I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13099

Using Machine Learning to Predict Nosocomial Infections and Medical Accidents in a NICU

Marc Beltempo
Georges Bresson
Guy Lacroix

MARCH 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13099

Using Machine Learning to Predict Nosocomial Infections and Medical Accidents in a NICU

Marc Beltempo
McGill University Health Centre

Guy Lacroix
Université Laval and IZA

Georges Bresson
Université Paris II

MARCH 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Using Machine Learning to Predict Nosocomial Infections and Medical Accidents in a NICU

Background: Adult studies have shown that nursing overtime and unit overcrowding is associated with increased adverse patient events but there exists little evidence for the Neonatal Intensive Care Unit (NICU). Objectives: To predict the onset on nosocomial infections and medical accidents in a NICU using machine learning models. Subjects: Retrospective study on the 7,438 neonates admitted in the CHU de Québec NICU (capacity of 51 beds) from 10 April 2008 to 28 March 2013. Daily administrative data on nursing overtime hours, total regular hours, number of admissions, patient characteristics, as well as information on nosocomial infections and on the timing and type of medical errors were retrieved from various hospital-level datasets. Methodology: We use a generalized mixed effects regression tree model (GMERT) to elaborate predictions trees for the two outcomes. Neonates' characteristics and daily exposure to numerous covariates are used in the model. GMERT is suitable for binary outcomes and is a recent extension of the standard tree-based method. The model allows to determine the most important predictors. Results: DRG severity level, regular hours of work, overtime, admission rates, birth weight and occupation rates are the main predictors for both outcomes. On the other hand, gestational age, C-Section, multiple births, medical/surgical and number of admissions are poor predictors. Conclusion: Prediction trees (predictors and split points) provide a useful management tool to prevent undesirable health outcomes in a NICU.

JEL Classification: I1, J2, C11, C14, C23

Keywords: neonatal health outcomes, nursing overtime, machine learning, mixed effects regression tree

Corresponding author:
Guy Lacroix
Department of Applied Economics
HEC Montréal 3000
Chemin de la Côte-Sainte-Catherine
Montréal (Québec)
Canada H3T 2A7

E-mail: guy.lacroix@hec.ca

INTRODUCTION

Neonatal intensive care units (NICUs) must contend with ever changing caseloads, patient mix and unplanned admissions (Tucker et al. [1999]). Workforce management is thus challenging and nursing overtime is often used to meet required nurse-to-patient ratios (Berney and Needleman [2005], Beltempo et al. [2016]). The increasing use of overtime hours as a labor management strategy has become an important issue across NICUs in Canada (Canadian Association of Paediatric Health Care Centers [2013]) and elsewhere (Griffiths et al. [2014]). Indeed, nursing overtime has been found by some to be deleterious to adult patients' health (Bae [2013], Lin [2014], Cimiotti et al. [2012], Dorrian et al. [2006], Trinkoff et al. [2011]), although others have concluded otherwise (Cook et al. [2012], Evans and Kim [2006], Berney and Needleman [2005, 2006], Duffield et al. [2011]).

The lack of clear evidence linking overtime and patient health may be due to methodological factors (Bae and Favry [2013], Weinstein et al. [2008]). Indeed, most studies use cross-sectional data and contrast health outcomes stemming from heterogeneous units and/or hospitals. Such analyses are likely to omit important unobserved patient characteristics and unit-specific work arrangements. As for NICUs, given that the mix of neonatologists, fellows, residents, nurse practitioners, *etc.* varies greatly across hospitals, singling out the contribution of overtime on the health outcomes of neonates is clearly a difficult task.

In this paper, we focus on the CHU de Quebec NICU, a tertiary/quaternary referral center with a 51-bed capacity that tends to a population of 1.7 million over a territory of 452,600 km^2 in Canada. Focusing on a single unit removes some of the aforementioned variations in specialty mix across NICUs. We study the occurrence of health care associated infections and medical incidents/accidents (henceforth HCAI and MA, respectively) among all neonates admitted to the NICU between April 2008 and March 2013. Daily exposure to overtime and regular hours of work, as well as numerous individual and NICU-specific covariates are used to predict the onset of the latter two outcomes. The analysis is based upon a generalized classification and regression tree approach.

DATA

The CHU de Quebec NICU is a Level-III referral center. At the time of the study, a total of 165 registered nurses were employed in the NICU, of which 68% (n=112) worked 8-hour shifts and 32% (n = 53) worked on 12-hour shifts. Nurse staffing was determined before each shift according to patient acuity, planned admissions, and electives procedures/tests. When nurses were deemed in shortage, management initially turned to available off-duty nurses. Next, a pool of floating nurses was relied upon. Finally, it resorted to voluntary and mandatory overtime if necessary.

Overtime is defined as all hours worked beyond the regular work schedule (Fédération Interprofessionnelle de la Santé du Québec [2011]).[1] Patient characteristics were collected using the hospital clinical database Med-Echo. It included gestational age, birth weight, sex, Apgar score, multiple pregnancies and type of delivery. Daily administrative data on overtime and regular hours of work, daily patient census and number of admissions were collected using the local administrative database Logibec. Information on HCAI was collected using the local infectious disease database TDR. Finally, information on MA was retrieved from the Gesrisk database.[2]

[1]This occurred whenever a nurse either started her shift earlier than planned or finished later than scheduled. Working beyond 16 consecutive hours per day was forbidden.

[2]Reporting the information on the timing as well as the type of MA is mandatory.

All newborns admitted during the study period were included conditional on having spent at least three days in the NICU. If an infant had more than one episode of bacteremia, these were considered separate events if they occurred more than 14 days apart. The date of the infection was determined as that at which the blood culture was obtained.

Descriptive Statistics

The total number of infants admitted in the NICU over our sample period is 7,438. Of those, 1,972 were omitted since their stay was shorter than three days. The final sample thus includes 5,466 neonates and represents over 101,621 infant/days over the sample period.[3] Table 1 provides the means of the main variables used in the models. Infants who contracted an infection or suffered a medical accident had either, or both, a lower gestational age and birth weight, and were more likely to have been delivered by C-Section (infection). The table also shows that approximately 6.7% (resp. 4.1%) of neonates were victim of a MA (resp. HCAI). From the NICU's point of view, the probability of observing a MA or a HCAI in any given day was 27.6% and 15.4%, respectively. This translates into 0.58% and 0.32% when computed daily and per neonate. The average length of stay is also considerably longer for neonates with either a MA or HCAI. In June 2012, management implemented a series interventions aimed at reducing overtime (*Overtime Reform*). In particular, it hired 15 full-time registered nurses and converted 10% of existing positions from 8-hour shifts to 12-hour shifts (Beltempo et al. [2016]). The table shows that proportionately more MA occurred in the months that followed the reform relative to either the population of neonates or to those with a HCAI. The table also shows that regular and overtime hours of work are positively related to the occurrence of either outcome. The relationship between hours of work and outcomes is further investigated in Figure 1 in which we depict the (smoothed) daily variations in overtime and regular hours, respectively. The figure highlights the negative correlation between regular hours of work and overtime. Further, peaks and troughs almost always mirror one another. The bottom panel depicts the daily frequencies of HCAI and accidents. Taken together, the three panels provide *prima facie* evidence that these events may be loosely related to hours of work. Yet the influence of other variables needs to be netted out in order to determine the precise link between work schedules and medical outcomes, if any. It is highly likely that the link between health outcomes, neonates' characteristics and work arrangements are fairly complex. Machine learning methods are particularly well-suited to analyze such complex interactions.

METHOD

Classification and regression trees (CART) are machine-learning methods used to construct decision trees. A decision tree is a is flowchart-like predictive model in which leaves represent classifications (outcomes), non-leaf nodes are predictors, and branches represent conjunctions of features that lead to the classifications. CART models involve selecting the best predictors and determining appropriate split points (nodes) in each of them.

The data at our disposal present a novel and interesting feature. Indeed, the outcomes (MA/HCAI) are observed daily at the neonate level, whereas work arrangements and NICU characteristics are observed daily. Since the time spent in the NICU varies across neonates, our panel dataset is unbalanced. Traditional CART models are ill-suited to analyze such data. Fortunately, Hajjem et al. [2017] have recently

[3]The appendix provides further details on the data and statistical tests.

Table 1 Sample Means

Variable	Accident Infection/ No	Accident Yes	Infection Yes
		NEONATES	
Sex (Female=1)(%)	43.84	48.13	43.19
Gestational Age	35.47	32.31	29.81
Weight (Grams)	2556.70	1962.64	1457.69
Apgar $>$ 7 at 5 Min. (%)	88.89	71.78	65.73
C-Section (%)	40.19	54.77	62.91
DRG Severity	2.15	3.13	3.41
First Birth (%)	73.74	78.84	85.45
DRG (% Surgical)	6.35	31.95	32.86
Length of stay (Days)	17.54	55.96	70.87
Overtime Reform (%)	16.42	21.83	15.32
Total (MA/HCAI)		490	274
Neonate Frequency (%)		6.66	4.11
Daily Frequency (%)		27.62	15.44
Daily/Infant Frequency (%)		0.575	0.322
		NICU	
Daily Admissions	4.32	4.49	4.47
Bed Occupancy	50.19	50.71	50.58
Daily Regular Hours	514.14	538.03	517.13
Daily Overtime Hours	20.80	25.78	26.74

proposed the Generalized Mixed-Effects Regression Tree (GMERT) which is suitable for unbalanced binary outcomes such as ours. In addition, GMERT allows the inclusion of random effects to account for neonate-specific unobserved variables that may impact their outcomes. As with most machine learning methods, the GMERT algorithm splits the sample into three subsets: a learning set, a validation set and a test set. The appropriate number of leaves is determined by cross-validation. [4] Finally, the dimensionality of the tree is reduced by ranking the predictors according to a so-called variable importance indicator (VIMP). The VIMP is a normalized score between 0% and 100% and variables with low values are removed from the tree since they have little predictive power. We thus use GMERT to predict the binary outcome y_{it} (*MA or HCAI*) of neonate i during its t-th day in the NICU, conditional on a series of predictors.

RESULTS

GMERT is used to predict the two outcomes (MA and HCAI) separately. The algorithm is fed the following common set of predictors for both outcomes: (neonate characteristics) *weight*, *gestational age*, *mother's first birth*, *sex*, *c-section*, *twins*, *Severity of Illness Index (1-4)*, *DRG Surgical/Medical*, *Apgar < 7 at 5 min.*, (NICU characteristics) *bed occupancy*, *# admissions*, *Reform*. The analysis of MA further includes daily regular and overtime hours in the NICU as well as a dummy variable equal to one if the infant incurred a HCAI prior to the MA. The analysis of HCAI assumes the infection occurred when results from the blood culture were obtained. Although there is no consensus on the lead time and

[4] The appendix provides additional details on GMERT.

sequence of events that may trigger the onset of a HCAI, we include the daily prior three-day moving average number of regular and overtime hours in the analysis (*Regular.Hours.MA3, Overtime.Hours.MA3)* (see (Hugonnet et al. [2007], Polin et al. [2012]). Finally, we include a dummy variable to account for the occurrence of a MA prior to the HCAI.

It it conceivable that individual characteristics that are unfavorable to the occurrence of a MA/HCAI may be compensated for by detrimental environmental factors, and *vice versa*. In addition, the levels at which the favorable/detrimental factors operate are not know *a priori* and may in fact interact in a highly non-linear fashion. Figures 2a and 2b report the main variables found to influence the occurrence of both outcomes. The predictors are arranged in descending order of importance. While GMERT includes as many as 15 predictors in each outcome, the algorithm retains only eleven of them when predicting either MA or HCAI. Furthermore, there are almost as many environmental factors as there are individual characteristics that play an important role in predicting the occurrence of either outcome. In fact, regular hours of work and overtime are two of the three most influencial predictors. Regular hours ranks first in predicting MA and second for HCAI. Not surprisingly, the algorithm identifies birth weight and gestational age as the main infant-specific drivers of the health outcomes. Predictors that have little predictive power are omitted in the pruning process as stressed above and do not appear in the figures. The average relative variable importance is equal to 25.34% in the MA model (dashed line) and 44.84% in the HCAI model. Figures 2a and 2b identify the main predictor variables in the entire MA and HCAI trees. Although *hours of work* and *overtime* stand out as the main variables, it does not necessarily follow that the main split nodes occur along these predictors. We investigate this issue further by focusing on the prediction trees for MA and HCAI as exhibited in Figures 3 and 4, respectively. The first number inside each node corresponds to the predicted probability of reaching the latter.[5] The number of observations (n) in appears just below the node. Finally, the splitting values appear in bold characters below the nodes. To ease reading, paths that lead to high probabilities of adverse outcomes are emphasized (> 75%).

The MA tree is illustrated in Figure 3. It contains 32 leaves. The primary split occurs with respect to the DRG *severity* level. Thus neonates who were deemed as having a "High" or below DRG appear on the main left-hand side branch. Those with a "Very High" DRG are located on the right-hand side branch. The second split involves Medical/Surgical intervention (left) and infection prior to the MA (right). Hence the first two levels involve neonate-specific factors. The ensuing paths involve a mixture of neonate and NICU specific factors. The tree contains six left-hand paths which likely lead to a MA and as many as seven on the right-hand side. The paths show different combinations of predictor variables and different split thresholds.

The HCAI tree is illustrated in Figure 4. It contains 31 leaves. The primary split occurs with respect to birth weight at a threshold of 1195 grams. The second level splits occur on the left-hand side with respect to DRG severity level (Low/Medium), and with respect of regular hours of work (threshold equal to 618) on the right-hand side. The tree depicts thirteen paths which likely lead to a HCAI, seven on the left-hand side and six on the right-hand.

While MA and HCAI are relatively rare events at the neonate/day level, the classification trees depicted in Figures 3 and 4 manage to identify the routes through which these are likely to occur. In addition, the split nodes and their thresholds values are sensical and provide useful guidance to management when designing policies to mitigate these deleterious outcomes.

The GMERT algorithm manages to unearth two important features of the data. First, surprisingly,

[5]To save on space, probabilities smaller than 1% appear as "0" inside the nodes.

it shows that institutional features are just as important drivers as neonate-specific medical conditions in predicting MA and HCAI, and hence within the grasp of management. Second, traditional statistical analyses are assuredly incapable of identifying the highly non-linear relations between these predictors.

CONCLUSION

Predicting HCAI or accidents in a NICU is a complex task. Recent machine learning algorithms are well-suited to unearth potential correlations as they now allow to account for unbalanced panel data and discrete health outcomes, two frequent features of clinical data. Prediction trees are now relatively widespread in empirical research and are integrated in many software suites. From an operational point of view, prediction trees can complement traditional management tools in preventing undesirable health outcomes in the NICU.

REFERENCES

Sung-Heui Bae. Presence of nurse mandatory overtime regulations and nurse and patient outcomes. *Nursing Economic$*, 31(2):59–89, 2013.

Sung-Heui Bae and Donna Favry. Assessing the relationships between nurse work hours/overtime and nurse and patient outcomes: Systematic literature review. *Nursing Outlook*, 62(2):138–156, 2013.

Marc Beltempo, Guy Lacroix, Michele Cabot, and Bruno Piedboeuf. Factors and costs associated with the use of registered nurse overtime in the neonatal intensive care unit. *Pediatrics and neonatal nursing Open Journal*, 4:17–23, 08 2016.

B. Berney and J. Needleman. Trends in nurse overtime, 1995-2002. *Policy Polit Nurs Pract*, 6:183–90, 2005.

Barbara Berney and Jack Needleman. Impact of nursing overtime on nurse-sensitive patient outcomes in New-York hospitals, 1995-2000. *Policy, Politics, & Nursing Practice*, 7(2):87–100, 2006.

Canadian Association of Paediatric Health Care Centers. Benchmarking report 2013. October 2013.

Jeannie P. Cimiotti, Linda H. Aiken, Douglas M. Sloane, and Evan S. Wu. Nurse staffing, burnout, and health care-associated infection. *American Journal of Infection Control*, 40(6):486–490, 2012.

Andrew Cook, Martin Gaynor, Melvin Stephens Jr, and Lowell Taylor. The effect of a hospital nurse staffing mandate on patient health outcomes: Evidence from California's minimum staffing regulation. *Journal of Health Economics*, 31(2):340–348, 2012.

Jillian Dorrian, Nicole Lamond, Cameron van den Heuvel, Jan Pincombe, Ann E. Rogers, and Drew Dawson. A pilot study of the safety implications of Australian nurses' sleep and work hours. *Chronobiology International*, 23(6):1149–1163, 2006.

Christine Duffield, Donna Diers, Linda O'Brien-Pallas, Chris Aisbett, Michael Roche, Madeleine King, and Kate Aisbett. Nursing staffing, nursing workload, the work environment and patient outcomes. *Applied Nursing Research*, 24(4):244 – 255, 2011.

William N. Evans and Beomsoo Kim. Patient outcomes when hospitals experience a surge in admissions. *Journal of Health Economics*, 25(2):365–388, 2006.

Fédération Interprofessionnelle de la Santé du Québec. Convention collective 2011-2015, article 19.01. 2011.

P. Griffiths, C. Dall'Ora, M. Simon, J. Ball, R. Lindqvist, A. M. Rafferty, et al. Nurses' shift length and overtime working in 12 european countries: the association with perceived quality of care and patient safety. *Med Care*, 52(11):975–981, 2014.

Ahlem Hajjem, François Bellavance, and Denis Larocque. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328, 2014.

Ahlem Hajjem, Denis Larocque, and François Bellavance. Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126:114 – 118, 2017.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: prediction, inference and data mining*. Springer-Verlag, New York, 2009.

Stéphane Hugonnet, Jean-Claude Chevrolet, and Didier Pittet. The effect of workload on infection risk in critically ill patients. *Critical Care Medicine*, 35(1):76–81, 2007.

Haizhen Lin. Revisiting the relationship between nurse staffing and quality of care in nursing homes: An instrumental variables approach. *Journal of Health Economics*, 37:13 – 24, 2014.

Richard A. Polin, Susan Denson, Michael T. Brady, and . Strategies for prevention of health care–associated infections in the nicu. *Pediatrics*, 129(4):e1085–e1093, 2012.

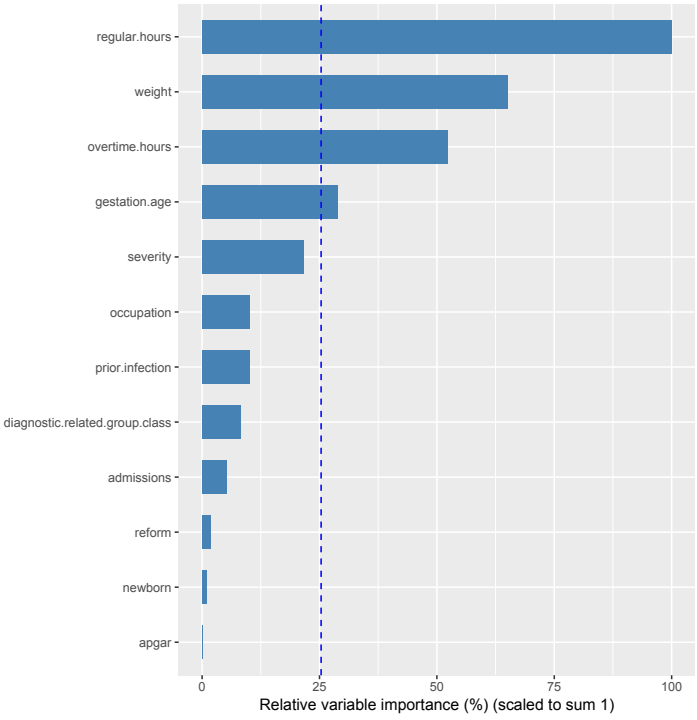
Alison M. Trinkoff, Meg Johantgen, Carla L. Storr, Ayse P. Gurses, Yulan Liang, and Kihye Han. Nurses' work schedule characteristics, nurse staffing, and patient mortality. *Nursing Research*, 60(1):1–8, 2011.

J. Tucker, W. Tarnow-Mordi, C. Gould, G. Parry, and N. Marlow. On behalf of the UK neonatal staffing study collaborative group. uk neonatal intensive care services in 1996. *Child Fetal Neonatal Ed*, 80: F233–34, 1999.

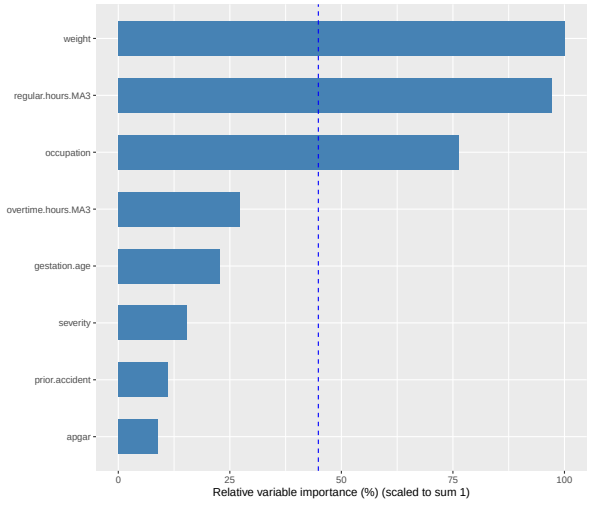
Robert A. Weinstein, Patricia W. Stone, Monika Pogorzelska, Laureen Kunches, and Lisa R. Hirschhorn. Hospital staffing and health care-associated infections: A systematic review of the literature. *Clinical Infectious Diseases*, 47(7):937–944, 2008.

Figure 1 Smoothed Daily Variations in Nursing Hours of Work, and NICU Daily # of HCAI and Nosocomial Infections

Figure 2 Relative Variable importance (%)
Generalized Mixed-Effect Regression Trees



(a) Accidents



(b) Infections

Figure 3 Generalized Mixed-Effects Regression Tree - Accidents

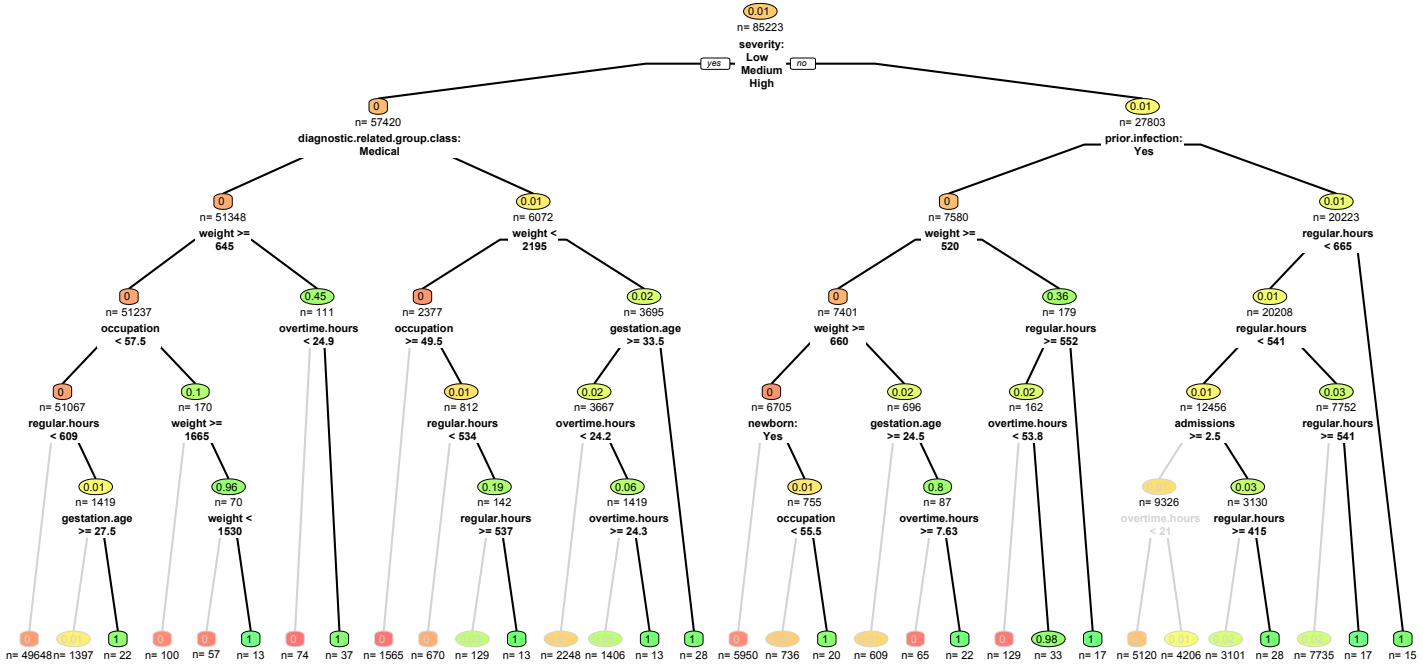
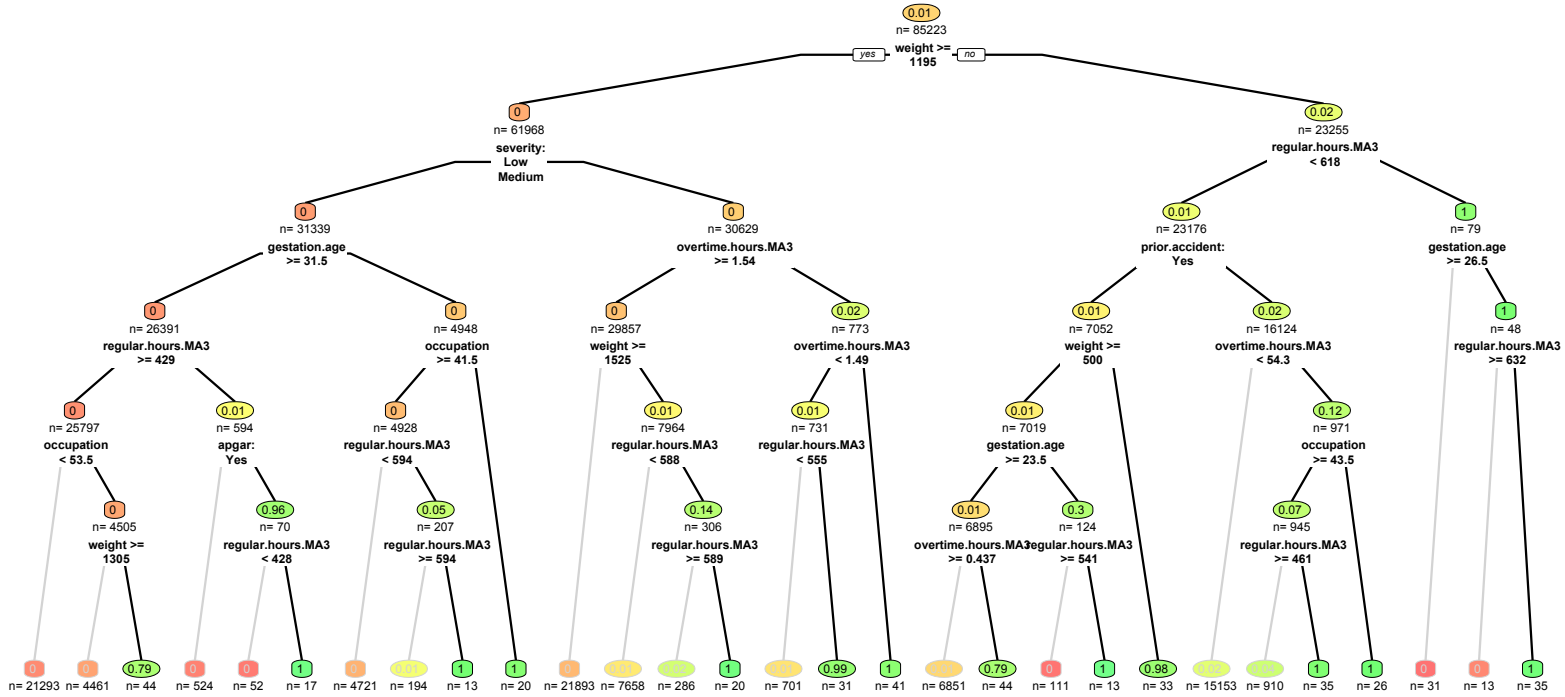


Figure 4 Generalized Mixed-Effects Regression Tree - Infections



APPENDIX

Generalized mixed effects regression tree with random intercept

As stated above, Hajjem et al. [2014] have proposed a mixed effects regression tree (MERT) method. It can appropriately deal with the possible random effects of observation-level covariates and can split observations within clusters since observation-level covariates are candidates in the splitting process. The main idea is to fit a tree after removing the random effects part of the model, update the estimates (or predictions) of the random effect and cycle until convergence. But MERT is designed for Gaussian response data. More recently, Hajjem et al. [2017] have proposed a tree based method, named "generalized mixed effects regression tree" (GMERT), which is suitable for non-Gaussian data (e.g., binary outcomes and count data). Following the steps of the generalized linear mixed models (GLMMs), the GMERT method can handle unbalanced clusters, and can incorporate observation-level covariates and their potential random effects. It allows observations within clusters to be split. For an unbalanced panel data set of infants ($i = 1, ..., N$) followed during T_i days ($t = 1, ..., T_i$), let y_{it} be the binary outcome variable and X_{it}, the $(1 \times K)$ vector of predictors. GMERT method includes random individual specific effects and starts with a logistic-mixed model

$$y \mid \beta, u \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(X\beta + Zu\right)\right)$$

where the notation $y \sim \text{Bernoulli}\,(p)$ is shorthand for the entries of y having independent Bernoulli distributions with parameters corresponding to the entries of p and $\text{logit}^{-1}\,(x)$ is shorthand for the logistic distribution $e^x/(1+e^x)$. y is the $(T \times 1)$ vector of binary outcomes, X is an $(T \times K)$ matrix of covariates, Z is an $(T \times NK)$ block-diagonal matrix of the X_i submatrices where $T = \sum_{i=1}^{N} T_i$. X and Z are called the fixed and random effects design matrices associated with β and u, the $(K \times 1)$ fixed effects and $(NK \times 1)$ random effects vectors. $X_{it,1}$ is the intercept and $X_{it,j}$, $2 \leq j \leq K$ are the other control covariates. The random intercept is defined by the sum $(\beta_1 + u_{i,1})$, the random slope for variable $X_{i,2}$ is the sum $(\beta_2 + u_{i,2})$, etc.

In our specific case, we only consider random intercept (*i.e.*, $u_{i,j} = 0, \forall j > 1$). So, Z is restricted to an $(T \times N)$ block-diagonal matrix of N subvectors $(T \times 1)$ of ones. Then, the GMERT reduces to a generalized mixed effects regression tree with random intercept (GMERT-RI). In the GMERT, proposed by Hajjem et al. [2017], the linear fixed part $X\beta$ is replaced by a function $f\,(X)$ which is estimated using a standard regression tree design (STD).[6] The overall data set is divided into three subsets: a learning subset, a validation subset and a test subset (corresponding respectively to 40%, 40% and 20% of the initial dataset).

A standard tree design (STD) — that does not account for random specific effects — is generally adjusted using the dataset resulting from the merge of the learning and validation subsets. To choose the right number of leaves, one proceeds by cross validation. We estimate the tree's performance by 10-block cross-validation for each level of relevant simplification. The complexity of the decision tree is defined as the number of splits in the tree and the complexity parameter (CP) is used to control the size of the decision tree and to select the optimal tree size. A good choice of CP for pruning the tree is often the leftmost value for which the cross-validation error (*i.e.*, the rate of misclassifications relative to the original score) lies below the "horizontal line".[7] Here, GMERT-RI is adjusted using the learning subset,

[6] The GMERT algorithm is detailed below. The R codes are available in the supplementary material of Hajjem et al. [2017]

[7] which represents the highest cross validation error less than the sum of the minimum cross validation error and the standard deviation of the error on that tree.

and then validated using the validation subset, *i.e.* the best GMERT-RI is selected based on minimum misclassification rate (MCR) observed on the validation subset.[8] The GMERT algorithm of Hajjem et al. [2017] is detailled below.

To select variables and reduce dimensionality, we can rank the predictors by some measure of importance and remove variables with low rank. Variable importance (VIMP) was originally defined using a measure involving surrogate variables (see for instance Hastie et al. [2009]). VIMP is calculated for each variable individually and the value is calculated as the sum of the decrease in impurity[9], it counts both when the variable appear as a primary split and when it appears as a surrogate. The relative variable importance is a number between 0 and 100%. For each variable, it is the VIMP of this variable divided by the maximum VIMP among all the variables. Then it is transformed into percentage scoring, the highest values as 100 and consecutively proportional until the lower values.

The GMERT algorithm

Recall that for the generalized mixed model (GLMM),

- $y_{it} \mid u_i$ belongs to the exponential family of distribution.
- $\mu_{it} = E\left[y_{it} \mid u_i\right]$ and $g\left(\mu_{it}\right) = \eta_{it} = X_{it}\beta + Z_{it}u_i$ for some known link function g. In our case, we use the logit link. $\mu_{it} = \frac{e^{\eta_{it}}}{1+e^{\eta_{it}}}$ and $g\left(\mu_{it}\right) = \log\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \eta_{it}$. So, $\mu_i = E\left[y_i \mid u_i\right]$ and $g\left(\mu_i\right) = \eta_i = X_i\beta + Z_iu_i$ with $u_i \sim N\left(0, \Sigma\right)$.
- $Cov\left[y_i \mid u_i\right] = \sigma^2 v\left(\mu_i\right)$ where σ^2 is a dispersion parameter and $v\left(\mu_i\right) = diag\left[v_{i1}, ..., v_{iT_i}\right] = diag\left[v\left(\mu_{i1}\right), ..., v\left(\mu_{iT_i}\right)\right]$ where $v\left(.\right)$ is a known variance function.

The generalized mixed effects regression tree (GMERT) model, proposed by Hajjem et al. [2014], can be written as $\eta_i = f\left(X_i\right) + Z_iu_i$ with $u_i \sim N\left(0, \Sigma\right)$ where the linear fixed part $X_i\beta$ is replaced by the function $f\left(X_i\right)$ that will be estimated with a standard regression tree model. A first-order Taylor-series expansion yields the linearized response variable, $\widetilde{y}_i = g\left(\mu_i\right) + \left(y_i - \mu_i\right)g'\left(\mu_i\right)$ and the mixed fixed effect regression tree (MERT) pseudo-model is defined as follows: $\widetilde{y}_i = f\left(X_i\right) + Z_iu_i + e_i$. The GMERT algorithm is basically the penalized quasi-likelihood (PQL) algorithm used to fit GLMMs where the weighted linear mixed effects (LME) pseudo-model is replaced by a weighted MERT pseudo-model. Therefore, the fixed-part $f\left(X_i\right)$ is estimated with a standard regression tree model. The GMERT algorithm of Hajjem et al. [2017] is the following:

[8] The misclassification rate (MCR) is given by $MCR = \left(\sum_{i=1}^{N^{(v)}}\sum_{t=1}^{T_i^{(v)}} \mid y_{it} - \widehat{y}_{it} \mid\right)/T^{(v)}$ where $\widehat{y}_{it} = \left(1 + \exp\left(-\widehat{f}(X'_{it}) - Z'_{it}\widehat{u}_i\right)\right)^{-1}$. $\widehat{y}_{it}$ is the predicted probability that $y_{it} = 1$. $\widehat{f}(X'_{it})$ is the predicted fixed component that results from the tree and $Z'_{it}\widehat{u}_i$ is its predicted random part corresponding to its cluster. $N^{(v)}$ is the number of clusters in the validation set, $T_i^{(v)}$ is the size of cluster i and $T^{(v)}$ is the total number of observations in the validation set.

[9] Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The Gini impurity can be computed by summing the probability p_i of an item with label i being chosen times the probability $(1 - p_i)$ of a mistake in categorizing that item. To compute Gini impurity I_G for a set of items with J classes, suppose $i \in \{1, 2, ..., J\}$ and let p_i be the fraction of items labeled with class i, then $I_G = 1 - \sum_{i=1}^{J} p_i^2$. In our case, $J = 2$ for accident (1) and no accident (0).

ALGORITHM 1. GMERT ALGORITHM (see Hajjem et al. [2017], p. 115):

1. **Initialization step.** Set $D = 0$. Initialize mean values $\widehat{\mu}_{it}^{(0)}$, $t = 1, ..., T_i$ and fit a weighted linear mixed effect (LME) pseudo-model using the linearized pseudo-responses, $\widetilde{y}_i^{(0)} = g\left(\widehat{\mu}_i^{(0)}\right) + \left(y_i - \widehat{\mu}_i^{(0)}\right) g'\left(\widehat{\mu}_i^{(0)}\right)$. Define the weights, $W_i^{(0)} = diag\left(w_{it}^{(0)}\right)$ where $w_{it}^0 = \left(v_{it} g'\left(\widehat{\mu}_{it}^{(0)}\right)^2\right)^{-1}$. Set $d = 1$ and let $\widehat{\sigma}_{(0)}^2$ and $\widehat{\Sigma}_{(0)}$ be the estimates of this weighted LME pseudo-model.

2. **Outer loop.** While non-convergence of $\widehat{\eta}_i$, do:

   2.1. **Inner loop.** While non-convergence of the generalized log-likelihood (GLL), set $d = d + 1$ and do:

   2.1.1. Update $\widehat{f}(X_i)$ and u_i using

   2.1.1.1. $\widetilde{y}_{i,(d)}^* = \widetilde{y}_i^{(D)} - Z_i \widehat{u}_{i,(d-1)}$,

   2.1.1.2. Let $\widehat{f}_{(d)}(X_i)$ be an estimate of $f(X_i)$ obtained from a standard regression tree algorithm with $\widetilde{y}_{i,(d)}^*$ as responses, X_i as covariates and W_i as weights, $i = 1, ..., T_i$,

   2.1.1.3. $\widehat{u}_{i,(d)} = \widehat{\Sigma}_{(d-1)} \left(W_i^{\frac{1}{2}(D)} Z_i\right)' \widehat{V}_{i,(d-1)}^{-1} \left(W_i^{\frac{1}{2}(D)} \widetilde{y}_i^{(D)} - W_i^{\frac{1}{2}(D)} \widehat{f}_{(d)}(X_i)\right)$ where $\widehat{V}_{i,(d-1)} = W_i^{\frac{1}{2}(D)} Z_i \widehat{\Sigma}_{(d-1)} \left(W_i^{\frac{1}{2}(D)} Z_i\right)' + \widehat{\sigma}_{(d-1)}^2 I_{T_i}$, $i = 1, ..., N$.

   2.1.2. Update the $\widehat{\sigma}^2$ and $\widehat{\Sigma}$ using

$$\begin{aligned}
\widehat{\sigma}_{(d)}^2 &= \frac{1}{NT} \sum_{i=1}^{N} \left(\widehat{\varepsilon}_{i,(d)}' \widehat{\varepsilon}_{i,(d)} + \widehat{\sigma}_{(d-1)}^2 \left[T_i - \widehat{\sigma}_{(d-1)}^2 \text{Trace}\left(\widehat{V}_{i,(d-1)}\right) \right] \right) \\
\widehat{\Sigma}_{(d)} &= \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{u}_{i,(d)} \widehat{u}_{i,(d)}^{R'} + \left[\widehat{\Sigma}_{(d-1)} - \widehat{\Sigma}_{(d-1)} \left(W_i^{\frac{1}{2}(D)} Z_i\right)' \widehat{V}_{i,(d-1)}^{-1} W_i^{\frac{1}{2}(D)} Z_i \widehat{\Sigma}_{(d-1)} \right] \right)
\end{aligned}$$

   where $\widehat{\varepsilon}_{i,(d)} = W_i^{\frac{1}{2}(D)} \left(\widetilde{y}_i^{(D)} - \widehat{f}_{(d)}(X_i) - Z_i \widehat{u}_{i,(d)}\right)$.

   2.1.3. Update the generalized log-likelihood (GLL) value using

$$\begin{aligned}
GLL\left(f(X, u \mid y)\right) = \sum_{i=1}^{N} \Big\{ & \widehat{\varepsilon}_{i,(d)}' \left(\widehat{\sigma}_{(d)}^2 I_{T_i}\right)^{-1} \widehat{\varepsilon}_{i,(d)} + \\
& \widehat{u}_{i,(d)}^{R'} \left(\widehat{\Sigma}_{(d)}\right)^{-1} \widehat{u}_{i,(d)} + \log \mid \widehat{\Sigma}_{(d)} \mid + \log \mid \widehat{\sigma}_{(d)}^2 I_{T_i} \mid \Big\}
\end{aligned}$$

3. **Updating step.** Set $D = D + 1$. Update $\widehat{\eta}_i$, $\widehat{\mu}_i$, $\widetilde{y}_i$, w_{it} and W_i using $\widehat{\eta}_i^{(D)} = \widehat{f}_d(X_i) + Z_i \widehat{u}_{i,(d)}$, $\widehat{\mu}_i^{(D)} = g^{-1}\left(\widehat{\eta}_i^{(D)}\right) = \left(1 + \exp(-\widehat{\eta}_i^{(D)})\right)^{-1}$, $\widetilde{y}_i^{(D)} = g\left(\widetilde{\mu}_i^{(D)}\right) + \left(y_i - \widetilde{\mu}_i^{(D)}\right) g'\left(\widetilde{\mu}_i^{(D)}\right)$, $w_{it}^{(D)} = \left(v_{it} g'\left(\widehat{\mu}_{it}^{(D)}\right)^2\right)^{-1}$ and $W_i^{(D)} = diag\left(w_{it}^{(D)}\right)$.

   As shown by Hajjem et al. [2017], the GMERT model can predicted the response for two categories of new observations: one that belongs to a cluster included (resp. not included) in the sample used to fit the model. To predict the response for a new observation from the first category, one uses both its corresponding fixed component prediction $\widehat{f}(X_i)$ and the predicted random part $Z_i \widehat{u}_i$

corresponding to its cluster. This is a cluster-specific estimate. For a new observation from the second category, one can only uses its corresponding fixed component prediction (*i.e.*, the random part is set to 0).