Development and validation of clinical prediction models to diagnose acute respiratory infections in children and adults from Canadian Hutterite communities.

By

Danielle Vuichard Gysin, MD

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Science

in Health Research Methodology

McMaster University

ABSTRACT

Acute respiratory infections (ARI) caused by influenza and other respiratory viruses affect millions of people annually. Although usually self-limiting a more complicated or severe course may occur in previously healthy people but are more likely in individuals with underlying illnesses. The most common viral agent is rhinovirus whereas influenza is less frequent but is well known to cause winter epidemics. In primary care, rapid diagnosis of influenza virus infections is essential in order to provide treatment. Clinical presentations vary among the different pathogens but may overlap and may also depend on host factors.

Predictive models have been developed for influenza but study results may be biased because only individuals presenting with fever were included. Most of these models have not been adequately validated and their predictive power, therefore, is likely overestimated.

The main objective of this thesis was to compare different mathematical models for the derivation of clinical prediction rules in individuals presenting with symptoms of ARI to better distinguish between influenza, influenza A subtypes and entero-/rhinovirus-related illness in children and adults and to evaluate model performance by using data-splitting for internal validation.

Data from a completed prospective cluster-randomized trial for the indirect effect of influenza vaccination in children of Hutterite communities served as a basis of my thesis. There were a total of 3288 first episodes per season of ARI in 2202 individuals and 321 (9.8%) influenza positive events over three influenza seasons (2008-2011). The data set was divided into children under 18 years and adults. Both data sets were randomly split by subjects into a derivation (2/3 of the dataset) and a validation population (1/3 of the dataset). All predictive models were developed in the derivation sets. Demographic factors and the classical symptoms of ARI were evaluated with logistic regression and Cox proportional hazard models using forward stepwise selection applying robust estimators to account for non-independent data and by means of recursive partitioning. The beta coefficients of the independent predictors were used to develop different point scores. These scores were then tested in the validation groups and performance between validation and derivation set was compared using receiver operating characteristics (ROC) curves. We determined sensitivities and specificities, positive and negative predictive values, and likelihood ratios at different cut-points which could reflect test and treatment thresholds.

Fever, chills, and cough were the most important predictors in children whereas chills and cough but not fever were most predictive of influenza virus infection in adults. Performance of the individual models was moderate with areas under the receiver operating characteristic curves between 0.75 and 0.80 for the main outcome influenza A or B virus infection. There was no statistically significant difference in performance between the derivation and validation sets for the main outcome.

The results have shown, that various mathematical models have similar discriminative ability to distinguish influenza from other respiratory viruses. The scores could assist clinicians in their decision-making. However, performance of the models was slightly overestimated due to potential clustering of data and the results would first needed to be validated in a different population before application in clinical practice.

ACKNOWLEDGMENTS

I would like to thank my thesis supervisor, Dr. Mark Loeb, for all his support, guidance and encouraging attitude during my graduate studies and the development of this thesis and for providing me the data set that served as basis for my thesis, Dr. Dominik Mertz for his excellent teaching sessions and helpful advice as well as my other thesis committee members, Dr. Marek Smieja and Eleanor Pullenayegum for sharing their expertise and supporting me during my thesis and my research internship.

I wish to thank my former supervisors from the University Hospital of Basel, Switzerland, Prof. Andreas Widmer and Prof. Manuel Battegay, for supporting me. I would like to express my thanks to the Margarete and Walter Lichtenstein foundation, University of Basel, Switzerland, for their generous funding and thus allowing me to take a break from clinical service and conduct research abroad.

I would also like to thank Ellen Amster for her encouragement and her insightful and knowledgeable advice.

Finally, I wish to thank my beloved husband Christian and my children Linus and Muriel, for their support and for undertaking this adventurous trip from Switzerland to Canada with me.

TABLE OF CONTENTS

Figure legend

Table legend:

Appendix:

Fig. 1 Conceptual framework for symptomatic predictors of influenza versus the common cold:

Exposure:

Predictors of interest:

- Fever
- Fatigue
- Chills
- Headache
- Cough
- runny nose
- sore throat
- myalgia
- runny nose
- sinus problems
- presence of 2 symptoms at the same time: fever plus cough, fever plus sore throat
- presence of 3 symptoms: fever plus cough plus sore throat

Other potential predictors, not systematically recorded:

- nausea
- vomiting
- diarrhoea
- chest pain
- dyspnoea
- sneezing

Surrogate markers:

- WBC (white blood cell) count
- CRP
- IL-6
- 25(OH)vitamin D
- VDR polymorphism

Moderators/modifiers

- Age
- Potential interactions with specific symptom combinations
- Different influenza strain types
- Influenza vaccine

Outcome measures:

1°: Influenza virus infection, laboratory confirmed with rt-PCR
2°: other respiratory viruses

Laboratory measures:

Specific influenza virus A/B PCR (gold standard)
respiratory virus panel PCR from nasal swab/nasopharyngeal swab including:

- Rhino-/Enterovirus
- Respiratory syncytial virus
- Coronavirus
- Parainfluenzavirus
- Human metapneumovirus

Direct antigen test
Blood: serum hemagluttinine inhibition titers for influenza

Potential covariates & confounders:

- Onset of symptoms/suddenness of onset
- Risk groups (high-risk, vaccine group, others)
- Gender
- Allergies/Asthma/cardiopulmonary comorbidities
- Co-infections between influenza and other respiratory viruses → probably need to be excluded? → could do a sensitivity analysis
- Antipyretics before consultation (probably not relevant for the first measurement, has not been evaluated (?)) → if taken but not measured would have diluted the actually measured effect
- Duration of symptoms

Epidemiology: influenza season/off influenza season

≍.

CHAPTER I

1.0 Introduction and background

1.1 Content review

Epidemiology and burden of influenza and non-influenza virus related acute respiratory infections (ARI):

Millions of people suffer from ARI due to either influenza or another respiratory virus annually which significantly impacts health-care costs and workforce productivity [1-4]. In a large population based surveillance study from 2009 to 2012/13, the cumulative incidence of influenza-like illness (ILI) visits at primary care clinics varied according to season between 14.2 and 30.4 per 1000 people with highest rates seen in children aged 0-17 years [5]. Young children have on average 6-8 ARI episodes per year compared to 2-4 episodes in adults [4]. With improved viral diagnostics over the past two decades, the detection rate and diversity of viral pathogens that were identified as causing ARI have substantially increased. The types of virus vary according to the population, age of the host, and season; however, rhinovirus, a picornavirus that includes over 100 serotypes, has consistently been found to account for an estimated annual proportion of 30-50% of ARI cases in contrast to influenza virus which only accounts for 5-15% [4, 6]. Influenza virus A and B belong to the family of orthomyxoviridae. Only influenza A has different subtypes based on serological and genetic differences; subtypes of hemagglutinine proteins H1, H2, and H3 as well as of neuraminidase N1 and N2 have most commonly caused epidemics and pandemics in humans [7]. Winter epidemics of influenza recur annually, though, the exact timing, the severity, and the distribution of circulating subtypes may vary considerably; e.g. rates of influenza-associated hospitalizations were found to be highest in seasons in which influenza

A(H3N2) was predominately circulating [8]. Other viruses such as coronavirus, respiratory syncytial virus (RSV), or parainfluenza virus have their own pattern of seasonal occurrence [4, 6].

Clinical manifestations:

The incubation period after which symptoms arise can be less than 24 hours for rhinovirus but may extend up to 4 days in influenza virus infections [7, 9]. Systemic symptoms associated with ARI include fever, chills, headache, and myalgia, whereas cough, sore throat, hoarseness, stuffed or runny nose, and sinus pain are referred to as respiratory symptoms. It is commonly described that rhinovirus infections start with a sore throat followed by nasal congestion, discharge, sneezing and cough whereas systemic symptoms are less pronounced [10, 11]. An acute onset of fever or general malaise, myalgia, sore throat and cough has traditionally been associated with influenza virus infection [7].

Some pathophysiological mechanism may explain why symptoms of influenza and rhinovirus manifest differently: Influenza virus primarily replicates in the tracheobronchial epithelium and causes damage which could explain that cough is an early symptom in influenza virus infection. Furthermore, pyrexia has been attributed to influenza virus induced cytokine release [4, 12]. In contrast, the nasopharynx is the principal replication site for rhinovirus and as a result of interleukin-8 induced influx of polymorphonuclear cells rhinorrhea is a primary symptom in rhinovirus infection [13]; however, increased interleukin-8 levels were also found in children with RSV and influenza virus infections, which might explain, that rhinorrhea is generally a very frequent symptom in ARI [14]. Eventually, many symptoms seem to overlap and are likely mediated by host factors such as age, comorbidities and the immune system's previous viral exposure [12, 15].

In most instances, influenza is an uncomplicated, self-limiting ARI. Patients at risk for complications or more severe disease include elderly and very young persons, patients with cardiovascular, neuromuscular comorbidities, obese and also previously healthy people [16, 17]. Non-influenza viruses such as rhinovirus, RSV, or coronavirus, are increasingly recognized as causative agents of severe respiratory tract infections in older people or persons with comorbid conditions and disease burden in an elderly population has become more accentuated [18-20].

Diagnosis:

According to the guidelines of the Infectious Diseases Society of America (IDSA) for seasonal influenza, a specimen from the respiratory tract should ideally be obtained within 5 days after symptom onset since the rate of false-negative results will otherwise increase due to declining viral shedding. Nasopharyngeal swabs in older children and adults and nasal aspirates in youngest children, respectively, are preferred over sputum or oropharyngeal swabs. Real-time polymerase chain reaction (RT-PCR) is the most sensitive and specific test and enables to differentiate between influenza subtypes. Commercially available rapid influenza diagnostic tests based on antigen detection are quick but less sensitive than RT-PCR. Other methods include immunofluorescence, virus culture and serology; however, the latter two do not provide a timely diagnosis [16]. The Centers for Disease Control and Prevention (CDC) recommendation clearly emphasizes that testing should be performed only if the result changes clinical practice or patient management [21].

The diagnostic modalities for other respiratory viruses also include PCR, antigen detection, and viral culture. Because of the numerous serotypes, routine antigen detection is not available for rhinovirus and PCR is therefore the most useful diagnostic test [4]. Respiratory virus panels test which are based on multiplex PCR is able to detect up to 20 different respiratory viruses including subtypes within a few hours [22]. Nasopharyngeal aspirates or nasal wash specimens are considered methods of choice but throat or nasal swabs are often employed because of their greater practicability [4].

Prevention and treatment:

Annual influenza vaccination is regarded as the most effective measure to prevent illness and spread of the disease. Elderly people and any person with underlying immunodeficiency might not be able to mount an effective immune response. In clinical trials, influenza vaccination prevented laboratory confirmed influenza in 70-90% of healthy individuals < 65 years of age and reduced work absenteeism, provided that the vaccine and circulating viruses are well matched [23].

Antiviral treatment should principally be considered in those patients at risk for severe or complicated illness. However, antiviral agents can also be considered in otherwise healthy subjects in order to shorten duration of symptoms or to prevent transmission to susceptible individuals at high risk, provided that treatment can be administered within the first 48 hours of illness onset. Most important, treatment must not be delayed until receipt of confirmatory diagnostic [24].

Due to the variety of pathogens treatment of the common cold has rather been symptom orientated [4]. Furthermore, although seldom indicated, antibiotics are often inappropriately prescribed [25]. The development of more specific antiviral treatments has been slow but seems to continuously evolve and promising results regarding treatment of e.g. human rhinovirus infections have been found in asthma patients [26].

1.2 Data base

A prospective cluster randomized controlled trial that evaluated whether vaccinating healthy Canadian Hutterite children and adolescents with either inactivated trivalent influenza vaccine

(ITIV) compared to hepatitis A vaccine as a "control" prevents influenza virus infection in the other Hutterite members, served as a basis of my study. All work in this completed trial was performed according to the guidelines of good clinical practice and the study was approved by the following Research Ethics Boards: Hamilton Health Sciences/Faculty of Health Sciences, the University of Calgary, the University of Saskatchewan and the University of Manitoba. Results from the trial have previously been published [27].

The population included children and adults of Canadian Hutterite communities in the provinces Alberta, Saskatchewan, and Manitoba. The study began in September 2008, extended over 3 influenza seasons and ended in July 2011. Eligible colonies had to show interest in the study, had at least 10 members at high risk for complications of influenza virus infection, and were located in reasonable distance (within 150 km) from designated towns. Colonies were excluded if children and adolescents did not receive routine childhood immunizations or if vaccination programs were implemented by local public health offices that offered influenza immunization for everyone.

After enrollment, the colonies were randomized to either the intervention or comparison group and remained in this group throughout the study. Every year, before beginning of the influenza season, healthy children aged 36 months and older and adolescents until the age of 15 years received either ITIV or hepatitis A according to the assignment of their respective colony. Exclusion criteria that applied to the original trial intervention and comparator group only comprised known allergies to the vaccine compounds as well as adverse effects potentially related to immunization. All Hutterite members were prospectively followed for signs and symptoms of influenza virus infection. There were no exclusion criteria for other Hutterite members. Influenza vaccination status was annually recorded in every study member and data about sociodemographic factors and comorbidities were collected at baseline and updated annually if required.

Signs and symptoms of respiratory illness including body temperature were recorded by the individuals or a family representative on a daily basis using a standardized questionnaire. Identical thermometers were distributed among the study participants. Trained study nurses visited the study colonies two times per week, checked the diaries for missing data and completed them with a family representative if required. The exact start and end date for each symptom was entered into a case report form. An ARI was defined as having at least 2 of the following symptoms: chills, cough, ear ache, fatigue, fever (≥ 38.0 C), headache, muscles aches, runny nose, or sore throat. A respiratory specimen (usually a nasopharyngeal swab) was obtained by the study nurse in all individuals that fulfilled the criteria for ARI and had 2 or more signs that were new since the last visit and new past 7 days of the last obtained specimen.

Overall there were 4640 individuals enrolled over 3 influenza seasons (2008-2011). Most study participants were followed over all three seasons and many individuals had several ARI episodes in the same season.

The data set available to investigate potential predictor of entero-/rhinovirus infections consisted of the season 1 data mentioned above where a respiratory virus panel assessing 16 different viruses was performed in most of the samples sent for influenza virus diagnosis.

Repeated ARI episodes occurring in the same individual were regarded as non-independent. Therefore, in order to minimize correlated data, only the first ARI episode in every season was considered.

1.3 Review of comparable literature

Clinical predictors of influenza are important in order to take rapid action in treatment and prevention. Accurate clinical diagnostic is especially useful to reduce unnecessary and potentially expensive rapid laboratory diagnostics; furthermore, empirical overtreatment with neuraminidase-inhibitors should be avoided since there is obviously no benefit for patients with an influenza-negative ARI [28]. However, a systematic review comprising 6 studies (7105 patients) showed that the accuracy of case definitions was imperfect in distinguishing influenza from ARI due to other viruses when participants were included irrespective of their age. Interpretation of the findings was challenged due to variable disease prevalence and inclusion criteria as well as different definitions of the fever cut off and variable duration of symptoms in the individual studies. Only studies before the outbreak of SARS were included. Overall, among a broad spectrum of studies, fever and cough significantly increased the likelihood of influenza, at least among elderly individuals [29]. A more recent review including 12 studies concluded that the simple heuristics such as the fever and cough rule, or the fever and cough, and acute onset rule have moderate accuracy [30].

It has previously been shown that symptoms vary according to age: gastrointestinal (diarrhoea/vomiting) symptoms has been most frequently seen in children whereas cough seems to be a predominant symptom in the elderly population. Other manifestations, however, such as fever, chills, headache, and rhinitis, showed no age-dependency [7]. Symptom manifestations may also differ according to influenza A subtype as shown in one study that stratified the predictive models according to influenza A subtypes: fever, rhinorrhea, cough, and myalgia were the most important predictors of influenza A/H3N2, whereas fatigue was the only significant positive predictor for A/H1N1 influenza; and in contrast to influenza A/H3N2, myalgia was negatively associated with A/H1N1[31]. Reviewing clinical signs and symptoms in patients during the 2009 pandemic influenza A/H1N1 revealed that mild illness without fever occurred in 8-23% of infected patients whereas fever was the predominant symptom in hospitalized patients [32]. A large population-based study found that children with sore throat were more likely to have either seasonal influenza A/H1N1 or influenza A/H3N2 than 2009 pandemic influenza and that adults with sore throat were more likely to have seasonal influenza A/H1N1 than pandemic influenza 2009, but that the proportions of patients with fever among the different subtypes were equal [33].

Overall, clinical predictors for influenza have broadly been investigated but often during one season only precluding consideration of temporal/seasonal variations [34-36]. Two studies used simple univariate or bivariate comparisons only [37, 38]. None of these studies [31, 34-40] applied any form of internal validation such as using split-halves or bootstrapping and the entire data set was used to create the models; performance of these predictive models, therefore, is likely overestimated. A major concern in most studies is the risk of selection bias and, in particular, spectrum bias due to having fever as a prerequisite for inclusion [34-39] or due to the fact that sampling of symptomatic patients was terminated after confirmation of the circulating influenza strains [40] which likely overestimated the importance of fever as predictor and obviates generalizability of the findings. Spectrum bias is a special concern in diagnostic test studies when the diagnostic test has been evaluated in a population with a different spectrum of disease or where the non-diseased population has a different spectrum of competing diagnoses than the population has for which the test is intended for [41, 42]. The cut-off value of a clinical prediction rule for the diagnosis of influenza derived from a population with a narrow spectrum of symptoms (e.g. all had fever) has likely a lower sensitivity but higher specificity when applied to patients with more general symptoms of ARI.

In order to guide influenza diagnostic testing and empiric treatment Ebell, Afonso, and colleagues developed a simple Flu score using multivariate analysis and derived a clinical decision rule based on a classification and regression tree (CART) analysis combining two cohorts of primary care outpatients and emergency care patients applying split-halves for validation [43, 44]. Although fever was the symptom in the primary decision node, there still remained a moderate risk of having influenza virus infection in the absence of fever, given that other systemic symptoms such as chills, myalgia or sweats were present; this population at moderate risk for influenza virus infection does not fulfill the commonly used case definitions of fever and cough or sore throat and would probably best benefit from further diagnostic testing.

Most studies did not differentiate between influenza A subtypes [37-40, 43] and variations in clinical characteristics among different influenza A subtypes are incompletely investigated.

To the best of our knowledge, the clinical prediction rules established by Ebell and Afonso et al. [43, 44] have never been validated in an external cohort.

Contrariwise, a prediction rule for entero-/rhinovirus-related ARI has, so far, only been established and validated in adolescents and adults but not in children [45]. Based on the fact that no rapid testing for entero-/rhinovirus in routine primary care practice is in use a prediction rule based on symptoms in children under 15 years of age could be useful.

1.4 Thesis objectives

A clinical prediction rule estimates the probability of an outcome or relates certain clinical features to the indication of a diagnostic test or the choice of treatment. Physicians therefore can use clinical prediction rules to classify patients according to their probability of a disease (assistive decision rule) or to decide whether or not there is benefit from treatment (directive decision rule) [46]. They further help clinicians to focus on clinical data that are important to obtain. This implies that a clinical prediction rule should include factors that not only have been shown in theory to be predictive of a disease but that are also feasible, realistic, and prompt to obtain; and the chosen outcome should be relevant and pertinent to clinical practice. Clinical predictors can include findings from history, physical exam and laboratory tests [47, 48].

The study of predictors to diagnose influenza- and non-influenza-related ARI in children and adults is of importance due to the high prevalence and burden of ARI including the risk of developing complications, the possibility to alter patient care through different measures (e.g. antiviral treatment, isolation precautions to prevent horizontal transmission), the low accuracy or unavailability of rapid diagnostic tests, and because of its potential impact on reducing unnecessary antiviral treatment. The aim is to evaluate symptomatic predictors (e.g. fever, cough, chills) and demographic factors e.g. vaccination against influenza, which are usually straight forward to obtain from the history of a patient or during the physical examination.

However, every newly developed prediction rule should be critically appraised with respect to how accurately it predicts the outcome and whether it has been externally validated and demonstrated benefit in an impact analysis before it can be considered to be applied in clinical practice. Yet, an often raised criticism is whether a particular prediction rule does indeed alter patient management or efficiency [46].

Various mathematical and non-mathematical models can be applied for developing a predictive model. Multiple regression models and recursive partitioning using regression or classification trees, are the ones most frequently cited among the mathematical models [49]. It has been shown in a similar work comparing various models to predict lower respiratory tract infections that non-mathematical models such as linear judgment or consensus models were outperformed by the mathematical models due to inferior discriminative ability [50]. This thesis therefore will focus on the comparison of different mathematical models. Such models can be parametric, if the relationship between the response variable and the explanatory variables is linear, or non-parametric, where the multivariable analysis does not require specification of a parametric model to develop a prediction rule in a systematic way.

The main objectives of this thesis were to identify relevant and timely measureable predictors from the literature and from what was available in the existing data set in order to (1) better characterise symptomatic predictors of influenza A and B virus infections in a population of children and adults with a broader spectrum of symptoms, (2) to explore potential differences in symptomatic presentation of influenza A(H1) and (H3) subtypes, (3) to develop and validate predictive scores for influenza virus infection, and (4) to evaluate the performance of the Flu score 3 developed by Ebell et al. in the Hutterite data set, and finally (5) whether symptoms of entero-/rhinovirus infections in children differ from other ARI.

From a methodological point of view we aimed to explore different statistical techniques to generate predictive models to diagnose influenza and non-influenza virus ARI for children and adults, to compare the performance of the different strategies for the derivation process by means of standard logistic regression, recursive partitioning, and Cox regression analysis. The performances of the models were compared by creating ROC curves from the scores or the predicted probabilities, respectively, from the derivation and validation set and examine the differences in the AUCs.

Eventually, our goal was to create a score that could be applied in clinical practice to assist physicians in their decision-making. Secondary objectives that evolved during the preparation of the data for the analysis were to learn methods how to deal with correlated or nested data.

All statistical analyses were performed using IBM SPSS for Windows version 23.0 [51] and R software version 3.3.0 [52].

Main research question:

Among children and adults from Hutterite communities with at least two of the following symptoms: Fever (≥38°C), cough, nasal congestion, sore throat, headache, sinus problems, muscle aches, fatigue, ear ache, or chills, as evaluated with a standardized, self-administered questionnaire, and in whom the diagnosis of influenza virus infection was uncertain at the time of first swab collection, to what extent can any of these symptoms discriminate between the presence or absence of influenza virus infection as established by PCR from a respiratory specimen as reference standard diagnostic?

Hypotheses:

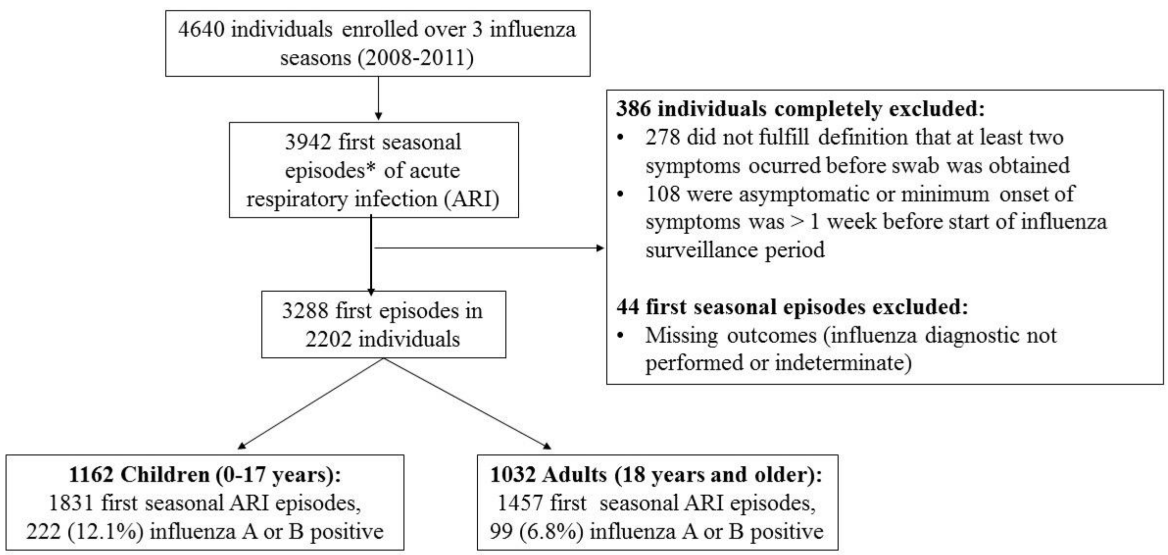
1) Based on previous experiences we hypothesized that the derived clinical prediction rule will perform worse when applied to the validation group.
2) There is no difference in the discriminative ability of a clinical prediction rule/influenza score developed by different methods including multiple logistic regression and recursive partitioning.

Setting:

Between December 2008 and June 2011, 3942 first seasonal ARI episodes in 2480 individuals were recorded. Overall, 108 individuals were excluded for all 3 seasons because they were actually asymptomatic or their minimum symptom onset was more than a week before start of the influenza surveillance period and another 278 subjects were excluded for all 3 seasons because they did not fulfill the eligibility criteria of at least two symptoms within 10 days before the respiratory specimen was obtained and 44 first seasonal episodes were excluded because they had a missing outcome (either influenza diagnostic was not performed or the result was indeterminate). There were eventually 3288 first seasonal ARI episodes in 2202 remaining subjects available which accounted for 321 (9.8%) influenza A or B positive events.

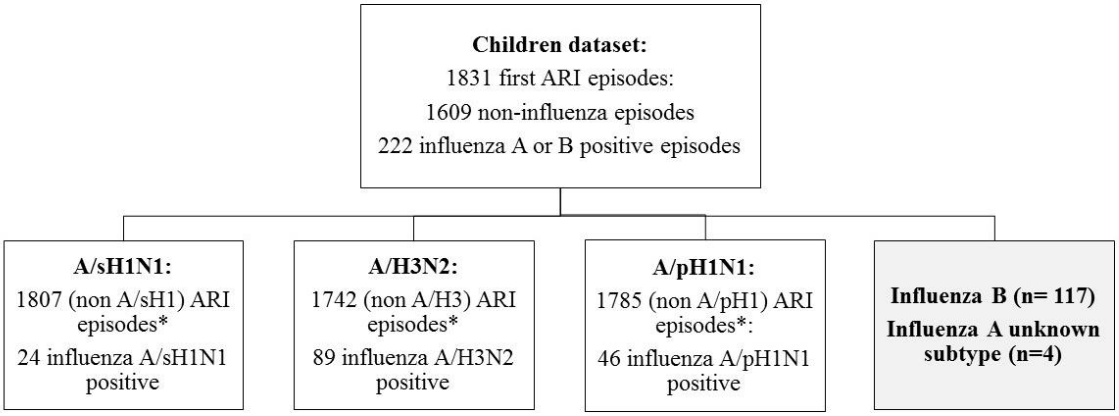
Based on data from the literature and own experience we decided that the analyses for the primary outcome, presence or absence of laboratory confirmed influenza virus (A or B) infection, and the secondary outcomes, presence or absence of specific influenza A subtypes, will be by age category. In order to preserve a reasonable number of observations, the data set was divided into children (age 0-17 years) and adults (18 years and older) (Figs 1.1-1.3) and the analyses that follow are all performed separately for children and adults.

Figure 1.1 Flow-chart of study population



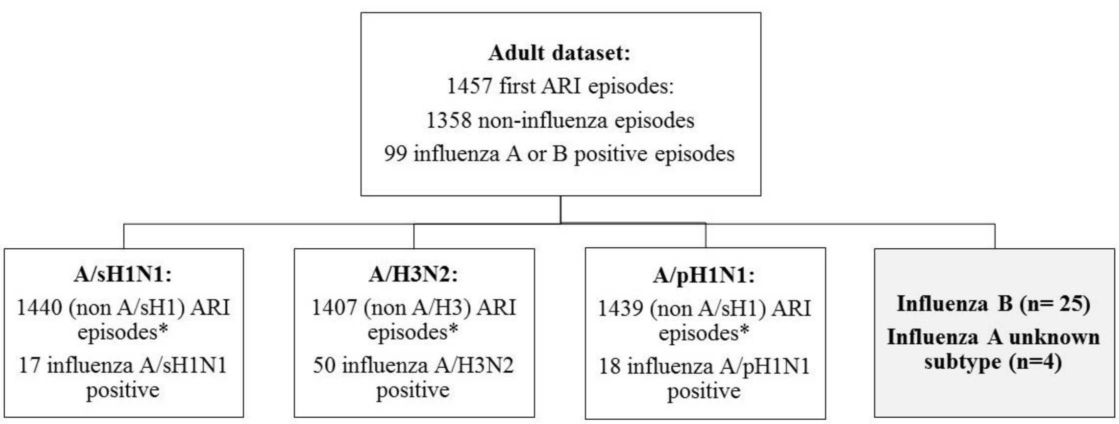
*total of first episodes of acute respiratory infection (ARI) over all 3 seasons, where ARI is defined as having ≥ 2 of the following symptoms: chills, cough, fever (≥ 38°C), ear ache, headache, muscles aches, runny nose, sinusproblems, sore throat.

Figure 1.2 Frequencies of first ARI episodes and number of episodes of influenza A or B infection subdivided into influenza A subtypes (seasonal A/H1N1, A/H3N2, pandemic influenza A/H1N1) and influenza B or unknown influenza A subtype in children.



* Including all influenza negative, influenza B, and other influenza A subtypes, respectively; s = seasonal, p = pandemic

Figure 1.3 Frequencies of first ARI episodes and number of episodes of influenza A or B infections divided into influenza A subtypes (seasonal A/H1N1, A/H3N2, pandemic influenza A/H1N1) and influenza B or unknown influenza A subtype in adults



* Including all influenza negative, influenza B, and other influenza A subtypes, respectively; s = seasonal, p = pandemic

Outcomes:

The primary outcome of interest is laboratory confirmed influenza A or B virus infection. Aiming for a balance between a sample size with enough observations of the outcome and minimal correlated data we considered only the first ARI episode in every season. This means that one individual had a maximum of 3 observations over the whole study period. Respiratory specimens were examined by real-time polymerase chain reaction (RT-PCR) (molecular method) for the presence of influenza viral RNA using the CDC Human Influenza Virus Real-Time RT-PCR Detection and Characterization Panel, which is known to have high sensitivity and specificity [27].

As secondary outcomes we investigated whether symptomatic predictors vary according to the subtypes of influenza A. The outcomes therefore are either influenza A/H1N1 seasonal, influenza A/H3N2, or influenza A/H1N1 pandemic, each compared versus all other influenza subtypes and non-influenza-related ARIs. Eventually, we also explored the diagnosis of laboratory confirmed entero-/rhinovirus infection against all other ARI as a binary response variable.

Predictors, potential confounders & effect modifiers

The predictors of interest were the individual symptoms: fever (≥38°C), cough, runny nose, sore throat, headache, sinus problems, muscle aches, fatigue, ear ache, or chills, as evaluated with a standardized, self-administered questionnaire. All symptoms were recorded on a daily basis as being either present or absent. However, for this analysis, only symptoms entered within the 10 days prior to the sampling of a respiratory specimen were considered for the analysis. Symptoms that were ongoing for a longer time were deemed unlikely to be related to influenza. Influenza viral shedding usually lasts on average 5 days but may vary according to subtypes and may be longer in children and immunocompromised patients [32, 53]. Acuity of symptom onset is frequently a criterion in clinical case definitions of influenza diagnosis and applies to symptoms that occurred within the last 48 hours of presentation. A new dichotomous variable was therefore created in the dataset that defined whether or not the symptoms started within 2 days before the swab specimen was obtained. Influenza immunization has been found to be protective against influenza virus infection and was, therefore, included as predictor of interest for the comparison between influenza A or B and non-influenza virus related ARI. Gender was deemed an important covariate to be adjusted for in the comparison between influenza and non-influenza ARI. The role of age either as a potential confounder or effect modifier was less clear. Categorized as children and adults it could potentially affect the strength of the relation between the various predictors and the outcome influenza virus infection and as such would be considered an effect modifier. However, age is also often considered a confounder. A mediator would be in the causal pathway between the exposure and the outcome; however, no such factor was identified (supplemental figure 1, appendix).

1.5 Validation of predictive models

When it comes to validation of a prognostic or predictive model two statistical approaches are of importance that help to determine the prediction accuracy or model performance: calibration and discrimination [54].

Calibration

When calibrating a logistic regression model we compare predictions and observations [54]. We can plot the observed proportions of patients with the disease ("cases") against expected (predicted) proportions of cases defined by ranges of predicted risk. If the observed proportions and predicted individual risks agree over the entire range of probabilities, the slope of the fitting line would be 1. The Hosmer and Lemeshow test is frequently proposed in addition to test the models' goodness-of-fit [55]. The original dataset itself is commonly well-calibrated; calibration is therefore more relevant when predicting cases in a new/external data set with the variables and estimates from the original data set, and comparing these proportions of predicted cases to the proportions of observed cases in this new data set.

Measuring Discriminability

Another important step is to determine the discriminative ability of the logistic regression model, e.g. how well the model correctly distinguishes between those with the outcome and those without. Complimentary measurements to summarize discriminability include sensitivity and specificity, Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC), or R-square [55]. We compared the ROC curves between the derivation and validation set as well applying the z-test as described by Hanley and McNeil [56].

Internal versus external validation

Every prediction model determined in a single data set generally overestimates its performance either due to overfitting, or because an important predictor was not measured [55]. It is therefore not enough to only show its predictive ability in the derivation data set. Instead, the model should ideally be validated by quantifying its predictive performance in another population. This is an important step before implementing a new prediction rule. If the performance in the validation set is poor, the model should be adjusted [57].

If this population is from a different centre than the original sample that was used for the development, the validation is usually referred to as "external validation" in contrast to "internal validation", where the researchers perform the validation within the same data set usually by applying a form of re-sampling of the observed data. Another approach is "temporal validation" which, in terms of stringency, can be regarded as between internal (= lowest stringency) and external (= highest stringency) validation and utilizes different observation periods for deriving and validating a model in the same population. In this work we planned to perform internal validation. Different methods are described to determine the validity of the derived estimates within the same data set: data-splitting, cross-validation, and bootstrapping [58].

Data splitting:

In data-splitting, the available data set is randomly divided into a derivation set used for the model development and the remainder of the data set is reserved solely for the purpose of validation. The ratio between the derivation and validation set is a trade-off between having enough observations for the model fitting, which at least initially involves a larger number of predictors including interaction terms, and the validation process, in which only the final, and hopefully reduced, model is being tested. Formalized procedures are available [59] and serve as the basis for a rule of thumb that the validation set usually consists of 1/4 to 1/3 of the full dataset [54].

We therefore randomly divided each, the children and adult data set, into 2 groups, a derivation and validation group. A 1:1 ratio is often used for data-splitting. However, in view of the considerable amount of independent variables, a larger derivation group of approximately 66% was chosen in order to have an adequate sample size of observed events for the primary analysis. Assuming, that there will be fewer variables in the final model a validation group of 34% of the original dataset was deemed appropriate.

Table 1.1 Frequencies of observations after data-splitting (random split, 2:1) in the children and adults data-set including all 3 seasons

	Children		**Adults**	
	Derivation (n=1241)	**Validation (n=590)**	**Derivation (n=950)**	**Validation (n=507)**
	n (%)	n (%)	n (%)	n (%)
Influenza A or B	152 (12.2)	70 (11.9)	67 (7.1)	32 (6.3)
Influenza A/sH1N1	15 (1.2)	9 (1.5)	11 (1.2)	6 (1.2)
Influenza A/H3N2	63 (5.1)	26 (4.4)	31 (3.3)	19 (3.7)
Influenza A/pH1N1	35 (2.8)	11 (1.9)	12 (1.3)	6 (1.2)

1.6 Other methodological aspects

Developing a point score to predict influenza and non-influenza virus related ARI

A score was generated based on the magnitude of the regression coefficients. After finding as set of best fitting models determined by the maximum likelihood estimation, different scores were developed following the steps proposed by Sullivan et al. [60]. The method simplifies when there are only dichotomous predictors in that each value of the beta coefficients is divided by a constant B. Since this constant reflects the number of regression units that correspond to one point, I divided all coefficients by the absolute value of the smallest beta-coefficient and rounded it to the nearest integer. I did not further multiply by e.g. 2 or 10 as frequently seen as I deemed it impractical for physicians to sum up larger numbers.

Sample size calculation and estimation of precision

Due to the retrospective nature of the study the sample size and the number of events were predetermined. Since the goal of this study is to identify symptomatic predictors which either correctly rule in or rule out the diagnosis of influenza virus infection, precision rather than power is the appropriate estimate; the former relates to a confidence interval where the lower and upper confidence limits of an accuracy index (e.g. sensitivity of a test) are reasonably narrow and cover a range of clinically meaningful values [61]. According to the formula provided by Obuchowsky [62], the estimated precision (width of the confidence interval) would be about 0.08 for a hypothesized sensitivity of our test of 90%, given an expected prevalence of influenza virus infection of 10% and a sample size of 1889 (Tab. 1.2); in other words, the 95% confidence limit for a sensitivity of 0.90 would be 0.86-0.94. However, this formula assumes independence of observations and the precision in clustered data is likely to be lower, depending on the size of the clusters and the degree of correlation within clusters [63].

Power- and sample size predictions in multivariable logistic regression are seldom exact and are driven by the number of independent variables including all relevant interaction terms. For sample size calculations rules-of-thumb that each independent variable requires at least 10 events are frequently operated [64]; and, vice versa, the number of predictors that can be included in a logistic regression model will be guided by the number of events in the data set. This assures to have enough degrees of freedom and avoids overfitting. Although simple it is probably the most transparent and comprehensible method to estimate the maximum number of predictors for a specified number of events [65]. As evident from Table 1.1, for the different influenza A subtypes data-splitting for validation is suboptimal due to the low number of observations that would allow inclusion of only a small number of predictors; inclusion of more predictors will likely result in an unstable model because of overfitting.

Table 1.2 Estimation of precision (width of the 95% confidence interval) with a given sample size, an expected prevalence of 10% and 7%, respectively, and a pre-specified (two-tailed) alpha of 0.05 for various hypothesized sensitivities of the clinical prediction rule.

	Children				**Adults**			
Sensitivity	0.95	0.90	0.85	0.8	0.95	0.90	0.85	0.8
Prevalence	0.10	0.10	0.10	0.10	0.07	0.07	0.07	0.07
alpha (α)	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
total sample	1889	1889	1889	1889	1482	1482	1482	1482
L	0.03	0.04	0.05	0.06	0.04	0.06	0.07	0.08

Formula for sample size [62], can be transformed to determine L: $n = z^2_{\alpha/2} V(\hat{\theta}_1)/L^2$

θ_1 = unknown accuracy (here the sensitivity) of the test under evaluation

α = type I error rate

L = desired width of one-half of the CI

Missing data:

We pre-specified that if more than 10% of the individuals would be excluded due to missing or invalid outcome measurement that we would scrutinize the excluded cases and compare their characteristics to those in the remaining dataset to evaluate for relevant discrepancies. For independent variables not missing completely at random we considered multiple imputation methods but this was not necessary since the dataset was almost complete with respect to the independent variables of interest.

Accounting for correlated data for the prediction of influenza virus infection

When considering that some of the subjects provided more than one respiratory sample due to a first ARI episode in a different season we have to find a way to account for these correlated data within subjects. Ignoring such within subject correlations may lead to incorrect estimation of the model parameters [66]. For non-normal response variables Generalized Estimating Equations (GEE) provide more efficient and unbiased regression estimates. One main characteristic of GEE, that also distinguishes it from other repeated measures approaches such as random-effects model, is that a unit-change in a parameter X describes an average response across the population rather than a cluster-specific individuals' response [66, 67].

Since GEE estimates are known to be robust even if the correlation structure was miss-specified but correlation of the data has to be accounted for in some way we chose an exchangeable correlation as the working correlation and used a "robust" estimator for standard errors for building a population average logistic regression model [68].

Another approach to analyse correlated data in logistic regression is the random effects model (synonyms: "cluster-specific" or "conditional" model). The clusters can be specific individuals but with multiple observations or related subjects with single observations. Applying a random-effects model is more appropriate when inferences are made on covariates which change within a cluster. The method of parameter estimation is usually based on the maximum likelihood [69].

In survival analysis clustering of data can be accounted for using robust variance estimation in which the assumptions of the distribution of error terms is more relaxed [70].

In summary, the increase in standard errors, may result in less significant parameter estimates than if independency of data was assumed.

References:

1. Russell, K., et al., *Update: Influenza Activity - United States, October 4, 2015-February 6, 2016.* MMWR Morb Mortal Wkly Rep, 2016. **65**(6): p. 146-53.
2. Bramley, T.J., D. Lerner, and M. Sames, *Productivity losses related to the common cold.* J Occup Environ Med, 2002. **44**(9): p. 822-9.
3. Nichol, K.L., et al., *Burden of influenza-like illness and effectiveness of influenza vaccination among working adults aged 50-64 years.* Clin Infect Dis, 2009. **48**(3): p. 292-8.
4. Heikkinen, T. and A. Jarvinen, *The common cold.* Lancet, 2003. **361**(9351): p. 51-9.
5. Fowlkes, A., et al., *Incidence of medically attended influenza during pandemic and post-pandemic seasons through the Influenza Incidence Surveillance Project, 2009-13.* Lancet Respir Med, 2015. **3**(9): p. 709-18.
6. Kirkpatrick, G.L., *The common cold.* Prim Care, 1996. **23**(4): p. 657-75.
7. Cox, N.J. and K. Subbarao, *Influenza.* Lancet, 1999. **354**(9186): p. 1277-82.
8. Bresee, J. and F.G. Hayden, *Epidemic influenza--responding to the expected but unpredictable.* N Engl J Med, 2013. **368**(7): p. 589-92.
9. Harris, J.M., 2nd and J.M. Gwaltney, Jr., *Incubation periods of experimental rhinovirus infection and illness.* Clin Infect Dis, 1996. **23**(6): p. 1287-90.
10. Gwaltney, J.M., Jr., et al., *Rhinovirus infections in an industrial population. II. Characteristics of illness and antibody response.* JAMA, 1967. **202**(6): p. 494-500.
11. Arruda, E., et al., *Frequency and natural history of rhinovirus infections in adults during autumn.* J Clin Microbiol, 1997. **35**(11): p. 2864-8.
12. Eccles, R., *Understanding the symptoms of the common cold and influenza.* Lancet Infect Dis, 2005. **5**(11): p. 718-25.
13. Turner, R.B., et al., *Association between interleukin-8 concentration in nasal secretions and severity of symptoms of experimental rhinovirus colds.* Clin Infect Dis, 1998. **26**(4): p. 840-6.
14. Oh, J.W., et al., *Interleukin-6, interleukin-8, interleukin-11, and interferon-gamma levels in nasopharyngeal aspirates from wheezing children with respiratory syncytial virus or influenza A virus infection.* Pediatr Allergy Immunol, 2002. **13**(5): p. 350-6.
15. Hendley, J.O., *The host response, not the virus, causes the symptoms of the common cold.* Clin Infect Dis, 1998. **26**(4): p. 847-8.
16. Harper, S.A., et al., *Seasonal influenza in adults and children--diagnosis, treatment, chemoprophylaxis, and institutional outbreak management: clinical practice guidelines of the Infectious Diseases Society of America.* Clin Infect Dis, 2009. **48**(8): p. 1003-32.
17. Mertz, D., et al., *Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis.* BMJ, 2013. **347**: p. f5061.
18. Falsey, A.R., et al., *Respiratory syncytial virus infection in elderly and high-risk adults.* N Engl J Med, 2005. **352**(17): p. 1749-59.
19. Walsh, E.E., J.H. Shin, and A.R. Falsey, *Clinical impact of human coronaviruses 229E and OC43 infection in diverse adult populations.* J Infect Dis, 2013. **208**(10): p. 1634-42.
20. Atmar, R.L., *Uncommon(ly considered) manifestations of infection with rhinovirus, agent of the common cold.* Clin Infect Dis, 2005. **41**(2): p. 266-7.
21. *Guidance for Clinicians on the Use of Rapid Influenza Virus Diagnostic Tests.* www.cdc.gov/flu/pdf/professionals/diagnosis/testing_algorithm.pdf. Last accessed, June 3, 2016.
22. Mahony, J., et al., *Development of a respiratory virus panel test for detection of twenty human respiratory viruses by use of multiplex PCR and a fluid microbead-based assay.* J Clin Microbiol, 2007. **45**(9): p. 2965-70.

23. Fiore, A.E., et al., *Prevention and control of influenza with vaccines: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2010.* MMWR Recomm Rep, 2010. **59**(RR-8): p. 1-62.
24. Fiore, A.E., et al., *Antiviral agents for the treatment and chemoprophylaxis of influenza --- recommendations of the Advisory Committee on Immunization Practices (ACIP).* MMWR Recomm Rep, 2011. **60**(1): p. 1-24.
25. in *Respiratory Tract Infections - Antibiotic Prescribing: Prescribing of Antibiotics for Self-Limiting Respiratory Tract Infections in Adults and Children in Primary Care*. 2008: London.
26. Hayden, F.G., *Advances in antivirals for non-influenza respiratory virus infections.* Influenza Other Respir Viruses, 2013. **7 Suppl 3**: p. 36-43.
27. Loeb, M., et al., *Effect of influenza vaccination of children on infection rates in Hutterite communities: a randomized trial.* JAMA, 2010. **303**(10): p. 943-50.
28. Dobson, J., et al., *Oseltamivir treatment for influenza in adults: a meta-analysis of randomised controlled trials.* Lancet, 2015. **385**(9979): p. 1729-37.
29. Call, S.A., et al., *Does this patient have influenza?* JAMA, 2005. **293**(8): p. 987-97.
30. Ebell, M.H. and A. Afonso, *A systematic review of clinical decision rules for the diagnosis of influenza.* Ann Fam Med, 2011. **9**(1): p. 69-77.
31. Carrat, F., et al., *Evaluation of clinical case definitions of influenza: detailed investigation of patients during the 1995-1996 epidemic in France.* Clin Infect Dis, 1999. **28**(2): p. 283-90.
32. Writing Committee of the, W.H.O.C.o.C.A.o.P.I., et al., *Clinical aspects of pandemic 2009 influenza A (H1N1) virus infection.* N Engl J Med, 2010. **362**(18): p. 1708-19.
33. Belongia, E.A., et al., *Clinical characteristics and 30-day outcomes for influenza A 2009 (H1N1), 2008-2009 (H1N1), and 2007-2008 (H3N2) infections.* JAMA, 2010. **304**(10): p. 1091-8.
34. Boivin, G., et al., *Predicting influenza infections during epidemics with use of a clinical case definition.* Clin Infect Dis, 2000. **31**(5): p. 1166-9.
35. Dai, X.Q., et al., *Clinical predictors for diagnosing pandemic (H1N1) 2009 and seasonal influenza (H3N2) in fever clinics in Beijing, China.* Biomed Environ Sci, 2012. **25**(1): p. 61-8.
36. Woolpert, T., et al., *Determination of clinical and demographic predictors of laboratory-confirmed influenza with subtype analysis.* BMC Infect Dis, 2012. **12**: p. 129.
37. Hulson, T.D., et al., *Diagnosing influenza: the value of clinical clues and laboratory tests.* J Fam Pract, 2001. **50**(12): p. 1051-6.
38. Ohmit, S.E. and A.S. Monto, *Symptomatic predictors of influenza virus positivity in children during the influenza season.* Clin Infect Dis, 2006. **43**(5): p. 564-8.
39. Monto, A.S., et al., *Clinical signs and symptoms predicting influenza infection.* Arch Intern Med, 2000. **160**(21): p. 3243-7.
40. Shah, S.C., et al., *Clinical predictors for laboratory-confirmed influenza infections: exploring case definitions for influenza-like illness.* Infect Control Hosp Epidemiol, 2015. **36**(3): p. 241-8.
41. Montori, V.M., et al., *Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests.* CMAJ, 2005. **173**(4): p. 385-90.
42. Lachs, M.S., et al., *Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection.* Ann Intern Med, 1992. **117**(2): p. 135-40.
43. Ebell, M.H., et al., *Development and validation of a clinical decision rule for the diagnosis of influenza.* J Am Board Fam Med, 2012. **25**(1): p. 55-62.
44. Afonso, A.M., et al., *The use of classification and regression trees to predict the likelihood of seasonal influenza.* Fam Pract, 2012. **29**(6): p. 671-7.
45. Monto, A.S., T.J. Bramley, and M. Sarnes, *Development of a predictive index for picornavirus infections.* Clin Infect Dis, 2003. **36**(3): p. 253-8.
46. Barrett, T.W. and D.L. Schriger, *Annals of emergency medicine journal club. Clinical prediction rules answers to the November 2009 journal club.* Ann Emerg Med, 2010. **55**(4): p. 380-9.

47. Reilly, B.M. and A.T. Evans, *Translating clinical research into clinical practice: impact of using prediction rules to make decisions.* Ann Intern Med, 2006. **144**(3): p. 201-9.
48. Wasson, J.H., et al., *Clinical prediction rules. Applications and methodological standards.* N Engl J Med, 1985. **313**(13): p. 793-9.
49. Cook, E.F. and L. Goldman, *Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis.* J Chronic Dis, 1984. **37**(9-10): p. 721-31.
50. Loeb, M., *Predictive Models of Lower Respiratory Tract Infections in Residents of Long-Term Care Facilities For The Elderly.* Unpublished thesis submitted to the School of Graduate Studies. McMaster University., 1997.
51. *IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.*
52. *R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.*
53. Loeb, M., et al., *Longitudinal study of influenza molecular viral shedding in Hutterite communities.* J Infect Dis, 2012. **206**(7): p. 1078-84.
54. Altman, D.G., et al., *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**: p. b605.
55. Royston, P., et al., *Prognosis and prognostic research: Developing a prognostic model.* BMJ, 2009. **338**: p. b604.
56. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.
57. Moons, K.G., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.* Ann Intern Med, 2015. **162**(1): p. W1-73.
58. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.
59. Picard, R.R. and K.N. Berk, *Data Splitting.* American Statistician, 1990. **44**(2): p. 140-147.
60. Sullivan, L.M., J.M. Massaro, and R.B. D'Agostino, Sr., *Presentation of multivariate data for clinical use: The Framingham Study risk score functions.* Stat Med, 2004. **23**(10): p. 1631-60.
61. Adams-Huet, B. and C. Ahn, *Bridging clinical investigators and statisticians: writing the statistical methodology for a research proposal.* J Investig Med, 2009. **57**(8): p. 818-24.
62. Obuchowski, N.A., *Sample size calculations in studies of test accuracy.* Stat Methods Med Res, 1998. **7**(4): p. 371-92.
63. Lee, E.W. and N. Dubin, *Estimation and sample size considerations for clustered binary responses.* Stat Med, 1994. **13**(12): p. 1241-52.
64. Norman, G., S. Monteiro, and S. Salama, *Sample size calculations: should the emperor's clothes be off the peg or made to measure?* BMJ, 2012. **345**: p. e5278.
65. Haynes, R.B., *Clinical epidemiology : how to do clinical practice research*. 3rd ed. 2006, Philadelphia: Lippincott Williams & Wilkins. xv, 496 p.
66. Ballinger, G.A., *Using generalized estimating equations for longitudinal data analysis.* Organizational Research Methods, 2004. **7**(2): p. 127-150.
67. Heck, R.H., S. Thomas, and L.N. Tabata, *Multilevel modeling of categorical outcomes using IBM SPSS*. Quantitative methodology series. 2012, New York ; London: Routledge. xvi, 439 p.
68. Hosmer, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression*. Third edition / ed. Wiley series in probability and statistics. 2013, Hoboken, New Jersey: Wiley. xvi, 500 pages.
69. Hox, J.J., *Multilevel analysis : techniques and applications*. 2nd ed. Quantitative methodology series. 2010, New York: Routledge an imprint of the Taylor and Francis Group.
70. Kleinbaum, D.G., et al., *Applied regression analysis and other multivariable methods*. Fifth edition. ed. 2013, Boston, MA: Cengage Learning. xix, 1051 pages.

CHAPTER II

2.0 The Logistic regression model

If the outcome takes only two possible values such as presence or absence of influenza virus infection the most commonly used method to describe the relationship between the response or dependent variable (denoted as Y) and one or more independent (or explanatory) variables (denoted as X_1, X_2, X_3, …, X_i) is logistic regression analysis.

2.1 Characteristics of logistic regression:

Multivariable (multiple) regression models are parametric models. Compared to linear regression where the relationship is specified between a continuous outcome and various independent variables, the outcome in logistic regression is a binary response variable (Y), e.g. the presence or absence of influenza virus infection and the logit is used as link function to establish a linear relationship between the logit of Y and the numerous predictors [1].

Logistic regression differs from linear regression not only by the dichotomous response variable but also by the underlying assumptions. In linear regression one of the assumptions is that *the mean value of Y, $\mu_{(Y|X)}$, is a straight-line function of X* [1] and can be described with the following equation:

$$\mu_{(Y|X)} = \beta_0 + \beta_1 X. \qquad (2.1.1)$$

In logistic regression, however, the conditional mean of Y given X reflects the probability (p) of Y=1 given X where p is bounded between 0 and 1. In order to obtain a straight-line function between X and the response variable Y, p needs to undergo the so called *logit transformation* which is defined as:

$$\text{logit (p)} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \text{ and } p\,(Y{=}1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \qquad (2.1.2)$$

The logit (p) may be continuous and ranges from $-\infty$ to $+\infty$.

Another difference between the linear and logistic regression model is that in logistic regression the outcome follows a binomial (not a normal) distribution.

The coefficients estimated in the logistic regression model inform about the relationships of the predictors and the outcome; a measure derived from these estimates is the *odds ratio* (OR). The OR in binary logistic regression compares the odds of having the disease in one group with a certain characteristic or exposure (e.g. being vaccinated against influenza) to the odds of having the disease in another group without that exposure (e.g. not being vaccinated against influenza). Where the odds of influenza virus infection in one group is defined as:

$$\text{Odds (D)} = \frac{pr\ (influenza)}{pr(no\ influenza)} = \frac{pr(influenza)}{1-pr(influenza)} \qquad (2.1.3)$$

The OR can be derived from the estimated *beta*-coefficients of the logistic regression model comparing group A (with the exposure) to group B (without exposure) by:

$$\text{OR}_{\text{A vs. B}} = \frac{odds\ (influenza) in\ group\ A}{odds\ (influenza) in\ group\ B} = \frac{e^{(\beta_0+\ \beta_1)}}{e^{\beta_0}} = e^{\beta_1} \qquad (2.1.4)$$

We can interpret this as: The estimated odds of contracting influenza for a person in group A is about e^{β_1} times that of a person in group B.

2.2 Model development:

We conducted univariable logistic regression for the comparison of baseline characteristics in the influenza and non-influenza group and tested the crude associations of variables of interest.

The aim was to find the most parsimonious sets of independent variables that are best able to predict the outcome. Because there is no automated model selection for GEE in SPSS logistic regression using stepwise forward selection was first performed with probabilities of 0.05 and 0.1 for entry and removal, respectively, based on the likelihood ratio test. Variables significant (P <

.05) in the univariable analysis were entered into the multivariable model to relate these predictors to the primary and secondary outcomes. From the predictors available in the dataset, we pre-specified to also include all those in the multiple logistic regression analysis that were deemed as clinically important in the literature and were used in previous prediction models: gender, fever, cough, and sore throat.

The Hosmer-Lemeshow goodness of fit test was applied and a p-value of ≥ 0.05 was regarded as good model fit. Multicollinearity for categorical variables was examined by the variance inflation factor (VIF). A VIF > 5.0 was considered inacceptable and one of the highly correlated variables had to be eliminated. In general, two-sided p-values < 0.05 were considered statistically significant.

Thus, logistic regression analyses was first performed in the derivation set. The final model parameters were analyzed by Generalized estimating equations (GEE) using an exchangeable correlation and robust estimators (= population average model). Only the final population average models were then applied to the test set.

For each final model a predictive score based on the beta-coefficients was generated and different cut-offs determined to find optimal sensitivities and specificities.

The differences in model performance between the derivation and validation sets were examined by means of ROC curves and differences in the AUCs were compared using the z-score test [2].

2.3 Results

2.3.1 Prediction of influenza A/B virus infection

The frequencies of the demographic and clinical predictors of influenza-positive and negative first seasonal episodes in the children and adult data set are listed in table 2.1.

Table 2.1. Characteristics of influenza positive and negative first seasonal episodes in the children and adult derivation set, influenza seasons 2008-2011. Values are indicated as numbers and percentages, n (%):

	Children (0-17 years)		**Adults (18 years and older)**	
	Influenza A or B negative, n=1089	**influenza A or B positive, n=152**	**influenza A or B negative, n=883**	**influenza A or B positive, n=67**
Female subjects	642 (59.0)	82 (53.9)	631 (71.5)	39 (58.2)
Influenza vaccine	802 (73.6)	103 (67.8)	275 (31.2)	11 (16.4)
Age category (6-16 yrs)	685 (62.9)	104 (68.4)	n.a.	n.a.
Chills	149 (13.7)	59 (38.8)	179 (20.3)	38 (56.7)
Cough	692 (63.5)	123 (80.9)	439 (49.7)	53 (79.1)
Ear problems	93 (8.5)	14 (9.2)	87 (9.9)	9 (13.4)
Fatigue	108 (9.9)	30 (19.7)	148 (16.8)	16 (23.9)
Fever	183 (16.8)	74 (48.7)	43 (4.9)	13 (19.4)
Headache	247 (22.7)	55 (36.2)	297 (33.6)	32 (47.8)
Myalgia	62 (5.7)	25 (16.4)	182 (20.6)	33 (49.3)
Runny nose	717 (65.8)	87 (57.2)	497 (56.3)	32 (47.8)
Sinus problems	164 (15.1)	26 (17.1)	358 (40.5)	20 (29.9)
Sore throat	569 (52.2)	82 (54.6)	485 (54.9)	41 (61.2)

Predictors of influenza A or B virus infection in children:

There were a total of 152 influenza positive and 1089 influenza negative episodes in the children derivation set. The following independent variables were statistically significant ($p < 0.05$) in the univariable analyses comparing influenza negative to influenza A or B positive episodes: Age > 5 years, chills, cough, fever and myalgia. These variables and the pre-specified predictors gender and sore throat, were entered into the logistic regression model. Applying stepwise selection the variables that remained significant in the model were age > 5 years, chills, cough, and fever. The final equation for the logistic regression model therefore was:

logit (p) = -3.881 + 0.566 (AGE>5) + 1.212 (CHILLS) + 1.125 (COUGH) + 1.57 (FEVER)

The estimated coefficients changed only slightly when GEE was applied; all coefficients remained statistically significant.

Interpretation of the exponentiated coefficients is the same as in standard logistic regression. According to this the following points were assigned to every predictor remaining in the final model: 1 for age over 5 years, 2 points for each of the predictors, chills and cough, and 3 points for fever (Table 2.2).

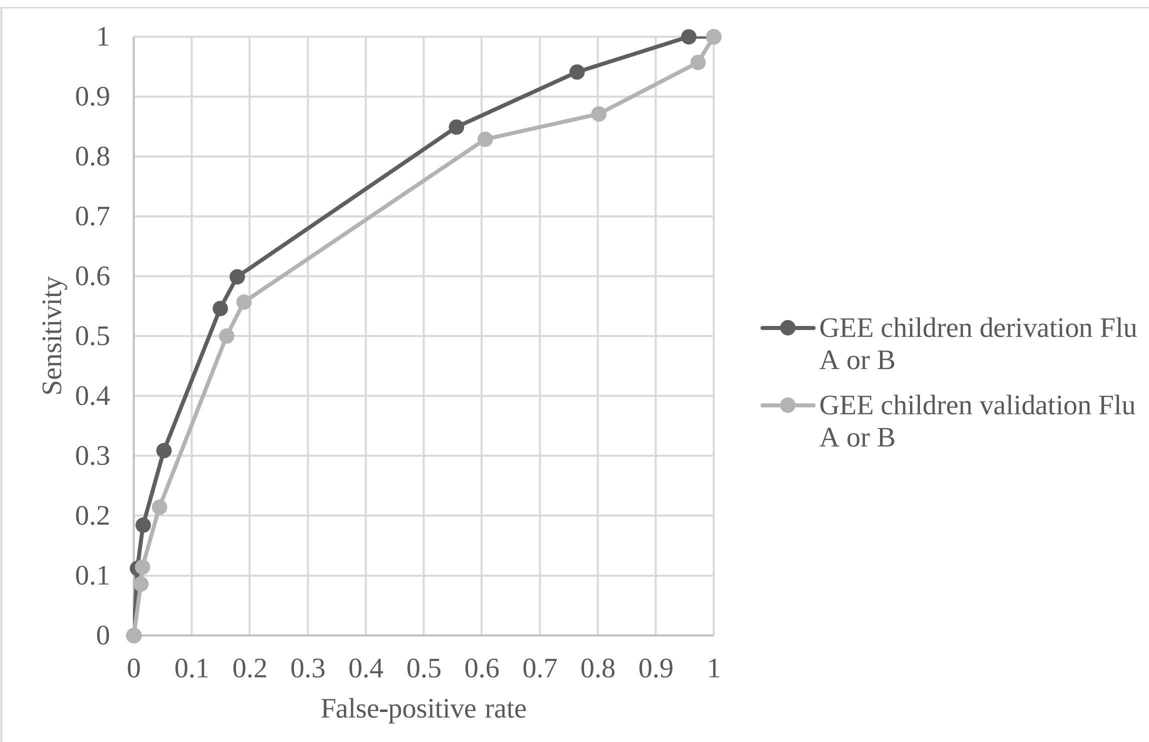
Table 2.2. Influenza A/B score derived from the children training set, GEE model:

	GEE						
	Beta	**SE**	**p-value**	**OR**	**95% CI of the OR**		**Influenza A/B score children**
					lower	**upper**	
Age over 5 years	0.559	0.230	0.006	1.8	1.2	2.6	**1**
Chills	1.202	0.231	<.0001	3.3	2.21	5.02	**2**
Cough	0.764	0.234	<.0001	3.1	2.01	4.74	**2**
Fever	1.362	0.220	<.0001	4.7	3.18	7.04	**3**
Intercept	-3.102	0.239					**-**

Comparison of ROC curves to evaluate the performance of the predictive model

Individual scores were calculated and ROC curves were constructed from the total scores in the derivation set. The same score was then applied to the validation set. The area under the curve (AUC) was 0.76 (95% CI 0.72 - 0.80) for the derivation cohort and 0.70 (95% CI 0.63-0.77) for the validation cohort (Fig. 2.1). The differences in the AUCs of the derivation and validation cohort was not statistically significant with 0.06 ($z = 1.383$, $p = 0.166$). With a score of ≥ 4 for influenza A or B positivity the sensitivity and specificity were 60% and 82%, respectively in the derivation set and 56% and 81%, respectively in the validation set.

Figure 2.1 ROC curves for the comparison of performance of the influenza A/B score in children



Predictors of influenza A or B virus infection in adults:

In the derivation data set there were a total of 883 influenza negative and 67 influenza A or B positive episodes. The following independent variables were statistically significant ($p < 0.05$) in the univariable analyses comparing influenza negative to influenza A or B positive episodes: Chills, cough, fatigue, fever, headache, myalgia, and sore throat. These variables as well as the predefined variable gender were entered into the logistic regression model applying stepwise selection. The combined symptoms fever and cough, as well as fever and sore throat, were also significant ($p<0.05$) in the univariable analysis with OR (95% CI) 10.84 (4.99-23.55) and 4.01 (1.57-10.23), respectively. However, due to collinearity, they were not entered into the logistic regression model.

The final equation for the logistic regression model therefore was:

logit (p) = -4.449 + 1.465 (CHILLS) + 1.534 (COUGH) + 0.968 (MYALGIA)

Generalized Estimating Equations (GEE) was performed to account for correlated measures. All variables remained statistically significant (Table 2.3). The score from the beta-coefficients, weighted by the lowest absolute value and rounded to the nearest integer assigned 2 points for each chills and cough and 1 point for having myalgia (Tab 2.3).

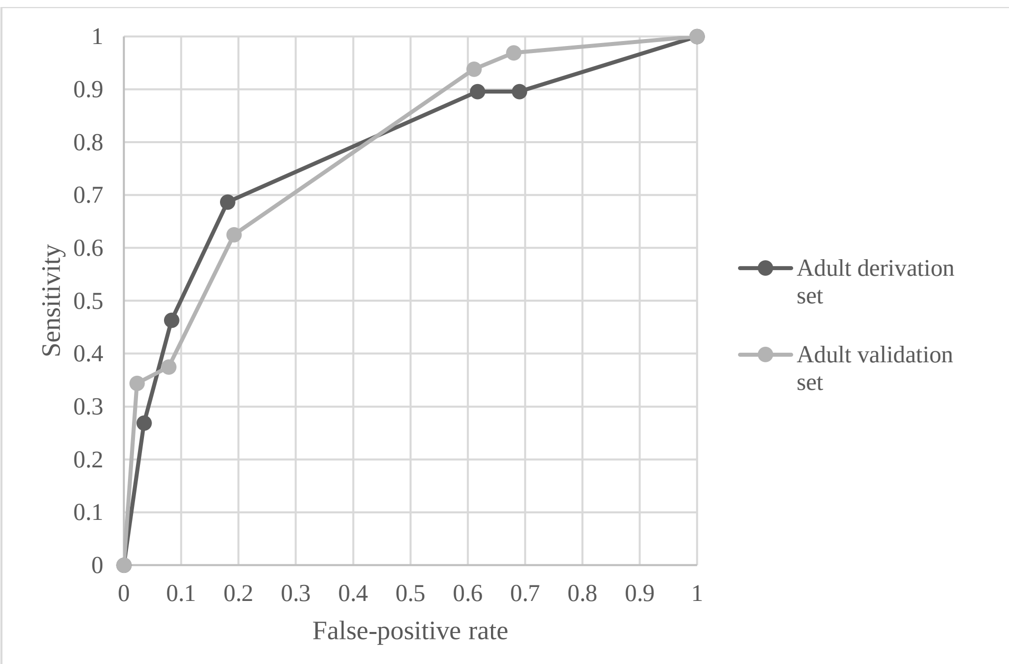
Individual scores were calculated and ROC curves were constructed from the total scores in the derivation set. The same score was then applied to the validation set. The area under the curve (AUC) was 0.78 (95% CI 0.72-0.85) in the derivation set and 0.79 (95% CI 0.71-0.87) in the validation cohort (Fig 2.2). The differences in the AUCs of the derivation and validation set was not statistically significant (difference: $area_{deriv}$-$area_{valid}$ = -0.01; $z = -0.169$, $p = 0.866$). The optimal sensitivity (69%) and specificity (82%) in the derivation set were found at a score of ≥ 3.

The corresponding sensitivity and specificity were 63% and 81%, respectively in the validation set.

Table 2.3. Influenza A/B score derived from the adult training set, logistic regression/GEE model:

	GEE						
	Beta	**SE**	**p-value**	**OR**	**95% CI of the OR lower**	**upper**	**Influenza A/B score adults**
Chills	1.465	0.287	<.0001	4.3	2.47	7.60	**2**
Cough	1.535	0.328	<.0001	4.6	2.44	8.83	**2**
Myalgia	0.967	0.287	0.001	2.6	1.50	4.62	**1**
Intercept	-4.452	0.367					-

Figure 2.2 ROC curves for the comparison of the performance of the influenza A/B score in adults



2.3.2 Prediction of influenza A/H3N2 subtype

Predictors of influenza A/H3N2 in children:

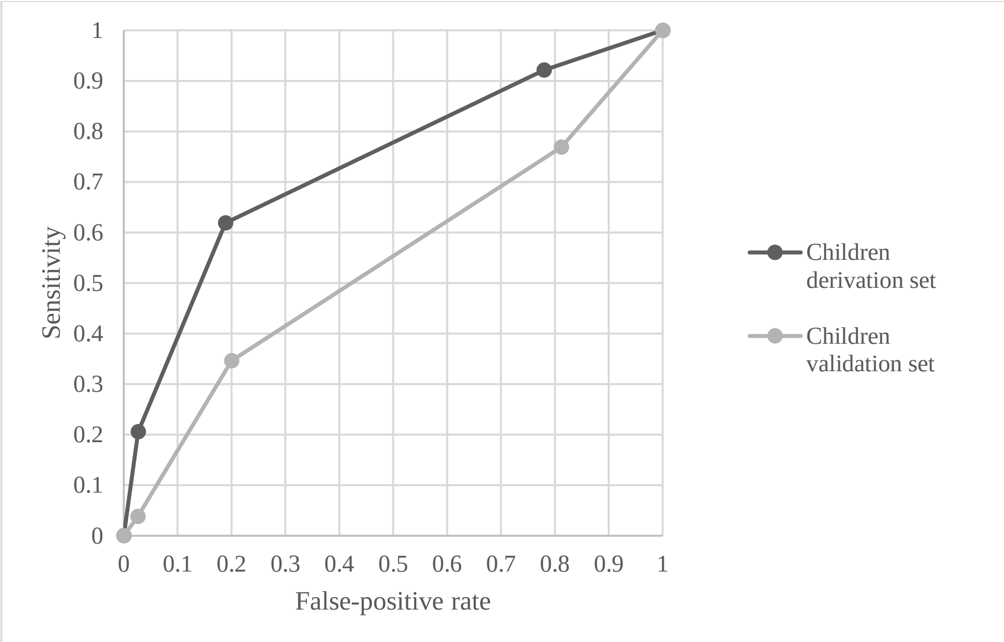
There were 63 (5.1%) influenza A/H3N2 positive episodes in a total of 1241 first ARI episodes. Of the 16 predictors of primary interest 5 (chills, cough, fever, myalgia, and acute onset) were found to be statistically significant ($p < 0.05$) in the univariable analyses comparing influenza A/H3N2 positive subjects to all other individuals in the derivation set. Including the pre-specified predictors gender and sore throat, multivariable logistic regression was performed using forward stepwise selection. The variables remaining in the final model were influenza chills, fever, and cough (Tab. 2.4). All predictors remained significant and the standard errors of the predictors changed only marginally when using GEE.

The scores assigned to every independent predictor were as follows: 1 for each, chills, fever and cough. The AUC of the ROC curve in the derivation set was 0.74 (SE: 0.04; 95% CI 0.67- 0.81) but was only 0.54 (SE: 0.07; 95% CI 0.41 -0.67) in the validation set (Fig. 2.3). This difference of 0.19 was statistically significant (z= 2.73; p= 0.006). Optimal sensitivity and specificity were 62% and 81%, respectively in the derivation set, and 35% and 80%, respectively, in the validation set, corresponding to a score of 2 points or higher.

Table 2.4 Influenza A/H3N2 score derived from the children training set

Model	**GEE**						
	Beta	**SE**	**p-value**	**OR**	**95% CI of the OR lower**	**upper**	**Influenza A/H3N2 score children**
Chills	1.307	0.308	<.0001	3.7	2.02	6.76	1
Cough	1.043	0.355	.003	2.8	1.41	5.69	1
Fever	1.257	0.278	<.0001	3.5	2.04	6.06	1
Intercept	-4.483	0.379					

Figure 2.3 ROC curves for the comparison of the performance of the influenza A/H3N2 score in children



Predictors of influenza A/H3N2 in adults:

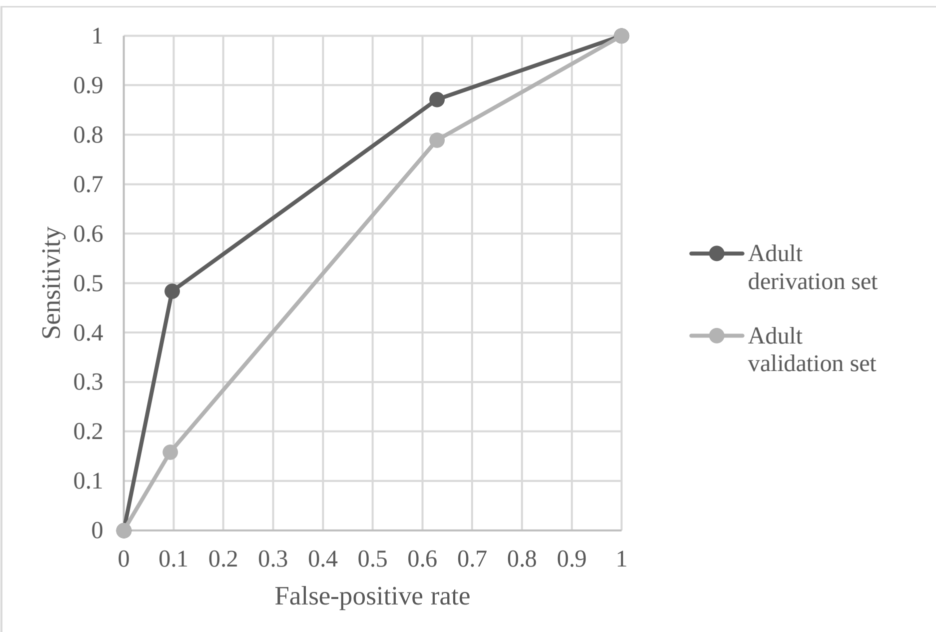
Of the 15 predictors of primary interest 5 (chills, cough, fever, myalgia, and ear problems) were found to be statistically significant ($p < 0.05$) in the univariable analyses comparing influenza A/H3N2 positive subjects to all other individuals in the adult derivation set. Including the pre-specified predictor gender and sore throat, forward stepwise multivariable logistic regression was performed. The variables remaining in the final model were chills and cough. Although fever was highly significant in the univariable analysis it was eliminated in the multivariable analysis. Both predictors remained significant after accounting for clustering (Tab. 2.5). The score assigned to the independent variables were as follows: 1 point for each chills and cough. The ROC curve derived from the total score had an AUC of 0.73 (SE 0.05, 95% CI 0.63-0.83) in the derivation set and an AUC of only 0.59 (SE 0.06, 95% CI 0.47-0.72) in the validation set for a statistically non-significant difference between the AUCs of 0.14 ($z = 1.605$, $p= 0.108$) (Fig. 2.4).

Optimal sensitivity and specificity when inspecting the ROC curves were 48% and 90%, respectively in the derivation set and 16% and 90%, respectively, in the validation set.

Table 2.5 Influenza A/H3N2 score derived from the adult training set:

	GEE						
	Beta	**SE**	**p-value**	**OR**	**95% CI of the OR lower**	**upper**	**Influenza A/H3N2 score adults**
Chills	1.436	0.377	<.0001	4.2	2.01	8.80	1
Cough	1.701	0.508	0.001	5.5	2.03	14.82	1
Constant	-5.085	0.556	<.0001	0.0	0.00	0.02	

Figure 2.4 ROC curves for the comparison of the performance of the influenza A/H3N2 score in adults



2.3.3 Prediction of entero-/rhinovirus (ERV) infections

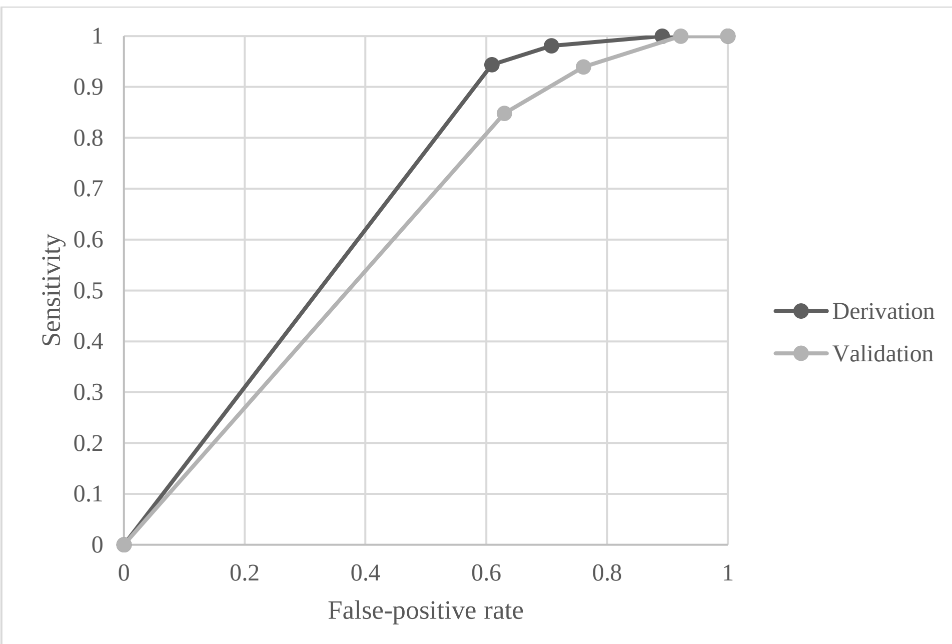
The data set comprised 653 children under 16 years with a first episode of ARI in season one. Of these, 14 children (2.1%) had a missing outcome (were not tested) and were therefore excluded from the analyses. Thus, there were 639 remaining children and of these 87 (14%) were tested ERV positive. Of the 552 children tested ERV negative, 114 (21%) were influenza A or B virus positive. At first, we randomly assigned approximately 60% of children to the derivation and 40% to the validation group. The derivation group consisted of 322 ERV negative and 54 ERV positive individuals, the validation group consisted of 230 ERV negative and 33 ERV positive children. The following independent variables were statistically significant in univariable analysis of the derivation group: chills (OR 0.15, 95% CI 0.04-0.62) and fever (OR 0.05, 95% CI 0.01-0.34); runny nose was marginally statistically significant (p= 0.050) with an OR 2.04 (95% CI 0.99-4.22) and was also entered in the multivariable model. We also pre-specified that cough and age category under 6 years, which were both not statistically significant, would be entered into the model as binary predictors (1=yes, 0=no).

Forward stepwise logistic regression selected chills and fever as significant predictors (Tab. 2.6). The Hosmer and Lemeshow test was not statistically significant (p= 0.932), indicating that the model had a good fit. The presence of either chills or fever significantly reduced the odds of an ERV infection. E.g. in children with fever the odds of an ERV infection was 0.06 the odds of children without chills holding the other variable constant (Tab. 2.6). A point score was created from the coefficients to classify each ARI case into either the ERV negative or ERV positive category assigning: -1 points for presence of chills and -2 point for fever. The ROC curve constructed from the total score had an AUC of 0.67 (95% CI 0.61-0.74) in the derivation set, and an AUC of 0.62 (95% CI 0.53-0.71) in the validation set (Fig. 2.5). The difference between the AUCs was not statistically significant (p=0.47). With a total score of 0 (= ARI without fever and chills) for the diagnosis of ERV infection, sensitivity and specificity were 94.4% and 39.1%, respectively in the derivation set, and 84.8 % and 37.0%, respectively in the validation set.

Table 2.6 Predictors of ERV infection in children in standard logistic regression

	Beta	SE	p-value	OR	95% CI of the OR		Score
Chills	-1.493	0.744	0.045	0.23	0.05	0.97	-1
Fever	-2.861	1.02	0.005	0.06	0.01	0.42	-2
Constant	-1.343	0.157					

Fig. 2.5 Comparison of ROC curves between derivation and validation set for the performance of the ERV score in children < 16 years



References:

1. Kleinbaum, D.G., et al., *Applied regression analysis and other multivariable methods*. Fifth edition. ed. 2013, Boston, MA: Cengage Learning. xix, 1051 pages.
2. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.

CHAPTER III

3.0 Decision tree models (DTM) and ensemble methods

3.1 Recursive partitioning

Recursive partitioning is a form of multivariable analysis which does not require specification of a parametric model to develop a prediction rule in a systematic way. One major advantage of decision tree analysis is the relative simple integration of complex interactions that are usually avoided in parametric models [1]. Outcomes are thereby predicted by dividing the data into disjoint (mutually exclusive) subsets based on the independent variables and “recursive” reflects the continued splitting within these subsets aiming to further improve predictions on the response variable until no further splitting is done. These so called *terminal* subsets form a partition of the primary split [2]. The generated algorithms are either classification or regression trees depending on whether the outcome of interest is categorical (binary) or continuous, respectively.

Generating the tree classifier: The principle of each splitting of a parent node into a daughter node is to reduce impurity. Each daughter node isolates subjects with a majority of either response to become “purer” than the parent subset [2]. The difference between the impurity of the parent node and the mean impurity in the two daughter nodes indicates the degree of impurity reduction [1].

The next question is how to find the best split: according to *Breiman et al*. a splitting rule specifies a goodness of split function Φ (s, t) for every $s \in S$ and node t where s denotes a split and S is the set of splits; for every t, one has to find the split s^* that minimizes tree impurity or, equivalently, reduces the estimated misclassification rate [2]. There exist different node impurity measures, i(t), such as the misclassification error, the Gini criterion (replaced in later work by the Gini diversity index), and the deviance (cross-entropy) [3, 4].

The Gini (diversity) index is a popular measure of quantification of each nodal impurity in CART and is the default in R and SPSS statistical software [4].

The Gini index has the form: $i(t) = \sum_j p\,(j|t)\ \sum_{k \neq j} p\,(k|t)$ (3.1.1)

and is the estimated probability of misclassification (= error rate) of observations that fall in node t, where $p(j|t)$ = estimated probability that an observation is actually in class j given that it falls in node t and $p(k|t)$ = estimated probability that an observation is in class i given that it falls in node t.

The Gini index can also be written as: $i(t) = 1 - \sum_j p^2\,(j|t)$ (3.1.2)

A Gini index of 0 signifies perfect discrimination, meaning that only class j observations are classified in node t [$p(j|t) = 1, p(k|t) = 0$] whereas it reaches its maximum node impurity when the probability to fall into a node for each class is equal.

The optimal split s is chosen that reaches the largest decrease in the Gini index. This can be written as: $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ (3.1.3)

Where $p_L i(t_L)$ and $p_R i(t_R)$ are the disjoint empirical probabilities that an observation falls into node t_L or t_R, respectively ($p_R = 1 - p_L$) [2, 4].

After the trees have grown as large as possible splits that are deemed uninfluential are eliminated, a process that is called “pruning” [4].

3.2 Ensemble methods

One limitation of a single tree model is its instability to small changes in the learning data which may result in high variability of the predictions [5].

Ensemble methods, also known as bagging and random forests, are methods in classification and regression trees (CART) where the predictions are based on a set of different trees created from random samples of the derivation set. These samples are slightly different from the original due to random variation. The random samples are usually drawn by the bootstrap method and a large individual tree is generated on every random sample usually without pruning or stopping. The predictions of these different trees are then combined and are considered to produce a less biased average prediction. One drawback of these methods, however, is that the generated trees are not as simple to interpret.

3.3 Model development

The same predictors of influenza virus infection were analyzed by recursive partitioning (RP) in order to classify patients with certain predictors into the risk groups of influenza virus infection or non-influenza virus related ARI.

All independent variables whether or not significant in univariable analysis were entered into the decision tree model for the prediction of ERV infection in children.

The primary approach for validation will be again, using data-splitting and perform RP to the learning set and test the model in the validation set There is only a dichotomous (e.g. yes/no) prediction in every node. The primary evaluation of model performance was the comparison of ROC curves of predicted probabilities generated from the derivation set to the ROC curves when the model was applied to the validation set. The differences in model performance were examined by comparing the AUCs using the z-score test [6].

In order to be more consistent with the other models and to have a tool to compare the predictive indexes between the different mathematical models at different cut-points, it was planned to generate a simplified score from the DTMs where the rank of the predictor, e.g. the node level at which the predictor appeared, actually determined the magnitude of the score (the higher in the node hierarchy the higher the score). One major downside, however, of these simplified scores is, that potential interaction effects are ignored which can lead to biased or false predictions. It was therefore planned that corresponding ROC curves were constructed from the scores to examine the discrepancy between the score and the predicted probabilities from the original models by visual guess and, provided there was no obvious discrepancy, to determine cut-points for sensitivity and specificity. However, since all models were rather complex and in view of the negation of the previously mentioned advantages of the DTM about discovering interaction effects which would not adequately be taken into account, no point scores were eventually assigned.

3.4 Results

3.4.1 Decision tree building for predicting influenza A/B

DTM for the prediction of influenza A/B in children

The CRT function in SPSS defined fever, chills, cough, runny nose, and gender as the most important predictors of influenza virus A/B infection in the derivation data set and created a rather complex tree with 8 terminal nodes (Fig. 3.1 a.- b.).

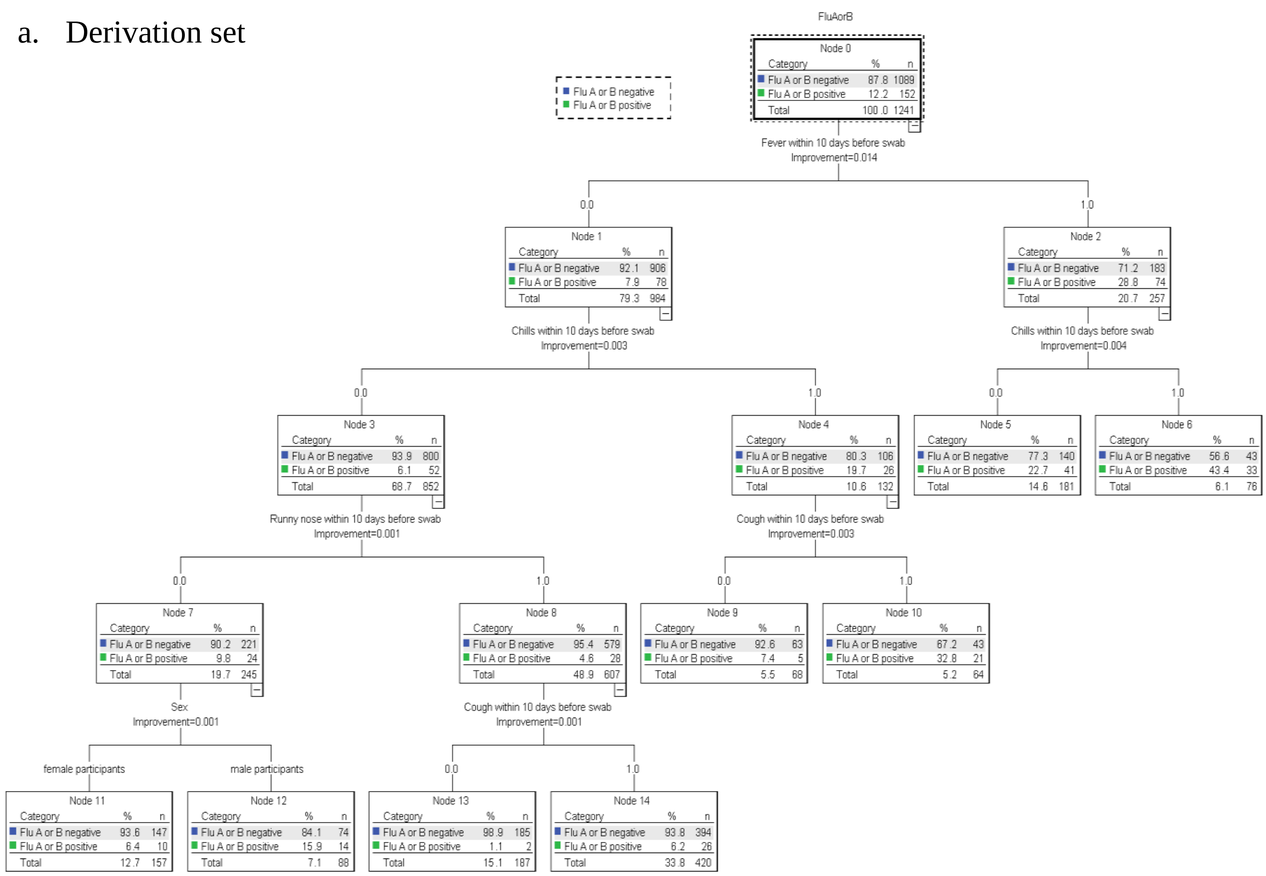
Overall, the null model without any predictors would incorrectly predict 12.2% of the cases as being influenza negative with an overall percentage classified correctly of 87.8%. Fever is the most important predictor, as it is listed in the first decision node. In those without fever 92.1% of all episodes will be correctly classified as influenza negative; however, the node also shows that 7.9% (78 influenza positive episodes) would be missed, which is more than half of all influenza positive (n=162) episodes. Assuming that those without fever and no chills were influenza negative would classify 93.8% correctly but would also incorrectly predict 23 (6.3%) episodes that were in fact influenza A/B positive. Contrariwise, among those with fever and chills 43% (14 episodes) would be classified correctly as influenza positive. The tree model reveals several potential interaction effects such as between fever and chills, chills and cough or chills and runny nose.

When pruning was applied, none of the predictors were selected since the model without any predictors already correctly classified 88.6% of all episodes as influenza negative as indicated by the null model without any predictor. This procedure, however, turned out to be not helpful, since the goal was to improve prediction of influenza positive episodes.

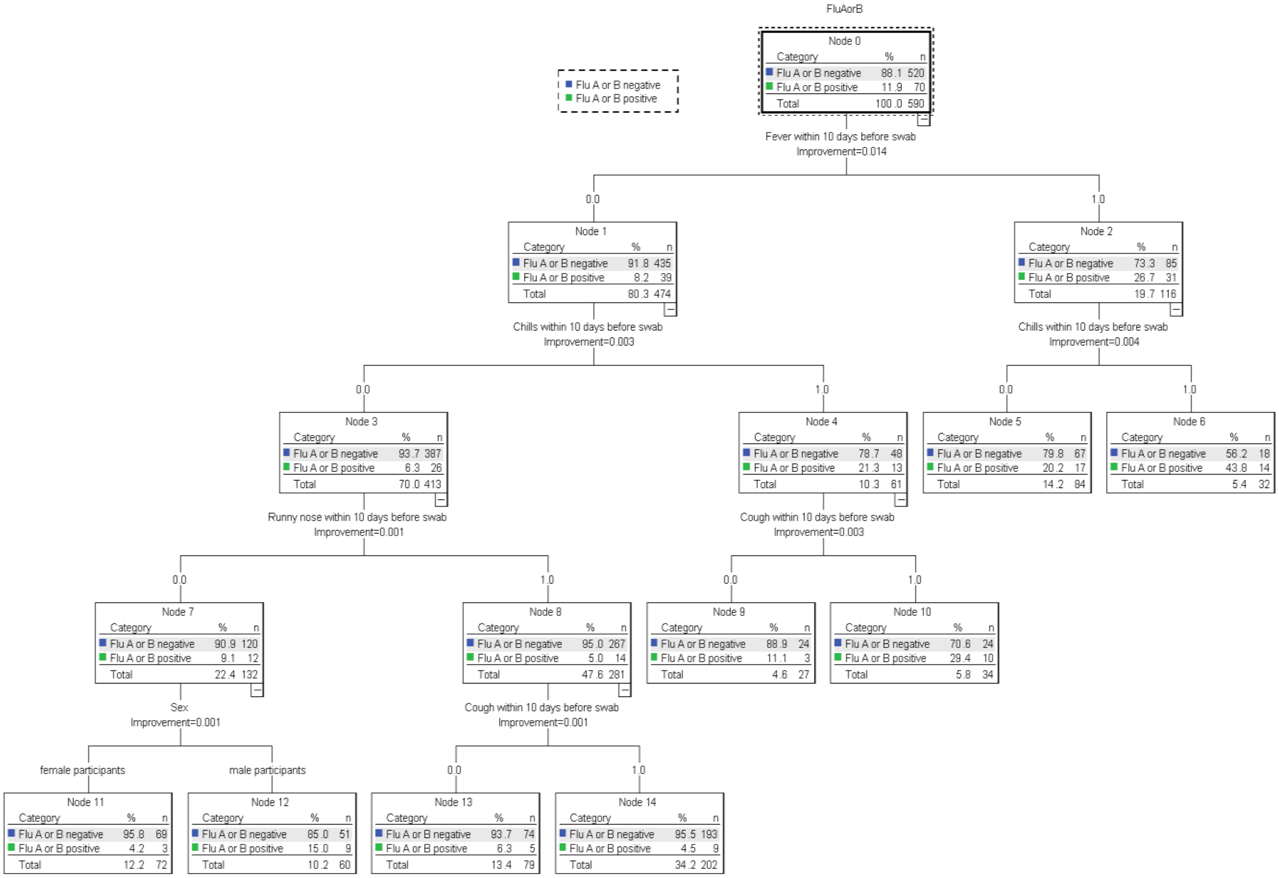
The model was then applied to the validation set and the percentages classified correctly were similar (88.1% in the validation set vs. 87.8% in the derivation set). In the terminal nodes, the percentages classified as influenza positive were similar between the derivation and validation set.

Figure 3.1 (a-b). Unpruned classification trees for predicting influenza A/B virus infection in children (a. Derivation set; b. Validation set).

a. Derivation set



b. Validation set

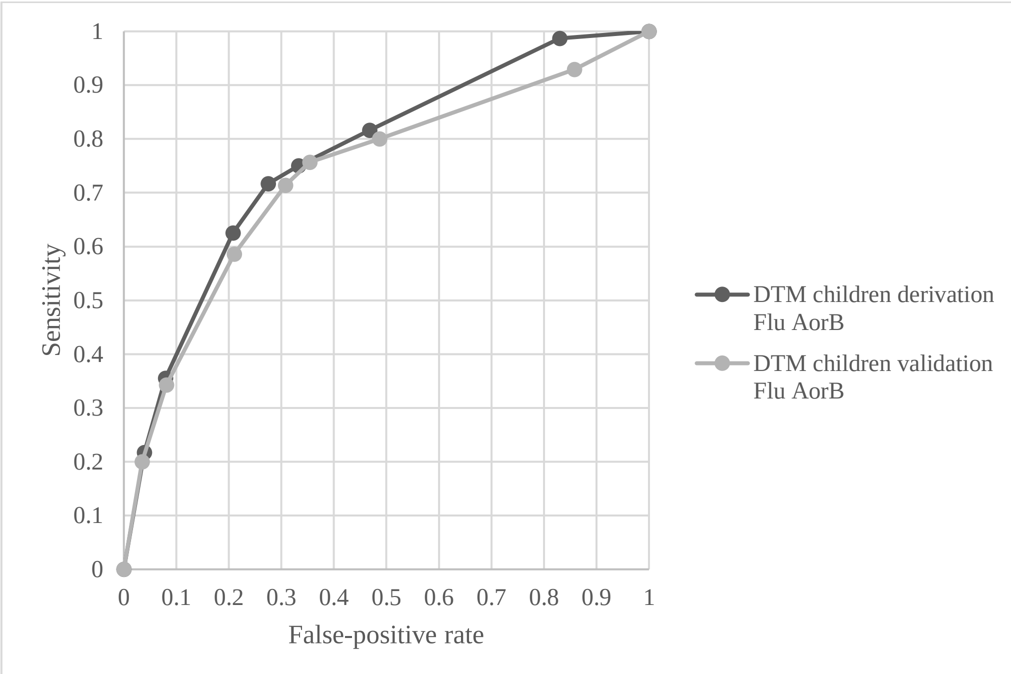


Comparison of the DTM performance

ROC curves were generated from the predicted probabilities of being influenza A or B positive and compared between the derivation and validation group.

The corresponding AUCs of the derivation and validation group were 0.77 (95% CI 0.73-0.81) and 0.74 (95% CI 0.67-0.80), respectively (Fig. 3.2). The difference between AUCs of 0.03 was not statistically significant (p=0.45). When eyeballing the ROC curves optimal sensitivities and specificities were 72% and 72%, respectively in the derivation set, and 71% and 70%, respectively, in the validation set.

Figure 3.2 Comparison of ROC curves generated from DTM to predict influenza A/B in children



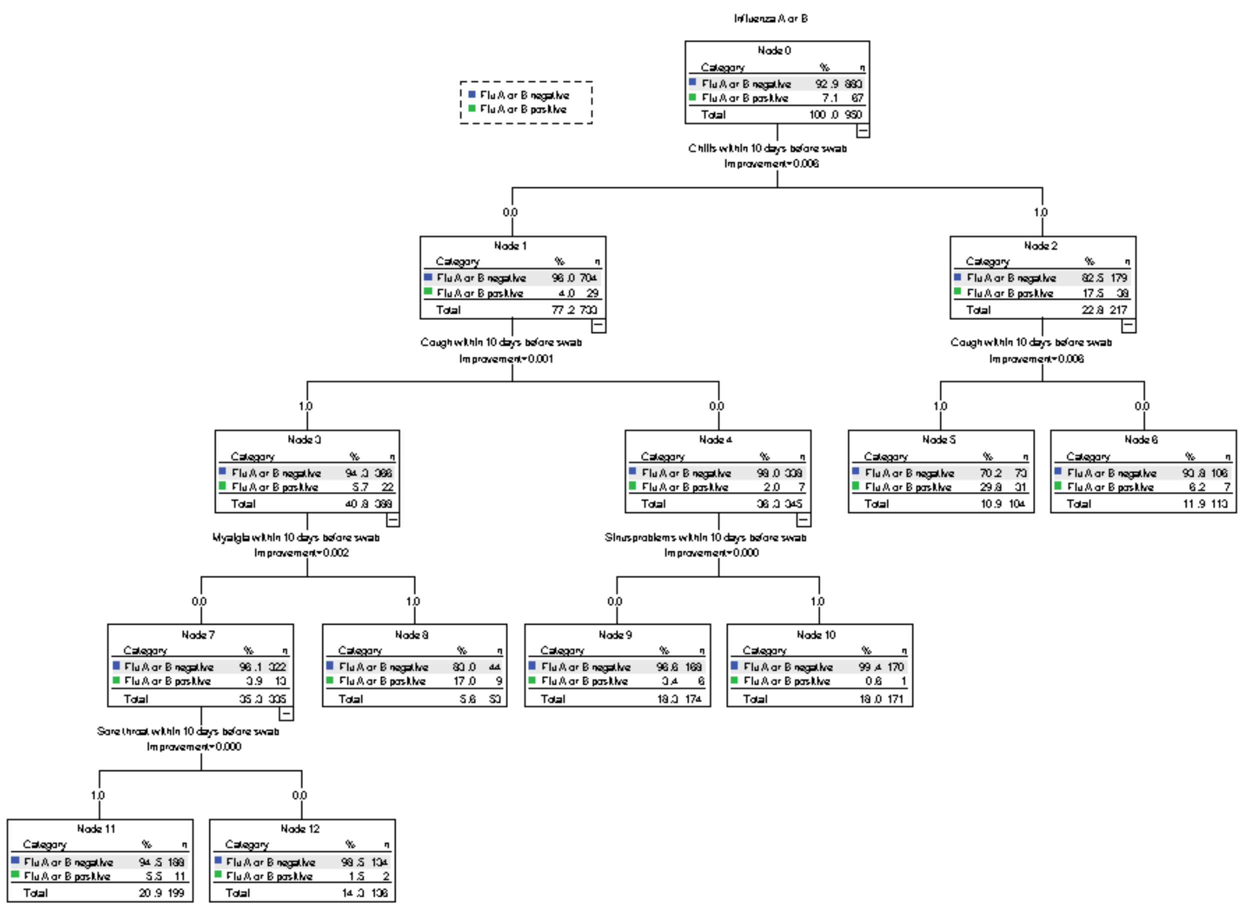
DTM for the prediction of influenza A/B in adults

The derivation set consisted of 883 influenza A/B negative and 67 influenza A/B episodes, whereas the validation set comprised 475 negative and 32 influenza A/B positive episodes. A decision tree model was built to predict the binary outcome influenza A or B infection by entering all independent categorical variables into the model. The results from the derivation set show that chills is the most important symptom and therefore appears in the primary decision node followed by cough and myalgia for a total of 7 terminal nodes (Fig 3.3 a-b).

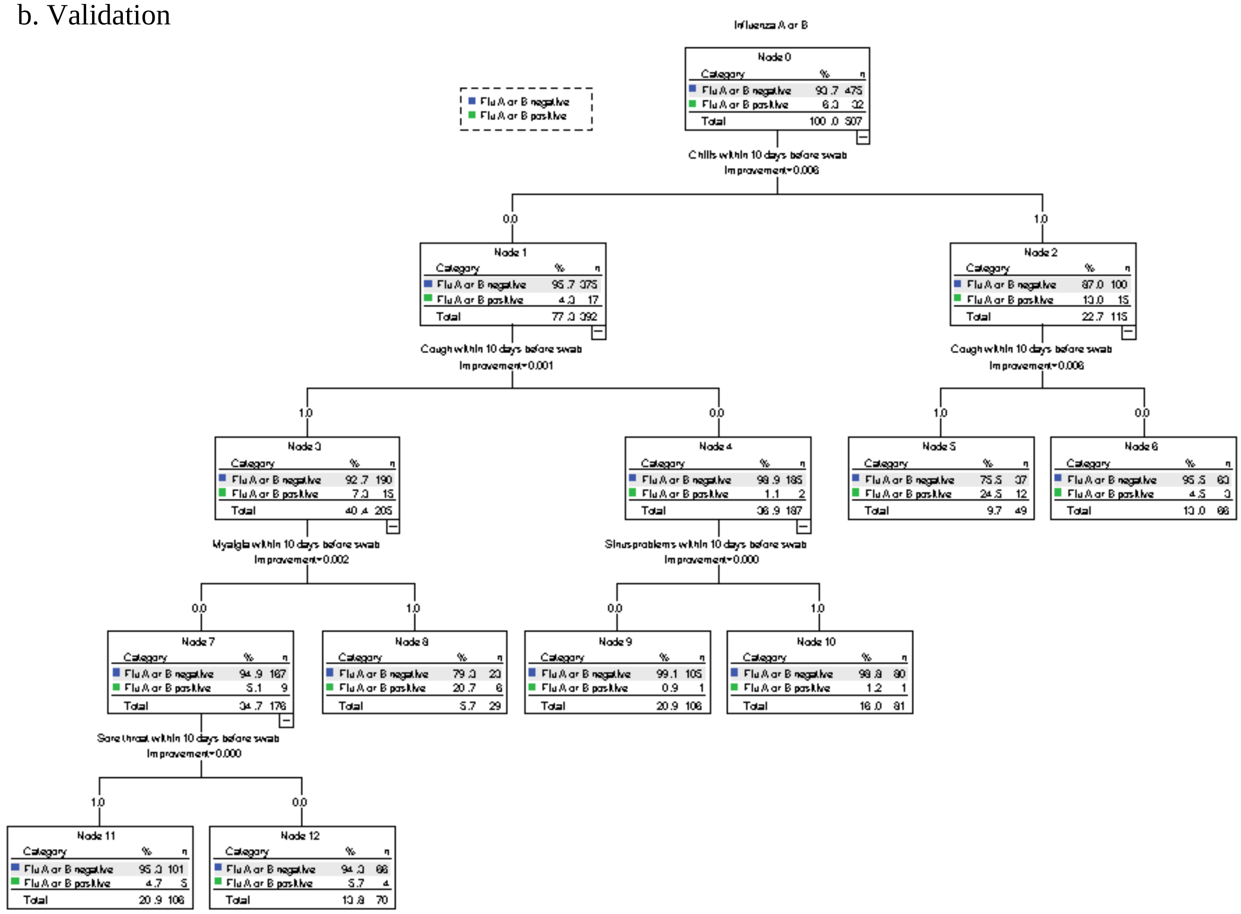
The model without any predictors would have been correct in 92.9% classifying all episodes as influenza negative. Considering all episodes without chills and cough but sinus problems the percentage of correctly classifying all episodes as influenza negative increases to 99.4%. However, among those without chills but cough and myalgia the proportion of influenza negatives episodes decreases to 83.0%. The highest proportion (29.8%) of influenza positive episodes was found among those presenting with chills and cough. The DTM also suggests, that there are potential interaction effects e.g. between cough and myalgia, since myalgia has only a predictive value for the diagnosis of influenza A or B in the presence but not in the absence of cough.

Figure 3.3 (a-b) DTM for the prediction of influenza A/B in adults

a. Derivation



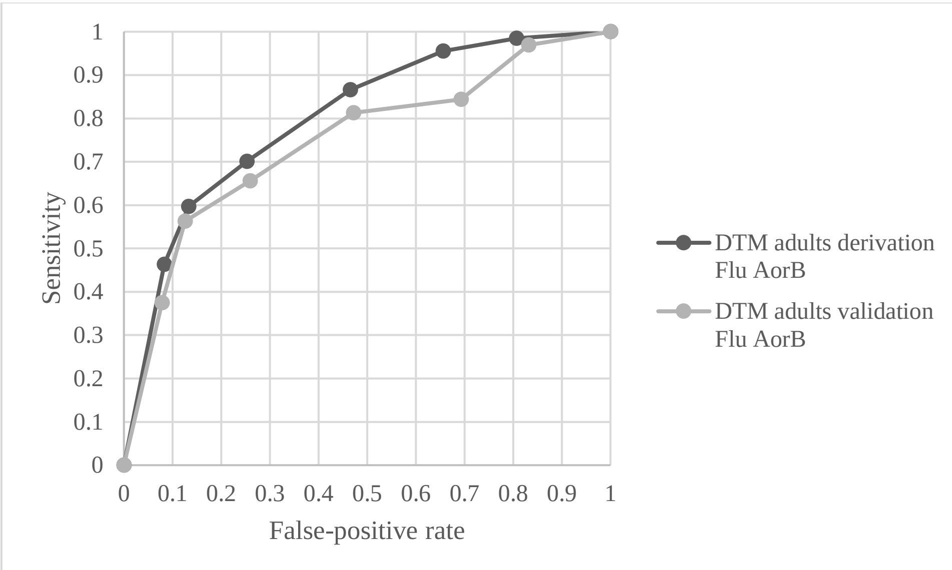
b. Validation



Comparison of DTM performance

The ROC curves generated of the predicted probabilities showed an AUC of 0.80 (95% CI 0.75-0.86) in the derivation set and an AUC of 0.75 (95% CI 0.65-0.85) in the validation set for a non-statistically significant difference (p=0.41). When eyeballing the curves optimal sensitivity and specificity were 70% and 75%, respectively, in the derivation set, and 66% and 74%, respectively, in the validation set (Fig 3.4).

Figure 3.4 Comparison of ROC curves between derivation and validation data set in adults to predict influenza A/B (Method: predicted probabilities from DTM)



3.4.2 Decision tree model for predicting influenza A/H3N2 subtype

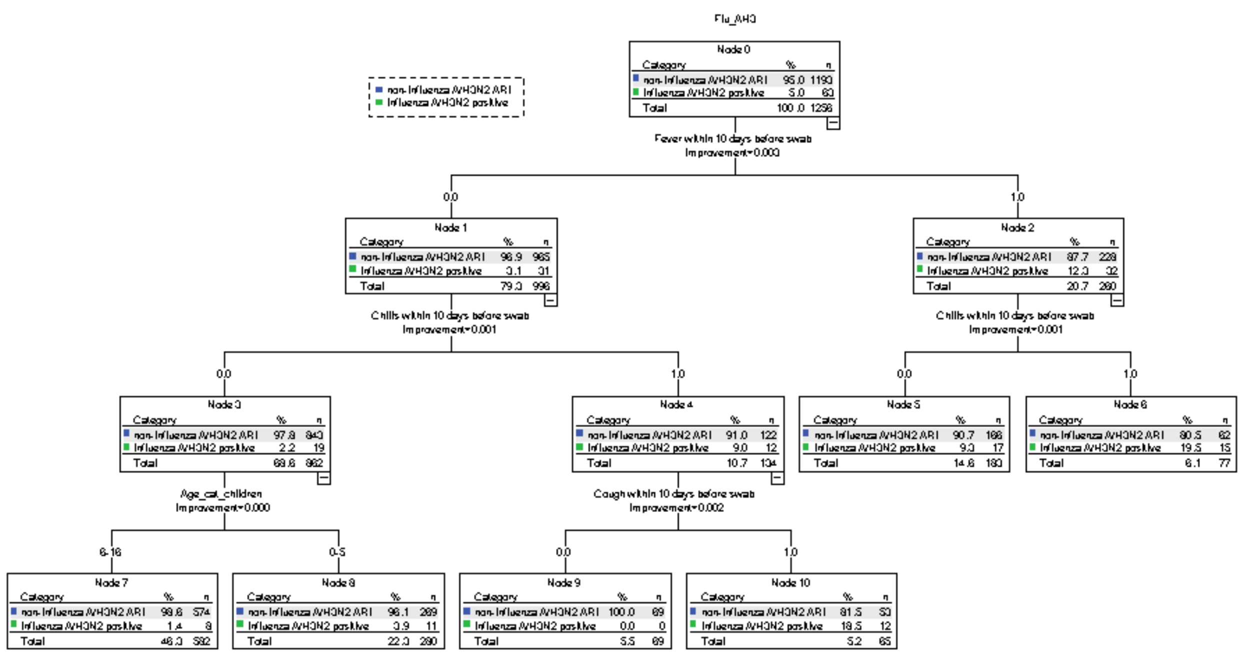
DTM for predicting influenza A/H3N2 in children.

The DTM for predicting influenza A/H3N2 versus all other ARIs was very similar than the one for influenza A or B and revealed that fever, chills, cough and age group were the most important predictor of influenza A/H3N2 virus infection (Fig. 3.5 a.-b.). The model has 6 terminal nodes. Considering the null model without any predictors, the risk of an influenza A/H3N2 positive episode was only 5% and the overall percentage predicted correctly was 95%. Among those episodes with fever and chills the risk of influenza A/H3N2 increases to 19.5%. Among those without fever but chills and cough the risk of influenza A/H3N2 increased to 18%. In contrast, the proportion of influenza positive episodes in children 6 years and older without fever and chills decreased to 1.4%. The proportion of influenza positive episodes among those with fever and chills was, however, lower in the validation set (12.1%).

ROC curves constructed from the predicted probabilities revealed worse performance of the model in the validation set with corresponding AUCs of 0.79 (95% CI 0.73-0.84) in the derivation and 0.62 (95% CI 0.51-0.74) in the validation set, respectively (Fig. 3.6). The difference of 0.16 between the two AUCs was statistically significant (z=2.34, p=0.019). Overall, the improvement in prediction was marginal and after pruning, all predictors were eliminated, implying that none of them was important enough to improve prediction of the outcome influenza A/H3N2 virus infection and the previously found importance of certain predictors are likely chance findings.

Figure 3.5 (a-b) Decision tree model for predicting influenza A/H3N2 in children

a. Derivation



b. Validation

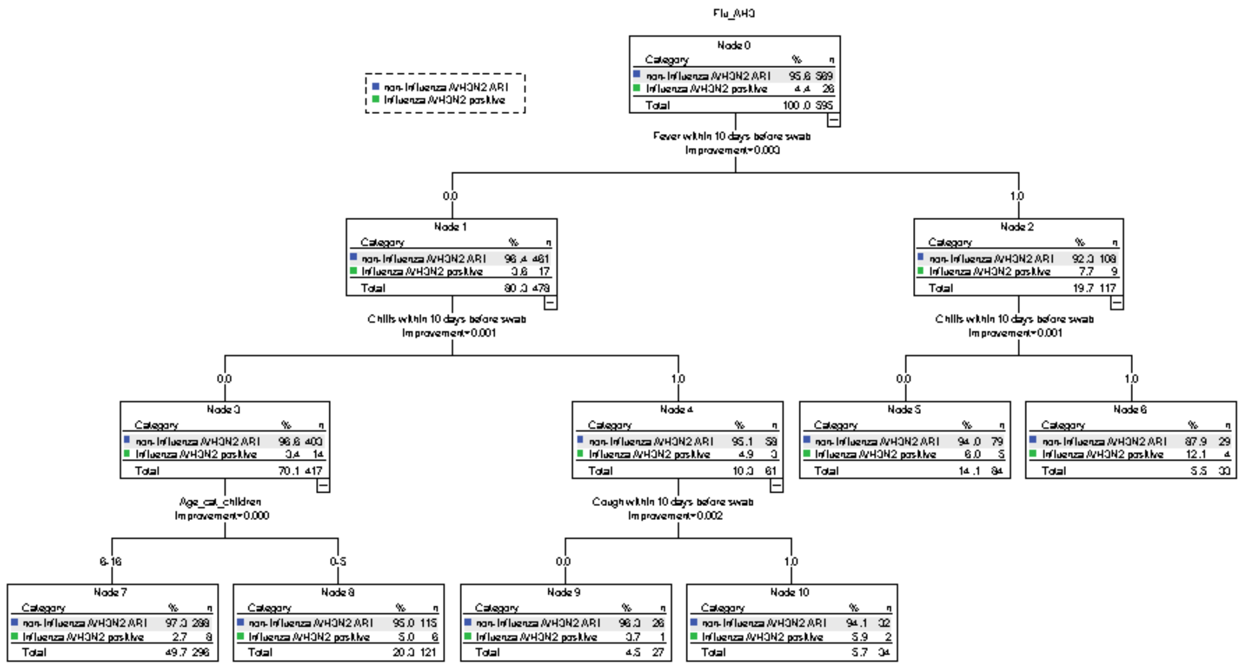
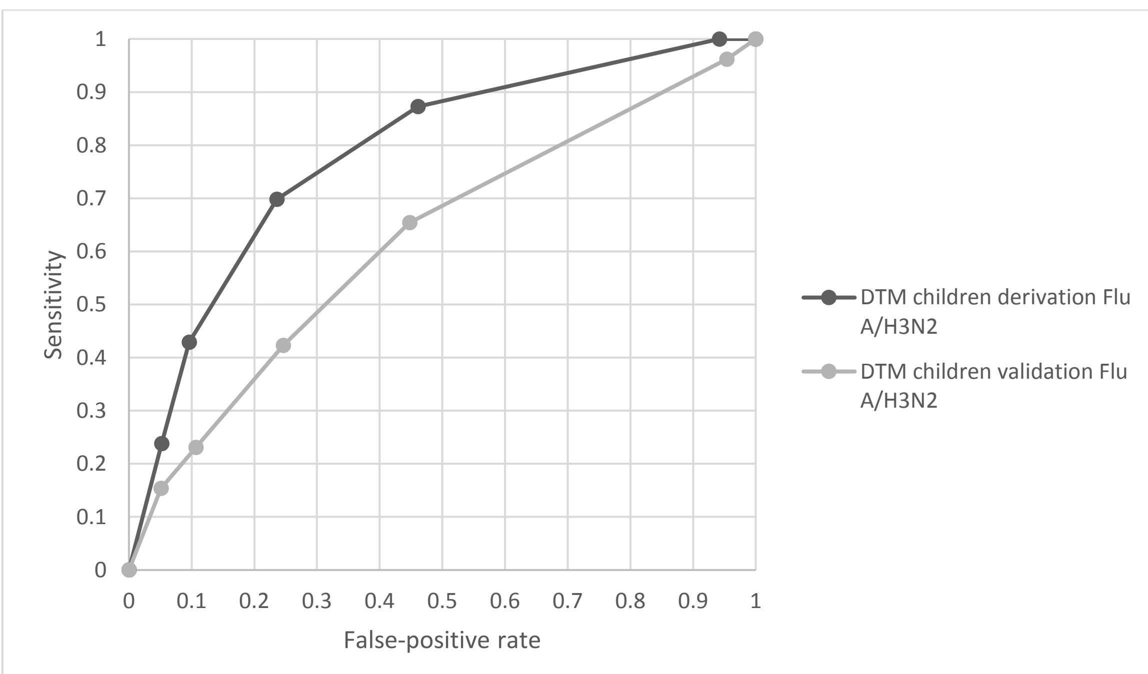


Figure 3.6 Comparison of ROC curves of unpruned DTM for the prediction of influenza A/H3N2 in children

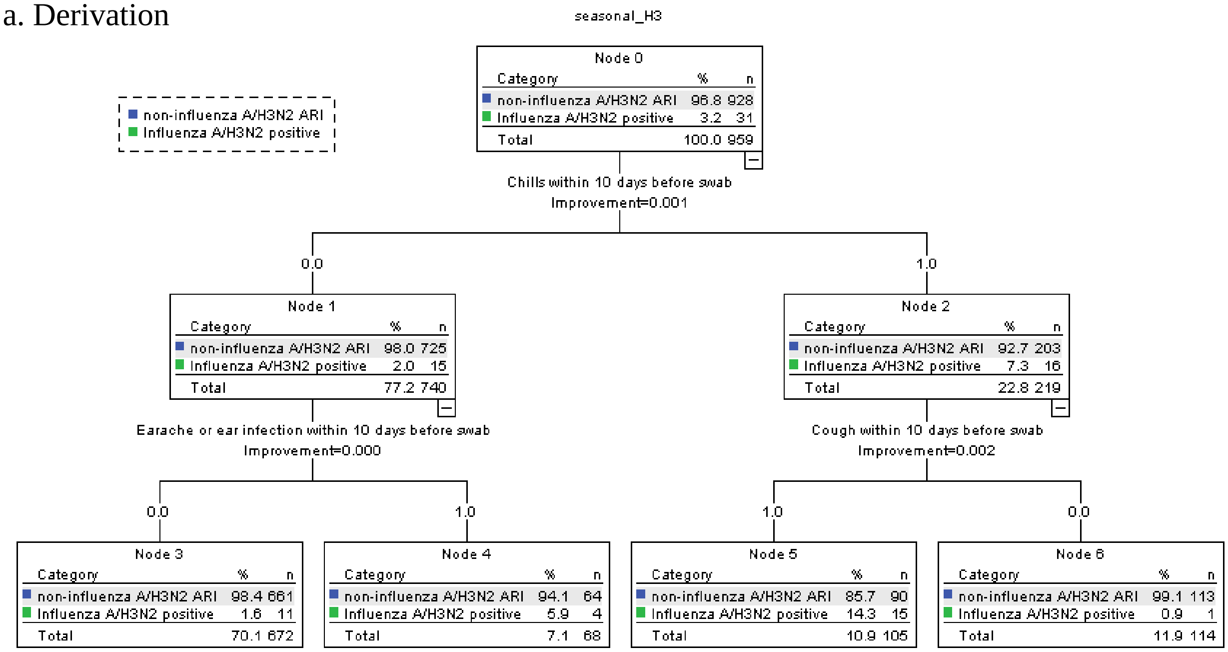


Decision tree for influenza A/H3N2 in the adult derivation set:

The DTM built for the prediction of influenza A/H3N2 created a tree with chills, cough, and ear problems (ear ache or ear infection) as the remaining predictors in the derivation set, whereas all other variables were eliminated. As in the DTM for influenza A/B in adults fever was not chosen as important variable. The overall percentage predicted correctly was 96.3% and the 2 remaining predictors only minimally improved the model (Fig. 3.7 a.-b.). In the absence of chills but presence of ear problems the risk of influenza A/H3N2 was lowest (2.6%) and increased to 6.1% in the presence of chills and cough. Whilst chills was chosen as most important predictor in the decision tree model, the normalized importance plot actually shows that fever and cough are the most important predictors. In the model, fever likely acted as surrogate variable (Fig 3.8). This means that the variable fever was only considered a substitute of chills and probably gave a second or third best split. One can try to either completely remove chills from the model or force fever into the model to examine whether the model has a better performance but other measures exist which are beyond the scope of this thesis [2, 4].

Figure 3.7 (a.-b.) Decision tree model for predicting influenza A/H3N2 in adults

a. Derivation



b. Validation

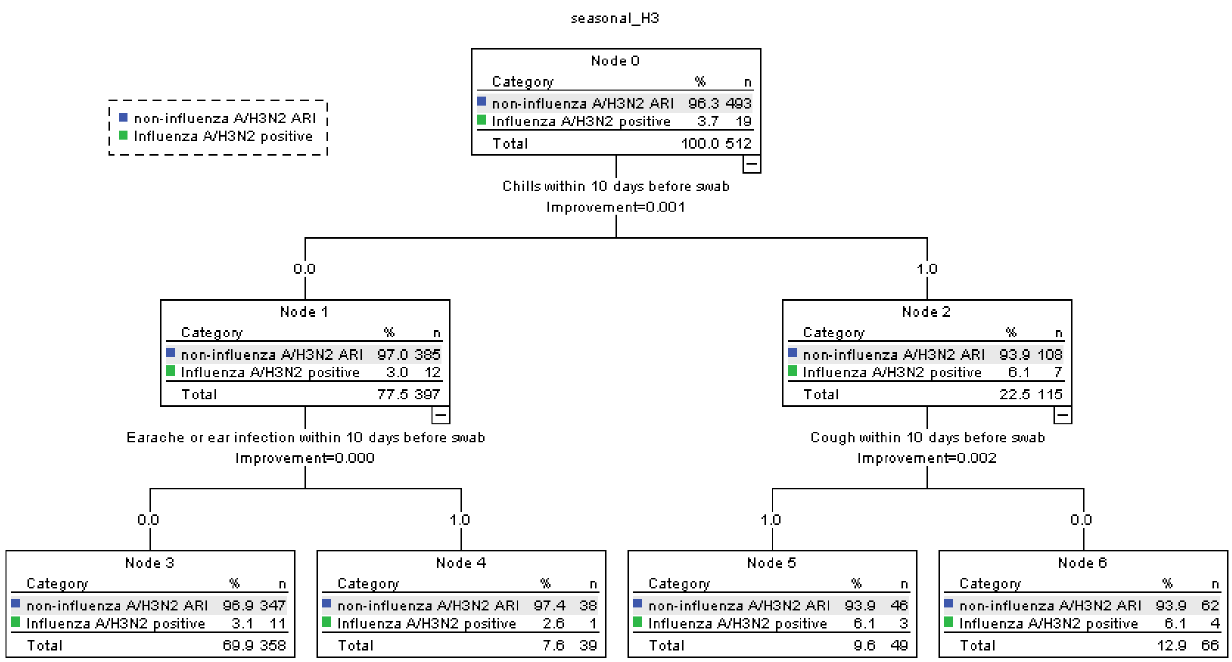
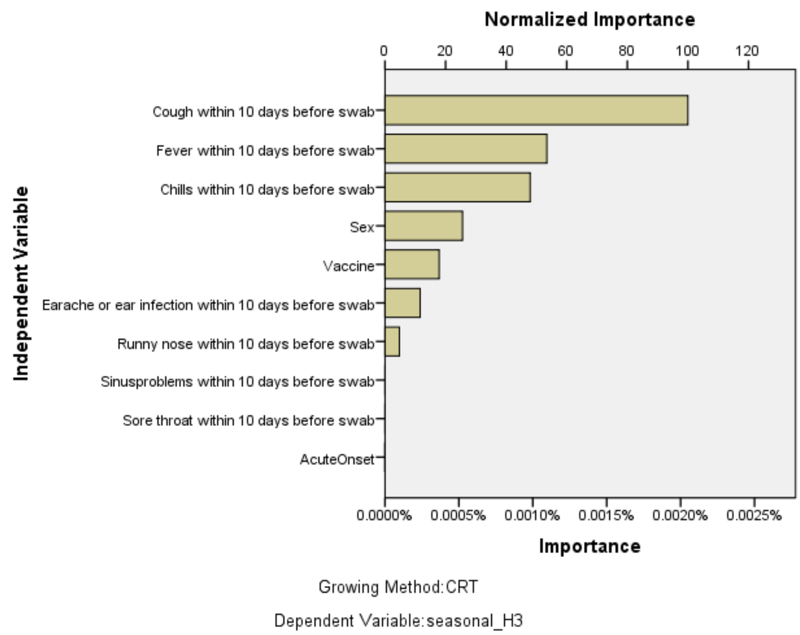


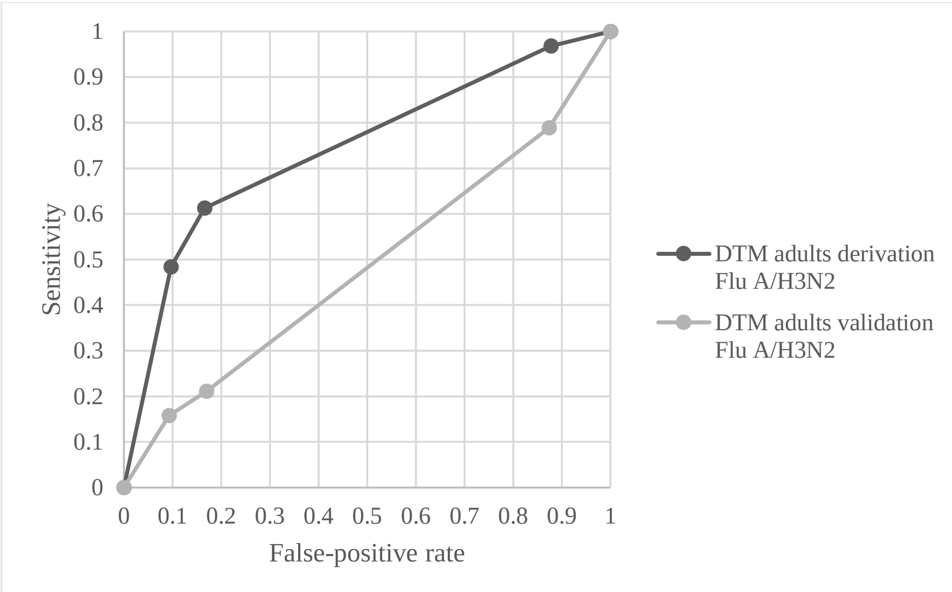
Fig. 3.8 Normalized variable importance plot.
The normalized importance of the variable X is the score when weighted by the variable that has the largest measure of importance. The variable having the largest measure of importance always has a score of 100.



Evaluating model performance by comparison of the ROC curves

The area under the ROC curves of the predicted probabilities was 0.74 (95% CI 0.64-0.84) in the derivation set but was only 0.49 (95% CI 0.34-0.64) in the validation set (Fig. 3.9). The difference between the areas of 0.26 was statistically significant (p= 0.002). Optimal sensitivities and specificities in the derivation set were 61% and 83%, respectively and 21% and 83%, respectively, in the validation set.

Figure 3.9 Comparison of ROC curves of unpruned DTM for the prediction of influenza A/H3N2 in adults

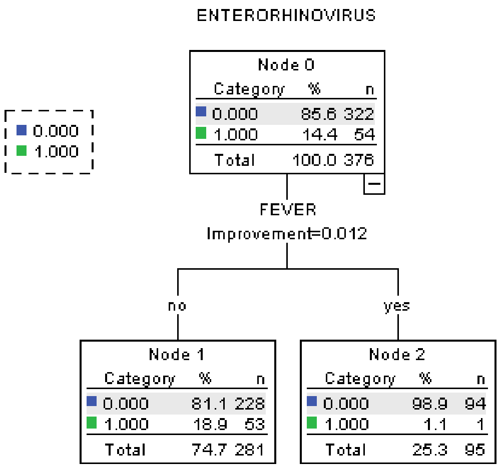


3.4.3 Decision tree for the prediction of entero-/rhinovirus (ERV) infection in children

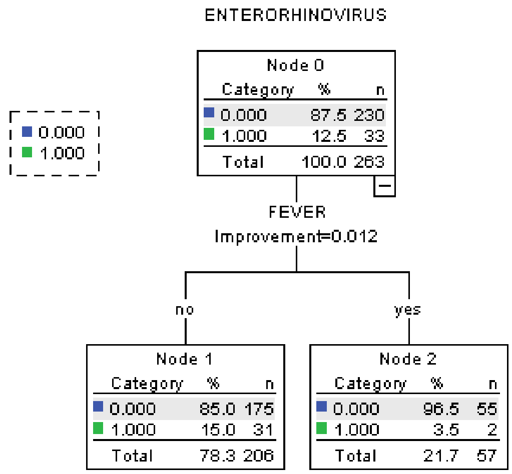
Of all parameters entered into the model only fever remained in the decision tree. Overall percentage predicted correctly was 85.6%. The DTM demonstrates that absence of fever increases the probability of an ERV infection (Fig. 3.10 a.-b). It also shows in the derivation set that presence of fever has a high negative predictive value of 98.9%. Comparison of the derivation and validation model, however, showed slightly worse performance of the validation model with 3.9% less individuals correctly classified as ERV positive in the absence of fever. Because of the few data points no ROC curves were constructed.

Fig. 3.10 (a.-b.) DTM for the prediction of ERV infection in children (a. derivation set; b. validation set).

a. Derivation



b. Validation



References:

1. Cook, E.F. and L. Goldman, *Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis.* J Chronic Dis, 1984. **37**(9-10): p. 721-31.
2. Breiman, L., *Classification and regression trees*. The Wadsworth statistics/probability series. 1984, Belmont, Calif.: Wadsworth International Group. x, 358 p.
3. Chambers, J.M. and T. Hastie, *Statistical models in S*. Wadsworth & Brooks/Cole computer science series. 1992, Pacific Grove, Calif.: Wadsworth & Brooks/Cole Advanced Books & Software. xv, 608 p.
4. Merkle, E.C. and V.A. Shaffer, *Binary recursive partitioning: background, methods, and application to psychology.* Br J Math Stat Psychol, 2011. **64**(Pt 1): p. 161-81.
5. Strobl, C., J. Malley, and G. Tutz, *An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests.* Psychol Methods, 2009. **14**(4): p. 323-48.
6. Hanley, J.A. and B.J. McNeil, *The meaning and use of the area under a receiver operating characteristic (ROC) curve.* Radiology, 1982. **143**(1): p. 29-36.

CHAPTER IV

4.0 Cox proportional hazard model and extended Cox model

4.1 Characteristics of the Cox proportional hazard model

We applied a survival model using the extended Cox proportional hazard (Cox PH) method described by Anderson and Gill to compare the time to first ARI in every season using robust standard errors to model the primary event of interest laboratory confirmed influenza A or B [1, 2]. The key decision in any survival analysis is to define the starting point for determining an individual's "true" survival time which ideally is close to the start of the exposure [3]. The minimum onset of symptom (MinOnset_Swab) was chosen as starting point in this study. Survival models are commonly applied for longer periods of follow-up, usually over several years, where censored data and different lengths of follow-up play an important role and cannot be ignored. The starting point is usually the beginning of the study or the start of an intervention. Our model, therefore, has a very short time-to-event. However, since the aim of the thesis was to compare different statistical models, we deemed a Cox PH model important to consider. Our main interest was though, to examine specific symptoms as potential predictors, which usually occur only a few days before the diagnosis of influenza is made and, therefore, we chose the minimum onset of symptoms as starting point rather than a more distant time point such as the beginning of the trial or the administration of influenza vaccination. Also, an individual level time point was deemed to be easier to understand. However, the question asked, is different from the other models. Important to note is, that the study nurse visited the communities for obtaining the respiratory specimens in nearly fixed intervals. Therefore, those episodes with laboratory confirmed influenza at one visit were considered interval-censored, whereas those episodes that remained influenza negative were considered right-censored.

Several statistical approaches exist to deal with interval-censored data [4]. Due to the low proportion of interval-censored events, we decided to simply consider them as right-censored and proceed with standard time-to-event analysis. This may, however, lead to biased estimates and even misleading results [4].

An individual could also have several different follow-up times provided that this subject had more than one ARI episode over the three seasons. Therefore, although the repeated events in an individual did not occur within the same follow-up period we considered these repeated events within the same subject to be possibly clustered with similar variance of the coefficients which could bias the standard errors of the coefficients downwards. This was accounted for by introducing a robust estimator which creates a population average estimate similar to the GEE model.

In addition, those with symptoms that did not qualify for ARI (e.g. only one symptom) were excluded from the analysis because these subjects were not allowed to provide a respiratory specimen for laboratory analysis as per protocol and hence, non-informative censoring, which is one of the major assumptions of the Cox PH model, would have been violated.

4.2 Model development

The same variables that were entered into the multivariable logistic regression model were entered into the Cox proportional hazard model using a forward stepwise selection approach. A coefficient of a categorical predictor measures the hazard of influenza in the presence of the factor when everything else is held constant. Model fit was examined by -2 log likelihood (-2LL) with smaller values showing better fit. The assumption that needs to be tested in every Cox PH model is that the hazard ratio for comparing any two particular choices of the predictor variables is constant over time. Major violations of the PH assumption (e.g. crossing of the curves) were tested by looking at the log-log survival curves for every predictor.

The analyses were performed on the derivation set and then applied to the validation set and the constructed ROC curves are compared and the optimal sensitivity and specificity determined.

4.3 Results

4.3.1 Hazard of influenza virus A/B infection

Prediction of the hazard of influenza virus A/B infection in children:

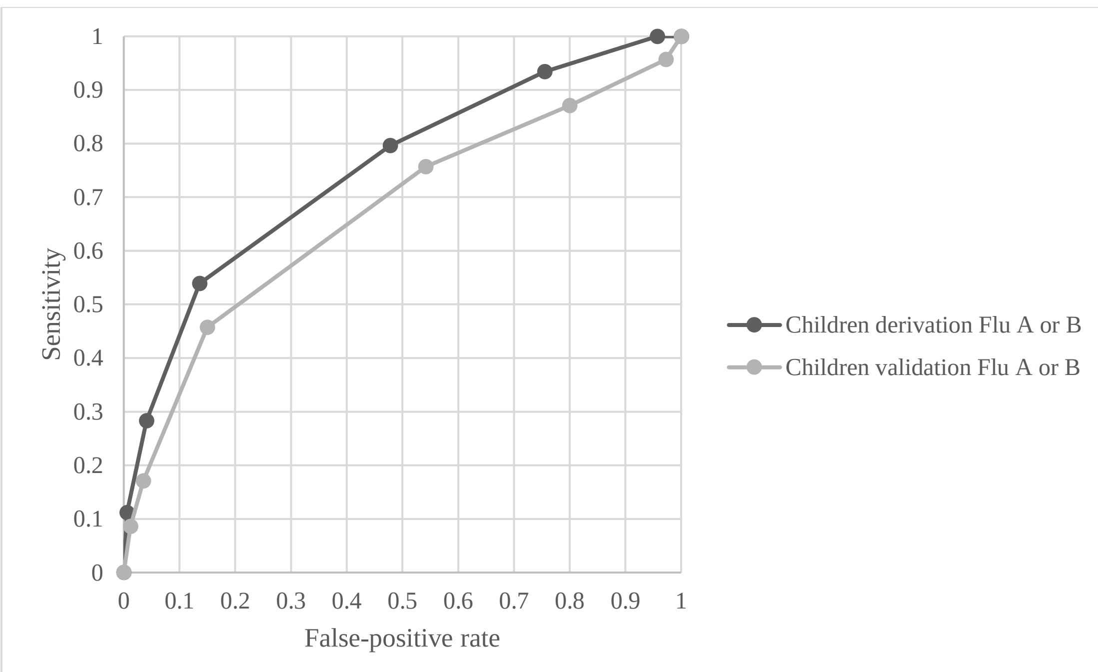
Variables that significantly influenced the hazard of influenza virus A/B infection in the standard Cox proportional hazard model were chills, cough, fever, and age category (over 5 years). Visual inspection did not show a major violation (e.g. crossing of the curves) of the PH assumption for any of the predictors in the final model. These predictors remained significant also after introducing the cluster variable (Tab. 4.1).

A score constructed from weighted coefficients significant in the extended Cox model assigned 1 point to each of the predictors chills and age group over 5 years, and 2 points to each of fever and cough. The corresponding ROC curves of the derivation and validation set had an AUC of 0.75 (95% CI 0.71-0.80) and 0.67 (95% CI 0.60-0.75), respectively. The difference between the two AUCs of 0.08 was, however, not statistically significant (p=0.06) (Fig. 4.1). In a range of total scores between 0 and 6 points optimal sensitivity (54%) and specificity (86%) were found at a cut-off value of 4 points or higher for the prediction of being influenza positive in the derivation set which corresponded to a sensitivity and specificity of 46% and 85%, respectively in the validation set.

Table 4.1 Variables influencing the hazard of influenza virus A/B infection in children, derivation set.

	coeff	HR	se(coef)	robust se	z	p-value	Influenza A/B score children
Chills	0.77	2.2	0.176	0.177	4.36	<.0001	**1**
Cough	0.939	2.6	0.21	0.211	4.46	<.0001	**2**
Fever	1.322	3.8	0.175	0.176	7.49	<.0001	**2**
Age category > 5 yrs	0.548	1.7	0.182	0.175	3.12	0.002	**1**

Figure 4.1 Comparison of ROC curves between derivation and validation data set (Model: extended Cox PH for the prediction of the hazard of influenza A or B) in children



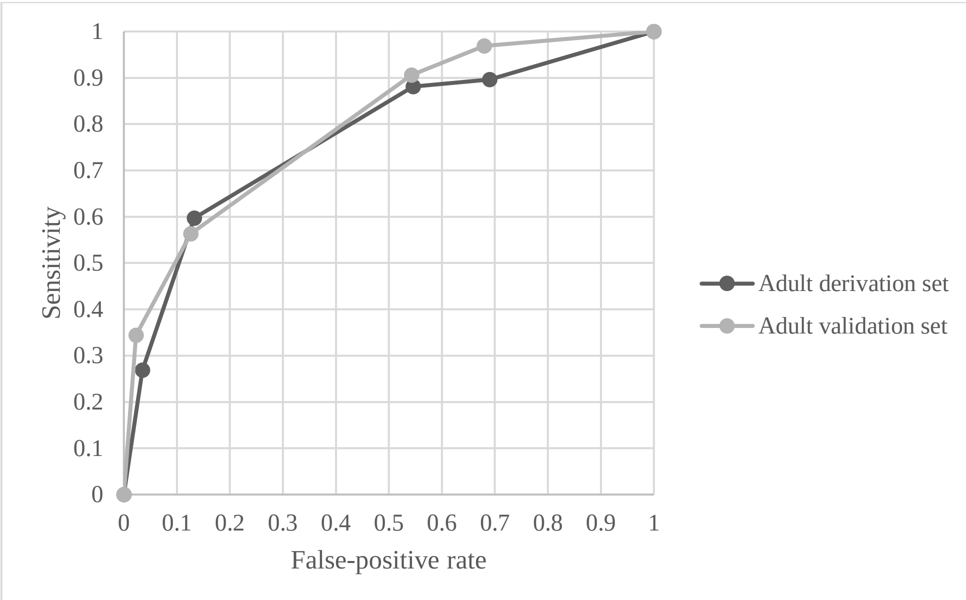
Prediction of the hazard of influenza virus A/B infection in adults:

Of the 8 significant and 1 pre-defined variable (sore throat) three variables influenced the hazard of influenza virus A/B infection in the standard and extended Cox PH model, including chills, cough, and myalgia. The -2LL was significantly reduced suggesting good model fit (Tab. 4.2). A score was constructed from weighted coefficients that remained significant in the extended Cox model that assigned 1 point to each, chills, and myalgia, and 2 points for cough. The corresponding ROC curves of the derivation and validation set had an AUC of 0.77 (95% CI 0.71-0.84) and 0.80 (95% CI 0.72-0.88), respectively with a statistically non-significant difference (p=0.66) between the two AUCs (Fig. 4.2). In a range of total scores between 0 and 4 points optimal sensitivity (60%) and specificity (87%) were found at a cut-point of 3 or higher for the prediction of being influenza A or B positive in the derivation set with a corresponding sensitivity and specificity of 56% and 87%, respectively, in the validation set.

Table 4.2 Variables influencing the hazard of influenza virus A/B infection in adults, derivation set.

	coeff	**HR**	**se(coef)**	**robust se**	**z**	**p-value**	**Influenza A/B score adults**
Chills	1.300	3.7	0.262	0.267	4.86	<.0001	1
Cough	1.436	4.2	0.305	0.298	4.82	<.0001	2
Myalgia	0.905	2.5	0.262	0.263	3.44	<0.001	1

Figure 4.2 Comparison of ROC curves between derivation and validation data set (Model: extended Cox PH for the prediction of the hazard of influenza A or B) in adults.



4.3.2 Prediction of influenza A/H3N2 subtype

Predictors of the hazard of influenza virus A/H3N2 infection in children:

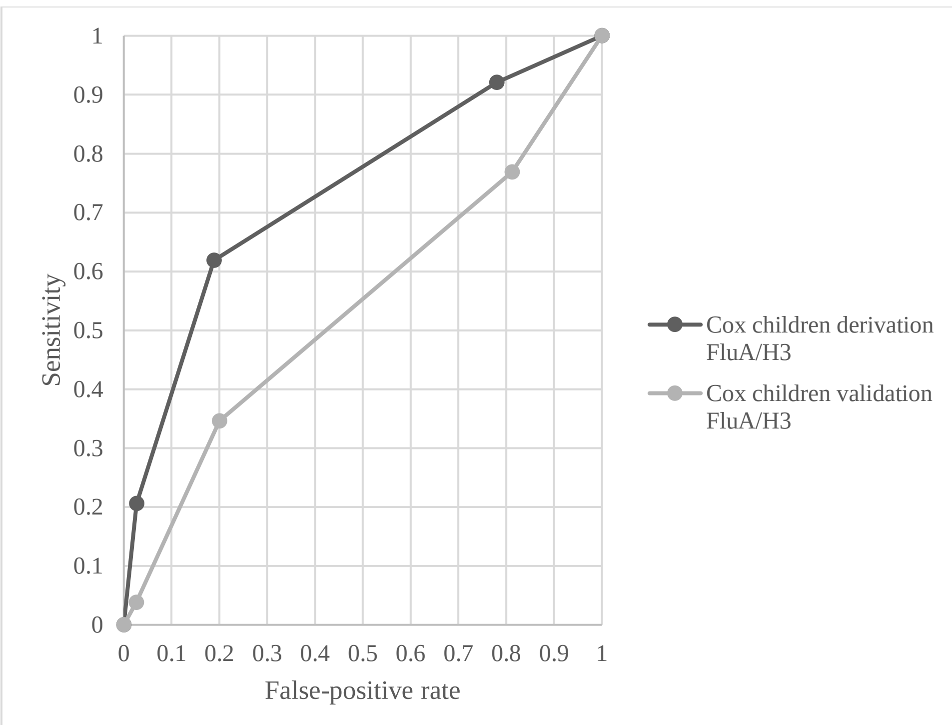
Significant predictors of the hazard influenza A/H3N2 infection in the standard and extended Cox proportional hazard model were Chills, cough, and fever. According to the -2LL the model had a good fit (Tab. 4.3).

When assigning 1 point for each, chills, cough, and fever the constructed ROC curve in the derivation data had an AUC of 0.74 (95% CI 0.67-0.81) and the AUC of the validation set was 0.54 (95% CI 0.41-0.67), respectively; a statistically significant difference of 0.20 (p=0.005) (Fig. 4.3). An individual's total score ranged from 0-3. We found that optimal sensitivity (62%) and specificity (81%) were at a cut-off value of 2 points (or higher) for the prediction of an influenza A/H3N2 hazard in the derivation set. However, in the validation set, sensitivity and specificity were decreased with 35% and 80%, respectively, at the same cut-point.

Table 4.3 Variables influencing the hazard of influenza virus A/H3N2 infection in children, extended Cox proportional hazard model, derivation data set

	coeff	**HR**	**se(coef)**	**robust se**	**z**	**p-value**	**Influenza A/H3N2 score children**
Chills	1.011	2.75	0.266	0.284	3.56	0.0004	**1**
Cough	0.873	2.394	0.325	0.336	2.6	0.009	**1**
Fever	1.2	3.32	0.261	0.271	4.43	<.0001	**1**

Figure 4.3 Comparison of ROC curves between derivation and validation data set (Model: extended Cox PH flu A/H3N2) in children.



Prediction of the hazard of seasonal influenza virus A/H3N2 infection in adults:

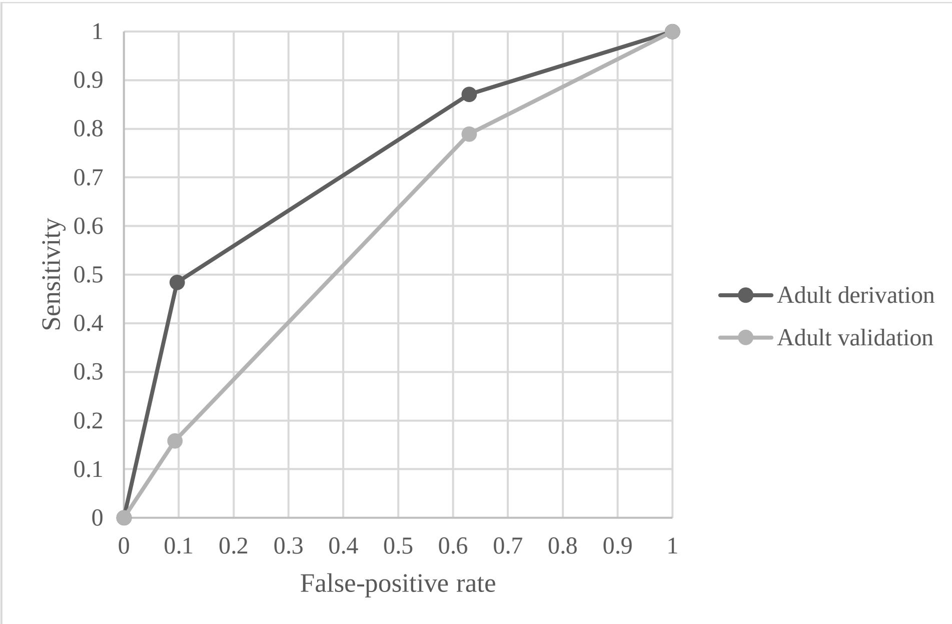
Of the 6 significant and 1 pre-defined variable (gender) only chills and cough influenced the hazard of influenza virus A/H3N2 infection in the standard and extended Cox proportional hazard model (Tab. 4.4). The -2LL was significantly reduced suggesting good model fit.

A score was constructed from significant coefficients with 1 point for each predictor, chills and cough. The corresponding ROC curves of the derivation and validation set had an AUC of 0.73 (95% CI 0.63-0.83) and 0.59 (95% CI 0.47-0.72), respectively (Fig. 4.4). The difference between the AUCs of 0.14 was not statistically significant (p=0.113). In a range of total scores between 0-2 points optimal sensitivity (48%) and specificity (90%) were found at a cut-point of 2 for the prediction of being influenza A/H3N2 positive in the derivation set. However, in the validation set the corresponding sensitivity and specificity at a cut-point of 2 were 16% and 91%, respectively.

Table 4.4 Predictors of influenza virus A/H3N2 infection in adults, extended Cox proportional hazard model, derivation data set.

	coeff	HR	se(coef)	robust se	z	p-value	Influenza A/H3N2 score adults
Chills	1.400	4.1	0.360	0.363	3.85	<0.001	1
Cough	1.631	5.1	0.492	0.497	3.28	0.001	1

Figure 4.4 Comparison of ROC curves between derivation and validation data set (Model: extended Cox PH flu A/H3N2) in adults.



References:

1. Andersen, P.K.G., R. D. , *Cox's Regression Model for Counting Processes: A Large Sample Study.* The Annals of Statistics, 1982. **10**(4): p. 1100-1120.
2. Ingel, K. and A. Jahn-Eimermacher, *Sample-size calculation and reestimation for a semiparametric analysis of recurrent event data taking robust standard errors into account.* Biom J, 2014. **56**(4): p. 631-48.
3. Kleinbaum, D.G. and M. Klein, *Survival analysis : a self-learning text*. 3rd ed. Statistics for biology and health,. 2012, New York: Springer. xv, 700 p.
4. Lindsey, J.C. and L.M. Ryan, *Tutorial in biostatistics methods for interval-censored data.* Stat Med, 1998. **17**(2): p. 219-38.

CHAPTER V

5.0 Application of the “Flu score 3” in the Hutterite data set

5.1 Generalizability of a prediction rule

Every prediction rule that has been internally validated should eventually be evaluated whether it provides accurate prediction in a new sample of patients. Internal validation may have been successful because the test patients were from the same population as the patients in the training set and the two populations were therefore likely to be more homogeneous compared to a new set of patients. The problem that arises is that important predictors may have been missed in the training and the test sample leading to underfitting of the model [1]. It is known that prognostic models often perform worse when applied to a different population. However, this may be evitable with sufficient attention to consistency in methods e.g outcome definitions or surveillance criteria [2].

We were interested in how accurately the Flu score 3 by Ebell and colleagues [3] developed in a population of 459 adult outpatients with influenza-like illness or ARI would predict influenza virus infection in adults of Hutterite communities presenting with ARI compared to our new prediction rule derived in the logistic regression model.

5.2 Characteristics of the Flu score 3

The study population of Ebell and colleagues was an assembly of two prospective cohorts, one from Switzerland, the other from the US. Data collection took place from 1999-2000. The Swiss cohort consisted of adult patients visiting a University primary care clinic, the US population were adult emergency department or urgent care ambulatory patients of a large tertiary care university hospital. Inclusion criteria in the Swiss cohort was influenza-like illness as determined by the primary care physician. US patients were consecutively included when presenting with signs and symptoms of ARI (cough, sinus pain, congestion/rhinorrhea, sore throat or fever). The outcome was influenza A or B determined by culture or PCR from a respiratory specimen. Symptoms were evaluated in both cohorts using a standardized questionnaire. Influenza prevalence in the Swiss and US cohort was 53% and 21%, respectively.

The prediction rule was derived from a random sample of 326 patients (70% of total study population) and was tested in the remaining 133 patients. They applied multiple logistic regression with backward elimination entering all independent parameters in the model that were statistically significant in univariable analysis. The simplified point score was constructed from the derived odds ratios (doubled value of the odds ratio to avoid half points) of the individual parameters. It comprised the following variables (point score): Acute onset within 48 hours (1 point), myalgia (2 points), chills/sweats (1 point), and fever plus cough (2 points). The risk of influenza was 8% in the lowest risk-group (0 to 2 points) for a LR 0.17. In the high-risk category (4 to 6 points) the risk of influenza was 59% with a LR of 2.7.

5.3 Assessing the performance of the Flu score 3

A formal validation where the ROC curves derived from the point score would be compared between the Ebell data set and the new data set to assess its generalizability was not possible as we did not have access to the original dataset. However, the results from the original study allowed us to compare the performance by comparing the sensitivities, specificities, positive and negative LR for the low-risk (0-2) and high-risk (4-6) categories derived from the application to the total original population to when applied to the Hutterite data set.

Due to the apparent resemblance of the populations and the same outcome definition (laboratory confirmed influenza virus infection) in the Ebell study and the present study we further hypothesized that the Flu score 3 would perform similarly to the influenza score derived from the logistic regression model in the current data set.

The adult influenza A/B score of the logistic regression model also involves chills (2 points), cough (2 points), and myalgia (1 points); however, fever is not part of the score.

5.4 Results

The point score was applied to the adult Hutterite data set in season 1 and the total scores were compared to the true category of presence or absence of influenza virus infection. Season 1 in the adult Hutterite data set consisted of 568 first ARI episodes. Fifty-one subjects (9%) were influenza A or B positive.

For better illustration 2x2 tables were reconstructed with the original data in table 5.1a. (data in **bold** were provided in the published manuscript) and compared to the 2x2 table derived after applying the Flu score 3 to the new data set (Tab. 5.1b). The first comparison (Table 5.1 a-b) depicts the performance if all subjects with a Flu score 3 of 0-2 would be categorized as influenza negative (“low-risk). The second comparison (Table 5.1 c-d.) shows the performance of the Flu score 3 if all individuals with a score between 4 and 6 would be categorized as influenza positive.

Table 5.1 (a-d). Comparison of data in 2x2 tables (a. low risk, original data set; b. low risk, new data set; c. high risk, original data set; d. high risk, new data set) with corresponding absolute numbers (n) and column percentages (%), respectively.

a. 2x2 table original data set (Ebell et al.)

	Flu score 0-2	Flu score 3-6	total
	n (%)	**n (%)**	**n (%)**
Influenza negative	137 (92)	165 (53)	302 (66)
Influenza positive	**12** (8)	145 (47)	**157** (34)
total	**149** (100)	310 (100)	**459** (100)

b. 2x2 table adult Hutterite data set (1st season)

	Flu score 0-2	Flu score 3-6	total
	n (%)	**n (%)**	**n (%)**
Influenza negative	439 (94)	78 (76)	517 (91)
Influenza positive	27 (6)	24 (24)	51 (9)
total	466 (100)	102 (100)	568 (100)

c. 2x2 table original data set (Ebell et al.)

	Flu score 0-3	**Flu score 4-6**	total
	n (%)	**n (%)**	**n (%)**
Influenza negative	227 (82)	75 (41)	302 (66)
Influenza positive	51 (18)	**106** (59)	**157 (**34)
total	278 (100)	**181** (100)	**459** (100)

d. 2x2 table adult Hutterite data set (1st season)

	Flu score 0-3	**Flu score 4-6**	total
	n (%)	**n (%)**	**n (%)**
Influenza negative	496 (93)	21 (60)	517 (91)
Influenza positive	37 (7)	14 (40)	51 (9)
total	533 (100)	35 (100)	568 (100)

Comparison of diagnostic indexes

Comparison of the diagnostic indexes showed that the application of the Flu score 3 in the adult Hutterite data set yields different results (Tab. 5.2). In the new data set, the proportion of individuals with a score between 0 and 2 that were misclassified (falsely predicted as being negative) in the low risk category was lower than in the original data set (0.06 vs. 0.08). Also, the true-negative rate was higher in the new data set than in the original data set (85% vs. 45%). On the other hand, the false-negative rate (1-sensitivity) was higher than in the original data set with 53% being missed compared to only 8% being missed.

The proportion of individuals with influenza that were correctly predicted by the score was 59% in the original data set but only 40% in the new data set. Overall performance of the flu score 3 in the new data set was better regarding specificity but this came at the cost of a lower sensitivity. Because of the lower prevalence of influenza A or B and therefore a lower pre-test odds in the new data set the post-test probability remained lower despite the higher LR of 6.8 compared to 2.7.

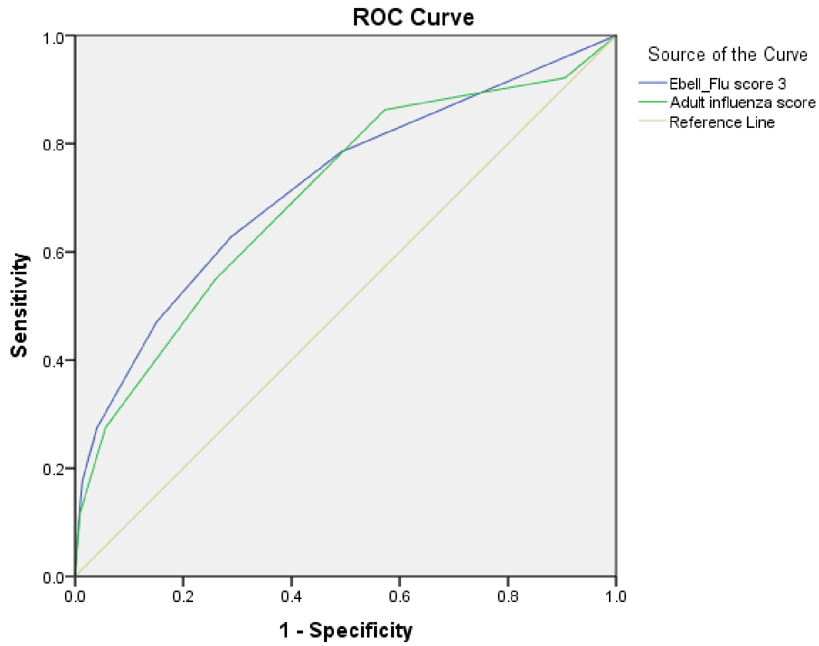
Table 5.2 Comparison of the diagnostic indexes

Original data set (Ebell et al.)						Adult Hutterite data set (1st season)					
Flu score 3	**Sens.**	**Spec.**	**1-NPV**	**PPV**	**LR**	**Flu score 3**	**Sens.**	**Spec.**	**1-NPV**	**PPV**	**LR**
Low risk (0-2 points)	0.92	0.45	0.08	n.a.	0.17	Low risk (0-2 points)	0.47	0.85	0.06	n.a.	0.62
High risk (4-6 points)	0.68	0.75	n.a.	0.59	2.7	High risk (4-6 points)	0.27	0.96	n.a.	0.40	6.8

Comparison of ROC curves

The area under the ROC curves constructed from the Flu score 3 was then compared to the area under the ROC curve generated from the logistic regression model in the adult data set and applied in the season 1 adult data set. Since the ROC curves are derived from the same underlying population the differences in the AUC were simply visually inspected (Figure 5.1). The AUC of the Flu score was 0.72 (95% CI 0.64-0.80) and the AUC of the adult influenza score slightly smaller with 0.70 (95% CI 0.62 -0.78). From eyeballing the two ROC curves and comparing the magnitudes of the AUCs the performance of these two scores seems almost equivalent in the Hutterite population of the influenza season 1.

Figure 5.1 ROC curve comparison of the Ebell Flu score and the own derived adult influenza A/B score



5.5 Conclusions

We conclude that the performance of a score derived in a different population and applied to a new data set with a lower prevalence of the disease and probably different case-mix is lower with respect to sensitivity and PPV which therefore needs to be considered when applying a prediction rule in a new population. The results also showed that the performance of the Flu score 3 was comparable to the performance of a score derived from the data set. Tests with an AUC of < 0.80 are, however, regarded as having moderate diagnostic or predictive power.

References:

1. Justice, A.C., K.E. Covinsky, and J.A. Berlin, *Assessing the generalizability of prognostic information.* Ann Intern Med, 1999. **130**(6): p. 515-24.
2. Charlson, M.E., et al., *Why predictive indexes perform less well in validation studies. Is it magic or methods?* Arch Intern Med, 1987. **147**(12): p. 2155-61.
3. Ebell, M.H., et al., *Development and validation of a clinical decision rule for the diagnosis of influenza.* J Am Board Fam Med, 2012. **25**(1): p. 55-62.

CHAPTER VI

6.0 Comparison of models:

6.1 Summary of models for the prediction of Influenza A/B virus infection

The predictors of influenza A/B virus infection derived from each model, the scores and the corresponding area under the ROC curves of the scores of the logistic regression/GEE, the recursive partitioning, and the extended Cox PH model are presented in tables 6.1 and 6.3 for children and adults, respectively.

Both tables depict the similar magnitudes of AUCs of the corresponding models which is also illustrated in figures 6.1 and 6.3 where the ROC curves of the various models are displayed separately for the derivation and validation set.

In order to provide an idea about the potential usefulness of the score for clinical decision making regarding the probability of influenza A/B virus infection depending on the magnitude of the score, a low and a high cut-off value was arbitrarily chosen, if possible, and displayed along with the sensitivities and specificities, PPV and NPV, positive and negative likelihood ratio (Tabs 6.2 and 6.4). This should give an idea of whether it would be save to rule out a diagnosis of influenza (probability below test threshold) or make the diagnosis likely enough to empirically treat (if needed) without further laboratory testing (probability at or above treatment threshold). Two examples illustrate the application of the score: A 10- years old child with cough, fever and myalgia would have a score of 5 in the logistic regression/GEE model. With a pre-test probability of 11% (= prevalence) and a pos. LR of 3.6 (derivation model), the post-test probability of influenza A/B virus infection would be approximately 31%. Similarly, an adult with chills and cough would have a score of 3 in the Cox PH model. The pre-test probability is assumed to be

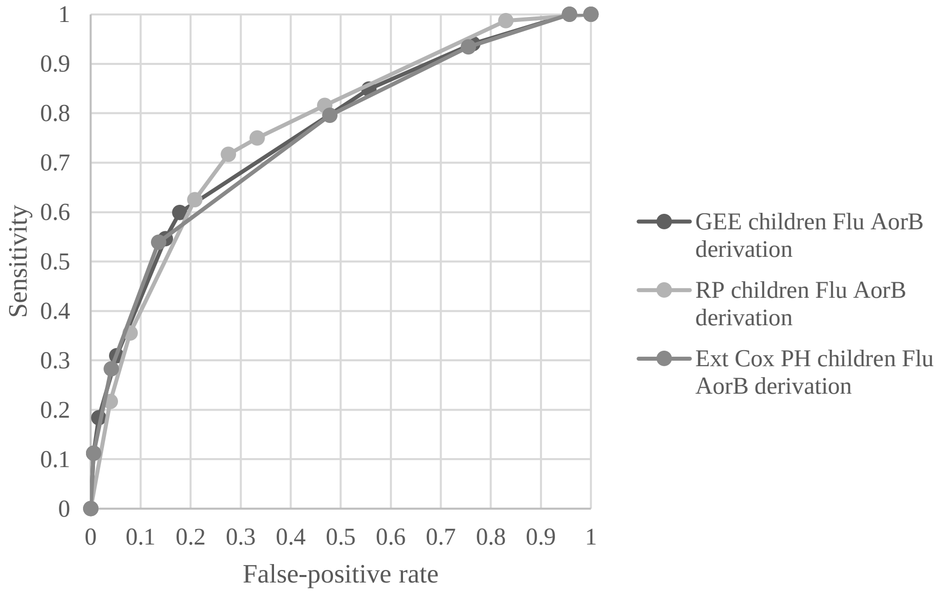
8%, the pos. LR in the derivation model is 4.6. It follows that the post-test probability of influenza A/B virus infection of this individual would be approximately 29%.

Table 6.1 Summary of models and predictors for the children data set, outcome: influenza A/B virus infection

Model	Predictors in the model	Score	Set	AUC (95% CI)	p-value	Cut-point	Sens	Spec
Logistic regression and GEE	**Age over 5 years** **Chills** **Cough** **Fever**	1 2 2 2	**Deriv**	0.76 (0.72-0.80)	0.166	4	60%	82%
			Valid	0.70 (0.63-0.77)		4	56%	81%
Recursive partitioning	**Fever** **Chills** **Cough** **Runny nose** **Sex**	n.a.	**Deriv**	0.77 (0.73-0.81)	0.450	n.a.	72%	72%
			Valid	0.74 (0.67-0.80)		n.a.	71%	70%
Extended Cox PH model	**Age over 5 years** **Chills** **Cough** **Fever**	1 1 2 2	**Deriv**	0.75 (0.71-0.80)	0.060	4	54%	86%
			Valid	0.67 (0.68-0.80)		4	46%	85%

Figure 6.1 (a-b). Comparison of the different models for the prediction of influenza A/B in children (a. derivation; b. validation set).

a.



b.

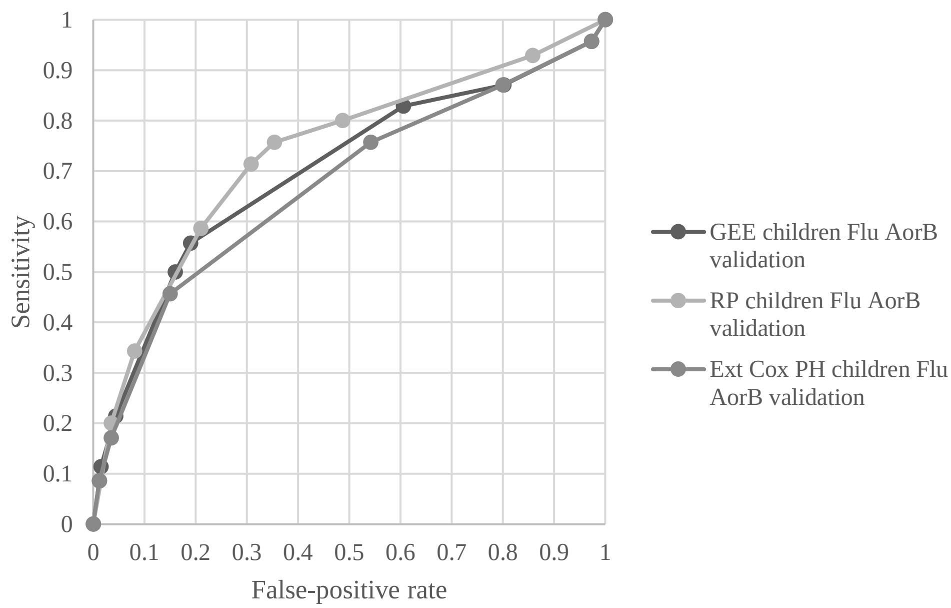


Table 6.2 Comparison of the predictive indexes for two different cut-points of the scores for the prediction of influenza A/B in children

Model	Predictors in the model	Score	Set	Cut-point	Sens	Spec	PPV*	NPV*	pos. LR	neg. LR
Logistic regression and GEE	Age over 5 years Chills Cough Fever	1 2 2 2	Deriv	3	85%	44%	16%	96%	1.5	0.34
				5	55%	85%	31%	94%	3.6	0.53
			Valid	3	83%	40%	15%	95%	1.4	0.42
				5	50%	84%	28%	93%	3.1	0.59
Extended Cox PH model	Age over 5 years	1	Deriv	3	78%	52%	17%	95%	1.6	0.42
	Chills	1		5	28%	96%	46%	92%	7.0	0.75
	Cough	2	Valid	3	76%	46%	15%	94%	1.4	0.52
	Fever	2		5	17%	97%	41%	90%	5.7	0.86

*assuming a prevalence (~pre-test probability) of influenza virus A or B infection of 11%.

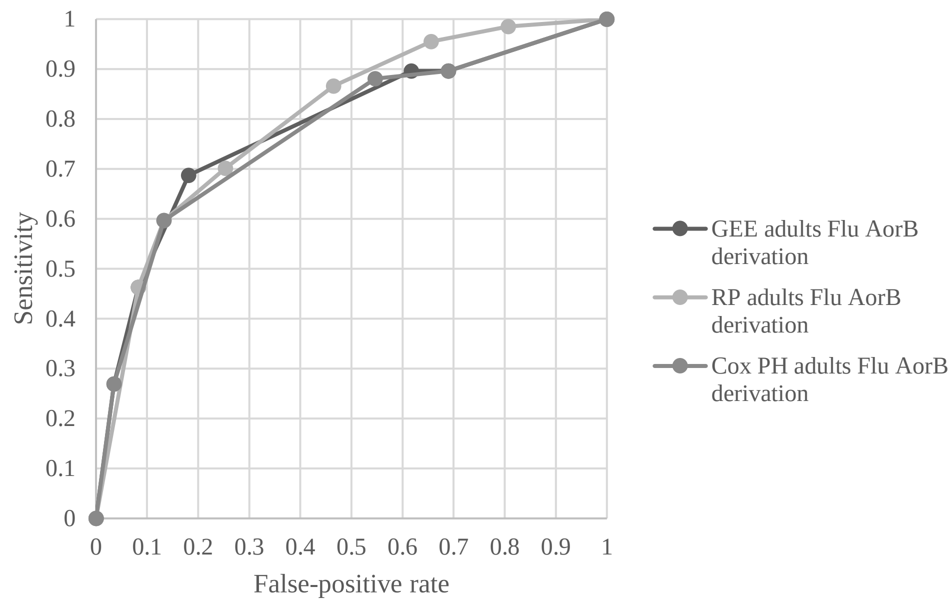
Table 6.3 Summary of models and predictors for the adult data set, outcome: influenza A/B virus infection

Model	Predictors in the model	Score	Set	AUC (95% CI)	p-value	Cut-point	Sens	Spec
Logistic regression and GEE	Chills Cough Myalgia	2 2 1	Deriv	0.78 (0.72-0.85)	0.866	3	69%	82%
			Valid	0.79 (0.71-0.87)		3	63%	81%
Recursive partitioning	Chills Cough Myalgia Sinus problems Sore throat	n.a.	Deriv	0.80 (0.75-0.86)	0.410	n.a.	70%	75%
			Valid	0.75 (0.65-0.85)		n.a.	66%	74%
Extended Cox PH model	Chills Cough Myalgia	1 2 1	Deriv	0.77 (0.71-0.84)	0.661	3	60%	87%
			Valid	0.80 (0.72-0.88)		3	56%	87%

n.a.= not applicable

Figure 6.2 (a-b). Comparison of ROC curves among different models for the prediction of influenza A/B in adults (a. derivation set; b. validation set).

a.



b.

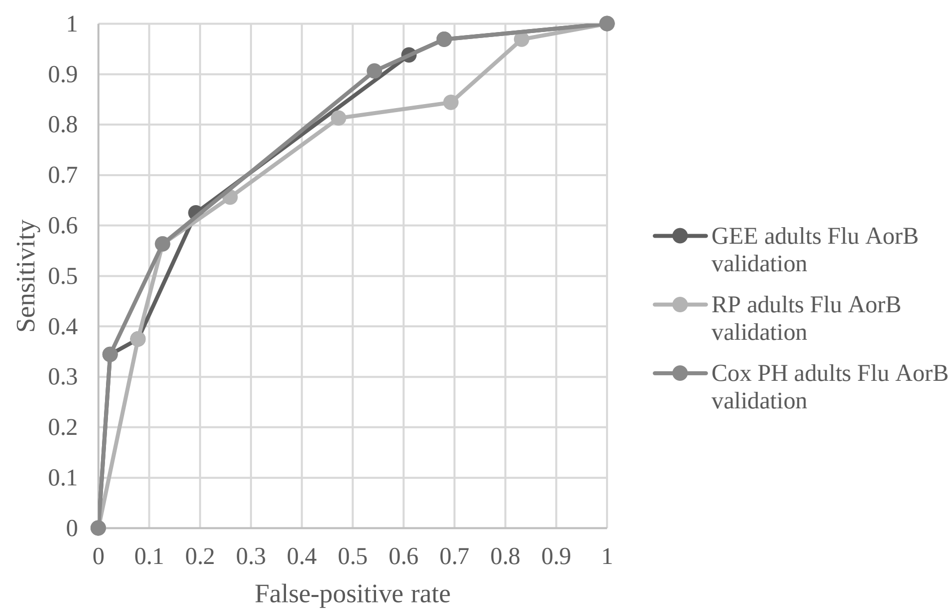


Table 6.4 Comparison of the predictive indexes for two different cut-points of the scores for the prediction of influenza A/B in adults.

Model	Predictors in the model	Score	Set	Cut-point	Sens	Spec	PPV*	NPV*	pos. LR	neg. LR
Logistic regression and GEE	Chills	2	Deriv	2	90%	38%	10%	98%	1.5	0.26
	Cough	2								
	Myalgia	1		4	30%	96%	34%	94%	5.8	0.59
			Valid	2	94%	29%	9%	91%	1.3	0.21
				4	38%	92%	26%	95%	4.8	0.67
Extended Cox PH model	Chills	1	Deriv	2	88%	45%	11%	98%	1.6	0.27
	Cough	2								
	Myalgia	1		3	60%	87%	26%	97%	4.6	0.46
			Valid	2	91%	46%	11%	99%	1.7	0.2
				3	24%	96%	24%	96%	4.3	0.5

*assuming a prevalence of 7%

6.2 Summary of models for the prediction of influenza A/H3N2 subtype

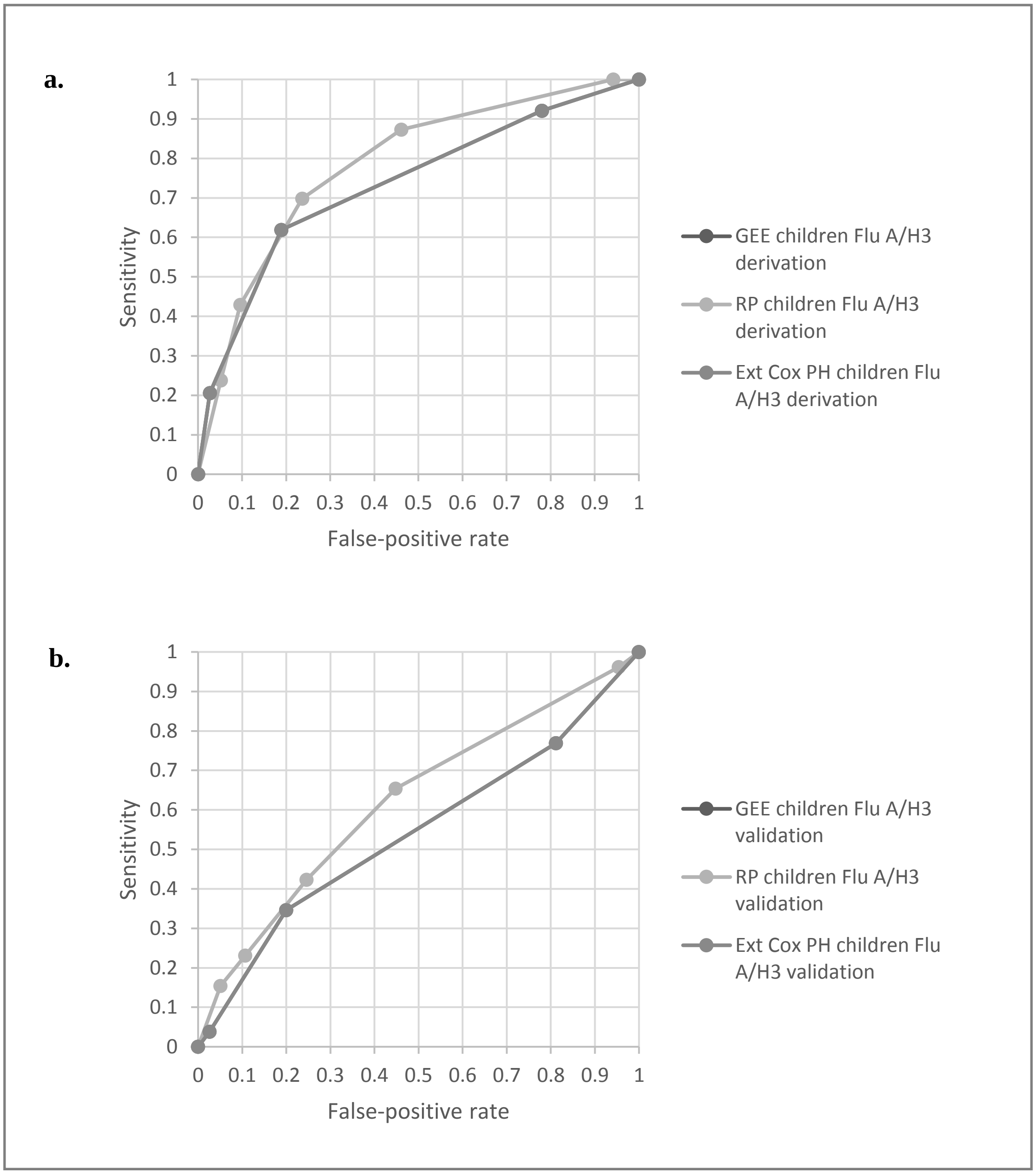
Prediction of influenza A/H3N2 in children

Fever, chills, and cough occurred in all three mathematical models for the prediction of influenza A/H3N2 (Tab. 6.5). None of the predictive models showed clear superiority regarding performance, although the recursive partitioning model with 4 predictors slightly outperforms the other two identical models in both, the derivation and validation set. (Fig. 6.3). It is, however, remarkable that all models performed significantly worse in the validation set, which is likely the result of a lower event rate in the validation set.

Table 6.5 Summary of models and predictors of influenza A/H3N2 in children

Models	Predictors	Score	Set	AUC (95% CI)	p-value	Sens	Spec
Logistic regression and GEE	**Chills** **Cough** **Fever**	1 1 1	**Deriv**	0.74 (0.67-0.81)	**0.006**	62%	81%
			Valid	0.54 (0.41-0.67)		35%	80%
Recursive partitioning	**Fever** **Chills** **Age group** **Cough**	n.a.	**Deriv**	0.79 (0.73-0.84)	**0.019**	70%	76%
			Valid	0.62 (0.51-0.74)		66%	55%
Extended Cox PH model	**Chills** **Cough** **Fever**	1 1 1	**Deriv**	0.74 (0.67-0.81)	**0.005**	62%	81%
			Valid	0.54 (0.41-0.67)		35%	80%

Figure 6.3 (a-b). ROC curve comparisons of the different models for the prediction of influenza A/H3N2 in children (a. derivation set; b. validation set)



Prediction of influenza A/H3N2 in adults.

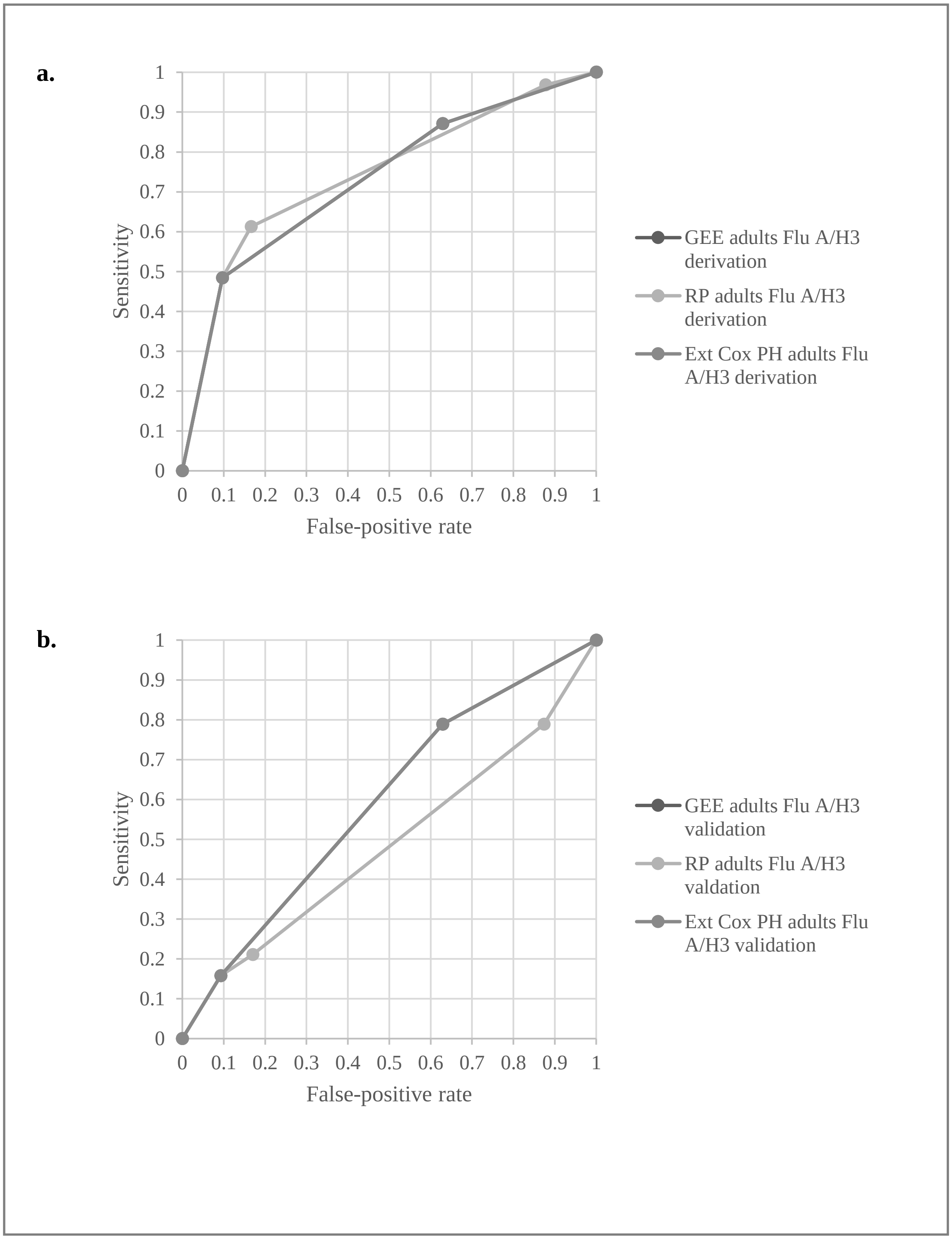
All 3 different models (logistic regression, recursive partitioning, and extended Cox model) were successfully built for the prediction of influenza A/H3N2 in adults (Tab. 6.6). All models included chills and cough as predictive variables whereas fever was eliminated during the model selection process.

The comparison of the ROC curves of the corresponding models is repetitive of what is shown in the children data set. Namely, the performances of the individual mathematical models is almost equivalent in the derivation set but markedly worse when applied to the validation set 6.4 (a-b).

Table 6.6 Summary of models and predictors of influenza A/H3N2 in adults

Model	Predictors of influenza A/H3N2	Score	Set	AUC (95% CI)	p-value	Sens	Spec
Logistic regression and GEE	**Chills** **Cough**	1 1	**Deriv**	0.73 (0.63-0.83)	**0.108**	48%	90%
			Valid	0.59 (0.47-0.72)		16%	90%
Recursive partitioning	**Chills** **Cough** **Ear problems**	n.a.	**Deriv**	0.74 (0.64-0.84)	**0.002**	61%	83%
			Valid	0.49 (0.34-0.64)		21%	83%
Extended Cox PH model	**Chills** **Cough**	1 1	**Deriv**	0.73 (0.63-0.83)	**0.140**	48%	90%
			Valid	0.59 (0.47-0.72)		16%	91%

Figure 6.4 (a-b). ROC curve comparisons of the different models for the prediction of influenza A/H3N2 in adults (a. derivation set; b. validation set)



6.3 Prediction of ERV infections in children

Presence of fever was the most important independent predictor negatively associated with ERV-infections in both, multiple logistic regression and recursive partitioning, whereas presence of chills was only a significant negative predictor in the logistic regression model. Performance of the models when applied to the validation set was slightly worse in both, standard logistic regression and recursive partitioning.

6.4 Discussion

The most predictive clinical and demographic variables that are attributable to influenza A or B virus infection or to an infection caused by the influenza A/H3N2 subtype have been explored by means of logistic regression/GEE, recursive partitioning, and Cox PH models with robust estimators. The results suggest, that in children, fever, chills and cough are the most predictive symptoms, whereas in adults chills and cough seem to be predictive of influenza virus A or B infections in this population of Hutterite community members with a broader spectrum of symptoms compared to most populations in previous studies in which part of the inclusion criteria was fever.

The models for predicting influenza A or B in children demonstrated that having at least 3 clinical symptoms is highly specific (84% to 97%) with PPVs between 28% and 46% but with low sensitivities between 17% and 55% implying that only a rather small proportion of all those who actually have influenza would be caught. Compared to children the two symptoms chills and cough have a similar high specificity of 87% or more with a PPV between 24% and 34% in adults. Fever was not an important predictor in adults as expressed by missing point scores whereas cough had the highest weight in adults with a minimum of 2 points assigned in every score. Despite the higher complexity of the DTM with higher numbers of important predictors the performances with respect to the visual comparison of the ROC curves and AUC was equivalent to the other models. The DTM has nicely illustrated, that interaction effects might be present which were not discovered by standard logistic regression and are, to the best of our knowledge, not described in the literature as such. Their biological plausibility and clinical relevance, however, remains uncertain.

Evaluating predictive models for influenza A/H3N2 revealed almost identical predictors as for the composite outcome influenza A or B. Fever had a positive weight in predicting influenza A/H3N2 in children but not in adults. This result can be explained by the fact that influenza A/H3N2 was the predominant strain in in season 1 and 3 in both, children and adults. Interestingly, fever was not statistically significant in univariable analysis of seasonal influenza and pandemic influenza. However, these results need to be interpreted with caution due to the low number of positive events and the inability to adjust for other predictors in a multivariable or decision tree model.

The one comparable study that evaluated specific predictors of influenza A virus subtypes in children and adults also found that fever was only a significant predictor of influenza A/H3N2 but not for influenza A/H1N1 [1].

The fever-and-cough rule in the study by Boivin et al. [2] had a sensitivity, specificity, PPV and NPV of 77.6%, 55.0%, 86.8%, and 39.3%, respectively. The same rule applied to the season 1 Hutterite children data set had a sensitivity, specificity, PPV and NPV of 36%, 90%, 44%, and 86%, respectively; and in the adult data set the same indexes were 22%, 98%, 45%, and 94% respectively for season 1. The discrepancies in the PPV and NPV between their and our result is explained by the large difference in the prevalence of influenza virus infection which was 72% in the study by Boivin et al. but only approximately 10% in our study which likely reflects a different case-mix in the population. Nevertheless, this rule seems useful to safely exclude patients with neither of these symptoms and would allow to initiate treatment with a neuraminidase inhibitor in those with both symptoms present without further testing. The low prevalence of influenza virus infection resulted in low PPVs even when the sensitivities and specificities were reasonably high. This challenges the clinical utility of any of these prediction rules and requires combination with a standard diagnostic test such as rapid antigen-based test. Although the fever-and-cough rule has been proven useful, these data show that it is still not perfect and individuals without fever, especially adults, would still have to be considered having influenza and would require further testing, in context with surveillance data confirming that influenza virus is circulating.

Although the differences in the AUCs between any models compared to the GEE model were almost negligible, it is still remarkable that the extended Cox PH model actually showed a similar magnitude in the AUC as the GEE model despite the fact that it was deemed the least appropriate model. The reason is probably that the time variable was actually meaningless in these models due to the underlying short course of influenza and other respiratory viral infections. Time was not appropriately linked between symptoms and having influenza or having no influenza, in that case, being censored, because the symptom duration relied on the fix interval when the nurses came and obtained the respiratory specimen. On the other hand, none of the key assumptions, which are censoring is non-informative and the assumption of proportional hazards (= constant relative hazard), were violated which would have invalidated the analyses. Therefore, in a study were the status of all patients is known at a fixed time the binary outcome can be analysed just as in logistic regression. This probably explains the equivalent performance of the model.

Overall, the 3 models were fairly consistent in their predictors and similar weights were assigned to the predictors. Validating the data using data-splitting showed worse performance in the validation group as is usually expected, especially in the case of overfitting.

Limitations:

The Cox PH model has some limitations since it predicted the time between the onset of symptoms and the different outcomes which is a different question compared to the other models. A more common approach would have been to treat symptoms as time-dependent covariates. Next, although technically correct, modeling interval-censored data with a standard Cox PH model may lead to biased estimates and other approaches e.g. an accelerated failure time model, should be considered to avoid misleading results [3]. However, dealing with interval-censored data is still a topic under investigation and most statistical software packages offer only limited functions [4].

Although, nested data has been taken into account in the derivation process using GEE and robust estimators, this was not true for the DTM. Furthermore, although the derivation and validation set were independent the validation set contained potentially nested data and testing the derived predictors by constructing ROC curves in populations in which the outcomes are potentially non-independent could lead to false estimation of the model performance resulting in either over- or underestimation.

It is also known, that simple data-splitting is a less stringent validation method than e.g. splitting the groups with respect to time [5]. But both of them are actually inefficient because only a proportion of the data set is being used. Bootstrapping has the advantage that it produces nearly unbiased estimates of predictive accuracy and the entire data set can be used. Bootstrapping with correlated data, however, has the disadvantage that it becomes computationally demanding and inefficient.

Any prediction rule needs to be tested and validated in an external group of patients to ensure that it maintains its predictive power. The results of this study, therefore, may only hold true in this population and are not generalizable. Application of a new clinical prediction rule, even if externally validated, requires a population similar to the one the rule has been developed and validated. It should also be assured that the clinical prediction rule has relevant influence on decision-making. Neither of these requirements are fulfilled by this study, with the latter usually requiring an impact analysis ideally by means of a randomized controlled trial.

A prediction rule should ideally be developed prospectively. This analysis was done in retrospect and therefore we had to deal with the information available in the data set. One could argue that the symptoms in children were not comprehensively enough evaluated since gastrointestinal symptoms such as vomiting and diarrhoea which frequently occur in children with influenza were not measured. This potential information bias is, however, with respect to the interpretation of the results, non-differential, and its impact is unknown.

Evaluation of predictors of the various influenza A subtypes and entero-/rhinovirus infections using separate binary outcomes could have been more efficiently explored using multinomial logistic regression with e.g. non-influenza as the reference category. This would have generated an overall model response to find out which predictors were important at all. The outcome categories have to be mutually exclusive. The decision to choose separate binary logistic regression models has been motivated by a clinical rational. In my different models, the aim was to obtain predictors that discriminate between influenza A/B and non-influenza cases, and between any particular influenza A subtype and all other pathogens including other influenza A subtypes, which would rather reflect a real life situation where patients present with ARI and we would like to predict the probability of an infection with a particular influenza A subtype among all the possible causing pathogens. These outcomes, however, are not mutually exclusive. In a multivariate model with 4 outcome categories e.g. (1) non-influenza, (2) seasonal A/H1N1, (3) Influenza A/H3N2, and (4) pandemic influenza A/H1N1 I would have had only one reference category, e.g. non-influenza virus, and I would have compared the predictors from a specific subtype only against the reference category, non-influenza virus, and the result would have been interpreted as the odds of having a particular influenza A subtype compared to the odds of having a non-influenza virus, thereby ignoring the other influenza subtypes which does not reflect reality.

It is also more challenging to generate individual predictive models using polytomous (or multinomial) logistic regression because, even if the coefficient is only significant in one comparison (e.g. fever in H3N2 vs. non-influenza but not in H1N1 vs. non-influenza) the predictor has to be retained in both equations or it has to be completely removed.

Point scores based on beta coefficients underlie the assumption of additive risk which excludes interaction effects. It is therefore important to emphasize that, in the point scores, potential interactions were not adequately taken into account by creating a simple additive score. The constructed ROC curves might therefore not adequately reproduce the performance since the true data generating mechanism is not additive.

Interpretation from regression might seem simpler but one disadvantage of the logistic regression model compared to recursive partitioning is that complex interactions remain undiscovered [6].

A disadvantage of the recursive partitioning model is, however, that if the primary daughter node has been chosen “incorrectly” everything else depends on this decision [6]. The low number of events and high number of predictors in the analyses of influenza A/H3N2 subtype rather produced instable models which likely represent chance findings. Applying ensemble methods such as random forest could overcome such problems. Eventually, choosing split-halves for validation reduced the sample size and therefore the confidence in the estimates.

6.5 Summary

This work aimed to predict influenza virus infection including subtypes and entero-rhinovirus infections applying different mathematical models and approaches for derivation and validation of the predictive scores. None of the models, however, was outperforming but all of them had reasonable performance. From a methodological perspective, given the small numbers of events, it would have been clearly more favorable to perform bootstrapping for derivation and validation and preserve the sample size or number of events, respectively, rather than splitting the data and reduce the number of events further thereby increasing the risk of overfitting. This comes, however, at the cost of computationally more demanding analyses. The derived scores would be simple to apply in clinical practice, but their generalizability and impact on clinical decision-making remains to be determined.

References:

1. Carrat F, Tachet A, Rouzioux C, Housset B, Valleron AJ. Evaluation of clinical case definitions of influenza: detailed investigation of patients during the 1995-1996 epidemic in France. Clin Infect Dis **1999**; 28(2): 283-90.
2. Boivin G, Hardy I, Tellier G, Maziade J. Predicting influenza infections during epidemics with use of a clinical case definition. Clin Infect Dis **2000**; 31(5): 1166-9.
3. Lindsey JC, Ryan LM. Tutorial in biostatistics methods for interval-censored data. Stat Med **1998**; 17(2): 219-38.
4. Gomez G, Calle ML, Oller R, Langohr K. Tutorial on methods for interval-censored data and implementation in R. Statistical modelling **2009**; 9(4): 259-97.
5. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med **1996**; 15(4): 361-87.
6. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods **2009**; 14(4): 323-48.