence relations between factors, it also quantifies the strengths of the relationships. The dependence relations are quantified by conditional probabilities. Finally, the system does not need information about all data to be useful; but with more data the more accurate the predictions. It is important to stress that the model is built from a database containing information on people who died from influenza. Thus, the probabilities are computed given that a person died from influenza. For instance, we may compute the probability distribution over age given the person died from some kind of influenza.

Therefore, we have built two models, each with a different purpose. The first model is an *Influenza Predictor*, built using the entirety of the CMF dataset. This will, given demographic factors, predict the probability a person is white male, for instance, given we know the person died from influenza. But to predict risk of influenza we would need to augment the data with data on people who did not die. The second model is an *Influenza Risk* model which is an analysis of a subset of the data, entirely dealing with the ICD-9 487.x codes. This model is used to determine specific risk factors for mortality associated with Influenza.

Below is a visual representation of a Tree-Augmented (TAN) Model of the *Influenza Risk* model: Illustrating progressive complexity, each oval represents a variable or factor. The factor with label "*icdcode*" represents the classification of Influenza, while each of the other factors represents the demographic information zip, age, racesex, year, county, and state.

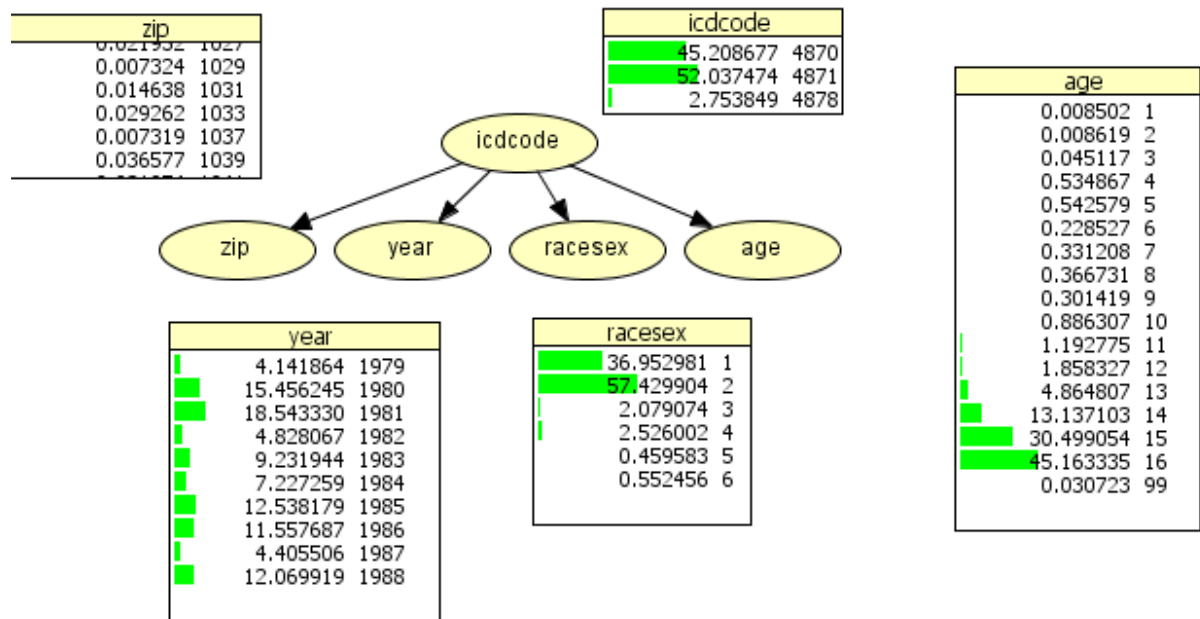


## Analysis

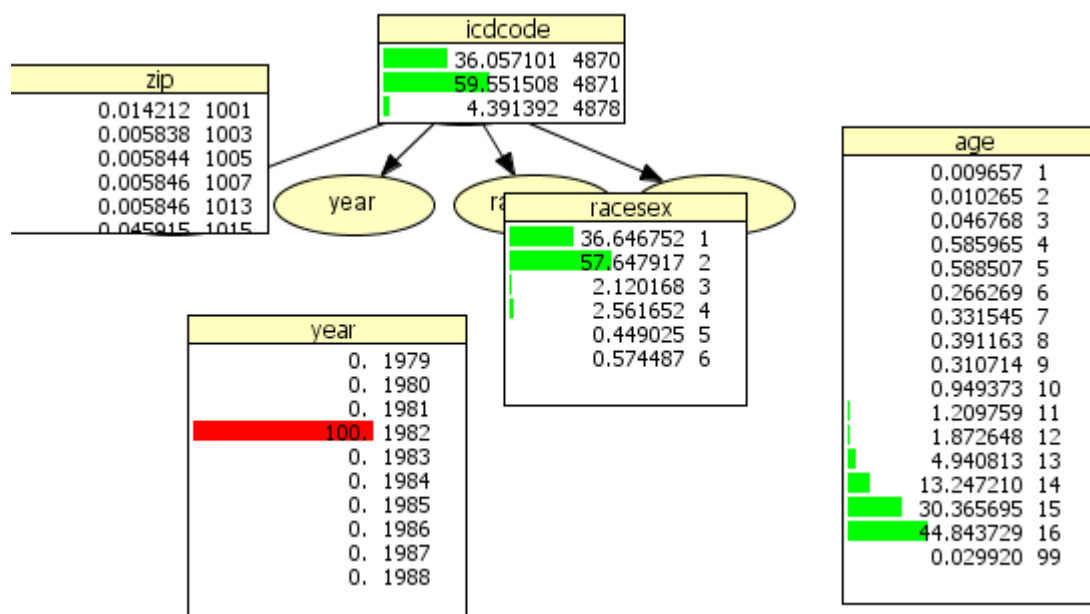
Value of Information (VOI): VOI analysis immediately points to the fact that the highest value of predictive information contained in the geographic information is associated with ZIP CODE (that is to say as greater than any other variable, even County, State). Specifically;

Analysis	Variable	Score
VOI	zip	0.1697
VOI	age	0.0048
VOI	year	0.0035
VOI	racesex	0.0010

Sensitivity analysis (in this case using the Hugin Explorer tool): This exemplary output chart illustrates that there is a probability mass concentration in ICD-9 Influenza type 487.1, White Female victims, 85+ years of age, in the years of 1980-81, 1985-86, and 1988.

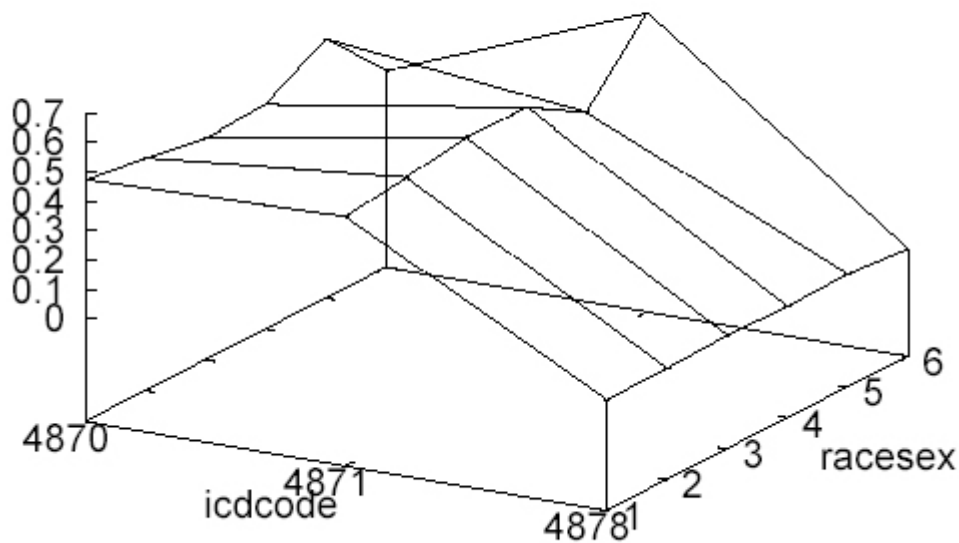


If the user is not satisfied with the strengths of the correlations between factors, they may manually adjust the values of the conditional probability distributions estimated from the data. Therefore, further looking at the specific *Year* of 1982 we find that the probability of diagnosis of “4871” to be greater than the average probability of “4871” being the cause of influenza mortality.



Represented as a plot we can see the probability mass concentration in the 487.1 diagnosis.

### What if analysis on racesex



The plot above shows the sensitivity of the probability of each type of influenza mortality as a function of the value of racesex. It is clear from the plot that the value of racesex has a almost no impact on the probability of the 487.8 diagnosis whereas the probability of both the 487.0 and 487.1 diagnosis is sensitive to racesex. The probability of 487.0 and 487.1 is most sensitive to racesex equal to Other Male (5) and Other Female (6).

### **Observations:**

According to our model:

- The two most common types of influenza mortalities have ICD-9 code "487.0" and "487.1" corresponding to influenza with pneumonia and influenza with other respiratory manifestations, respectively ("487.8" is influenza with other manifestations). For instance, Given that a person died from influenza the probability that it is influenza with pneumonia is 45.2%
- Given that a person died from influenza the probability that the person has an age of 75-84 is most likely at 30.5%.
- The probability that person who died from influenza is a white female is 57.4%. And for year 1982 this probability is highest at 57.6%.
- If a child of age 'one day' has died from influenza the probability that the child died from influenza with pneumonia is 93.6%.
- Given that a person died from influenza in the period 1979 to 1988, it is most likely that the person died in 1981 (18.5%).

## **Summary**

Our Bayesian Influenza model allows a thorough predictive analysis of Influenza mortality. We hope that derived models from this example may therefore have epidemiological value. We wish to extend this technology to the health community, as this case study has been purely illustrative of our technology.

By combining the use of Inference, Value of Information analysis and What-If Sensitivity analysis, we can focus on detailed parts of the model that are more correlative, thereby improving both accuracy and a sense of the ‘complete picture’. This technique could be extended to isolate data (such as specific zip code derived municipalities) that are worthy of greater attention by health professionals.

In future, we would like to extend this Influenza model to account for new and perhaps deadlier influenza strains, such as the ‘Avian’ type, as this type represents a pandemic threat to public health. We look forward to feedback on how to extend the model for these types.

Finally, the complete source files for this case study can be found at:

<http://www.poulinhugin.com/casestudy/Influenza.htm>

Note: The majority (non-graphical) aspects of our results can be duplicated using the Patterns and Predictions™ Trial version found at: <http://www.poulinhugin.com/trial/>

## **Authors**

Chris Poulin, Project Lead and Partner  
Poulin Holdings LLC, P.O. Box 15564, Boston, MA 02215 US  
Phone : +1 617 755 9049, E-mail: [chris@poulinhugin.com](mailto:chris@poulinhugin.com)

Anders Madsen, PhD, CEO  
Hugin Expert A/S, Gasværksvej 5, 9000 Aalborg, Denmark  
Phone : +45 9655 0790, E-mail: [anders@hugin.com](mailto:anders@hugin.com)

## **Citations**

The Economic Impact of Pandemic Influenza in the United States: Priorities for Intervention  
Martin I. Meltzer, Nancy J. Cox, and Keiji Fukuda  
Centers for Disease Control and Prevention, Atlanta, Georgia, USA  
<http://www.cdc.gov/Ncidod/eid/vol5no5/meltzer.htm>

## **Acknowledgements**

Martin I. Meltzer, MS, Ph.D.  
CDC/CCID/NCPDCID/DEISS  
Mailstop D-59, 1600 Clifton Rd. Atlanta, GA 30333

Chris Van Beneden, MD, MPH  
Respiratory Diseases Branch  
Centers for Disease Control and Prevention, Atlanta, GA

Gregory Peterson, Esq.  
20 Dean Road  
Wellesley, MA 02481

# # #