**A study of the transferability of influenza case detection systems between two large healthcare systems**

**–supplementary material–**

Ye Ye1,2, Michael M. Wagner1,2, Gregory F. Cooper1,2, Jeffrey P. Ferraro3,4, Howard Su1, Per H. Gesteland3,4,5, Peter J. Haug3,4, Nicholas E. Millett1, John M. Aronis1, Andrew J. Nowalk6, Victor M. Ruiz1, Arturo López Pineda7, Lingyun Shi1, Rudy Van Bree4, Thomas Ginter8, Fuchiang Tsui1,2\*

1. Real-time Outbreak and Disease Surveillance Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America. 2. Intelligent Systems Program, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America. 3. Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, United States of America. 4. Intermountain Healthcare, Salt Lake City, Utah, United States of America. 5. Department of Pediatrics, University of Utah, Salt Lake City, Utah, United States of America. 6. Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, Pittsburgh, Pennsylvania, United States of America. 7. Department of Genetics, Stanford University School of Medicine, Stanford, California, United States of America. 8. VA Salt Lake City Healthcare System, Salt Lake City, Utah, United States of America

**\* Corresponding author**

E-mail: tsui2@pitt.edu

# **S1 Text. Pseudocode of Greedy Feature Selection Wrapper with K2.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Greedy feature selection wrapper with K2** (*originalTrain*, *maxFeatureSize*, *threshold*)

**INPUT**:

* *originalTrain* is the training dataset with the following order: *age group*, *diagnosis*, *finding* with the highest information gain score, …, *finding* with the lowest information gain score.
* *maxFeatureSize* is the number of features in the *originalTrain* file.
* *threshold* determines whether an increase of AUC is significant.

*(We used 0.0001 in our experiment.)*

**OUTPUT**:

* *BN-Greedy-K2* is the final Bayesian network

*temp\_train* = {*diagnosis*, *age group* of encounters in training dataset}

*best\_train* = *temp\_train*

*featureList* = {*age group*}

*currentAUC* = 0

*lastAUC* = 0

**FOR** k from 1 to 10 /\*10-fold cross validation\*/

 {*temp\_train\_fold\_k*, *temp\_validation\_fold\_k*} = Sampling k th fold from *temp\_train*

 *BN\_temp\_train\_fold\_k* is the Bayesian network learned from *temp\_train\_fold\_k* with K2

 *tempAUC* = AUC of *BN\_temp\_train\_fold\_k* when testing with *temp\_validation\_fold\_k*

 *currentAUC* = *currentAUC* + *tempAUC*

**END FOR**

*currentAUC* = *currentAUC* / 10

*lastAUC* = *currentAUC*

**FOR** n from 3 to *maxFeatureSize*

 *temp\_train* = *best\_train* + nth featureof the *originalTrain* file

 *increase* = false

 *currentAUC* = 0

 **FOR** k from 1 to 10 /\*10-fold cross validation\*/

 {*temp\_train\_fold\_k*, *temp\_validation\_fold\_k*} = Sample k th fold from *temp\_train*

 *BN\_temp\_train\_fold\_k* is the Bayesian network learned from *temp\_train\_fold\_k* with K2

 /\* The configuration of K2 classifier is: maximum 2 parents, ordering: *age group*, *diagnosis*, finding with

 highest information gain score, …, finding with lowest information gain score.\*/

 *tempAUC* = AUC of *BN\_temp\_train\_fold\_k* when testing with *temp\_validation\_fold\_k*

 *currentAUC* = *currentAUC* + *tempAUC*

 **END FOR**

*currentAUC* = *currentAUC* / 10

 **IF** *currentAUC* - *lastAUC* >= *threshold*

 *increase*=true

 *lastAUC* = *currentAUC*

 *best\_train* = *temp\_train*

 *featureList* = featureList + {nth feature}

 **END IF**

**END FOR**

*BN-Greedy-K2* = the Bayesian network learned from *best\_train* with K2 algorithm

**RETURN** *BN-Greedy-K2*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*