Semistructured black-box prediction: proposed approach for asthma admissions in London

Ireneous N Soyiri1,2
Daniel D Reidpath1
1Global Public Health, School of Medicine and Health Sciences, Monash University, Malaysia; 2School of Public Health, University of Ghana, Accra, Ghana

Abstract: Asthma is a global public health problem and the most common chronic disease among children. The factors associated with the condition are diverse, and environmental factors appear to be the leading cause of asthma exacerbation and its worsening disease burden. However, it remains unknown how changes in the environment affect asthma over time, and how temporal or environmental factors predict asthma events. The methodologies for forecasting asthma and other similar chronic conditions are not comprehensively documented anywhere to account for semistructured noncausal forecasting approaches. This paper highlights and discusses practical issues associated with asthma and the environment, and suggests possible approaches for developing decision-making tools in the form of semistructured black-box models, which is relatively new for asthma. Two statistical methods which can potentially be used in predictive modeling and health forecasting for both anticipated and peak events are suggested. Importantly, this paper attempts to bridge the areas of epidemiology, environmental medicine and exposure risks, and health services provision. The ideas discussed herein will support the development and implementation of early warning systems for chronic respiratory conditions in large populations, and ultimately lead to better decision-making tools for improving health service delivery.

Keywords: asthma, health care, black-box forecast, chronic, epidemiology, environment, respiratory risk

Introduction

Asthma is a global public health problem and the most common chronic disease among children. It is underdiagnosed and undertreated, and constitutes a huge burden on individuals, societies, and institutions. Current estimates suggest that as many as 300 million people are affected worldwide, and the burden of this chronic respiratory condition is rising, particularly among children. Recent reviews on asthma reaffirm the highly heterogeneous nature of the disease, which is also influenced by a number of complex genetic and environmental factors. Many of these reviews have comprehensively addressed key factors which contribute to the manifestation and progression of asthma in individuals, as well as laboratory-based experiments. However, there is little information on the forecasting of asthma events with the purpose of providing early warning systems to help in the management of the condition at the population level.

Asthma has a predictable prognosis. However, its diagnosis remains a challenge because the disease is not clearly defined by a particular set of conditions, but by a mix of several dynamic factors. These include numerous and quite unpredictable...
underlying genetic and environmental factors, as well as occupational factors. As a result of the complex nature of the condition, some of the diagnostic techniques commonly used range from history and patterns of symptoms, physical examination, and lung function measurements including spirometry, through to skin test identification of allergens.

Several studies have shown changes in the global epidemiology of asthma. Developed countries have consistently shown dramatic increases in prevalence, and this change has more recently been observed in some less developed countries as well. The United Kingdom National Asthma Campaign reported that asthma affected over five million people, ie, about one in every five households. Subsequent reports indicated that the United Kingdom had one of the highest prevalence rates, at over 15%. In England, 67,077 people were hospitalized for asthma between April 2006 and March 2007, of whom more than 40% were children under the age of 15 years. According to the Hospital Episode Statistics of the Department of Health (January 2001–December 2006), London, which is the busiest and most densely populated area in the United Kingdom, recorded about 57,000 asthma-related hospital admissions over that period, giving a crude annual rate of around 9500 (ie, an absolute estimate). This situation presents asthma as an important condition of public health concern, with dimensions not just limited to the individual(s) affected, but also posing a significant burden on health care resources as well as society at large, and therefore a need for preparedness. A significant factor in making preparation is the capacity to forecast what to expect in terms of health systems demands.

The use of algorithms in forecasting, also known as black-box forecasting, provides a novel approach to achieve a better outcome compared with most traditional structural/causal modeling techniques. Black-box forecasting involves a theoretical association of predictors with outcomes, the utility of which is strictly based on forecasting performance. Because it is theoretical, and can be a computationally exhaustive process, this sometimes leads to overfitting and poor performance on unseen datasets. On the other hand, structural models account for specific indicators/variables and require substantive knowledge of the subject matter in order to construct an intelligent model. They also overemphasize specific causes in an environment in which the complete causal process is poorly understood. Given the two approaches, a balanced semistructured black-box approach can be useful mid path in developing predictive models for health forecasting, where there is some prior knowledge of the relationship between a health condition and its environmental mediators.

We provide a brief overview of asthma and its environmental causes, with perspectives from the United Kingdom pertinent to other countries with similar populations. One of the key aims of the review is the association between asthma and environmental factors which have potential roles in health forecasting. The other is on developing semistructured black-box approaches that are predictive and can hypothetically forecast asthma events.

### Literature search and study approach

In preparing this discussion paper, a scoping of the literature on asthma and associated environmental factors, as well as approaches adopted for managing the condition, was conducted using medical-related databases including PubMed (Medline), Web of Science, and Google Scholar. In addition, citation mapping was used to search and retrace the literature from the initially selected key papers and documents. All the papers and documents identified were synthesized and summarized according to the objective of this paper.

The paper presents reviewed literature on asthma, with a focus on environmental triggers which have been reported, particularly for their effects on respiratory health, including highlights from studies that support the links between environment and health. It further illustrates a wider scope of factors associated with asthma and the manifestation of its symptoms in a framework. The second part of the paper focuses on describing proposed statistical approaches for developing predictive forecasting models. One of these approaches (the negative binomial regression predictive model) is exemplified using data on daily admissions for asthma in London (2001–2006) as well as synthetically generated temporal dummy variables. The paper also briefly discusses a variety of selection strategies for predictive modeling.

### Asthma and the environment

Local environmental conditions are important in determining the impact or manifestation of asthma. Factors such as temperature, humidity, and air pressure, as well as air pollutants, all interact to affect the occurrence of asthma, but do not have exclusively independent effects on the condition. The impact of these complex environmental factors and their interrelationships in health has never been fully understood, even though understanding the key
pathways of some individual factors has played an important role in developing a number of therapies.11,12 The Department of Health and the then Health Protection Agency in the United Kingdom published a review of the health effects of climate change in 2008.33 In the report, two component issues of the environment, which are principally of interest, are weather and air quality, and these were both considered. It is noted that the constituent indicators of temperature, humidity, vapor/pressure, and relative humidity on the worsening of asthma symptoms.50–58 Many of these studies have used the association between weather and disease incidence, hospitalization, or mortality to examine the relationship. For instance, the effect of temperature on general practitioner consultations for respiratory disease was observed, and it was found that there could be up to 15 days of delayed effect of cold temperatures on the incidence of respiratory illness.59 Also, constant seasonal variability in asthma admissions among children was found in Athens, Greece, where relative humidity and atmospheric pressure were established as key determinants.60

The relationship between asthma and environmental conditions is affected in complex ways, and it is worth noting that these effects have different associations depending on location. Asthma events in Mexico, for instance, are associated with the rainy season, whilst in England and Wales, asthma events are more strongly associated with seasonal temperature change rather than rainfall.54,57 Furthermore, it has been observed in the United Kingdom and Taiwan that peaks in asthma events occur in the winter and autumn seasons, but not in summer.59,61,62 Given the importance of context, it becomes critical to understand local relationships between asthma, weather, air quality, and season. However, even when local relationships are well understood, it remains difficult to predict extreme asthma events, ie, unusual peaks in asthma events that fall outside the usual fluctuations associated with seasonal changes and variations in weather and air quality. Forecasting these risks is complex and uncertain, but also requires specific data on a very long-term basis.53 Meanwhile, the use of semistructured black-box approaches in forecasting routine and/or extreme asthma events has not been comprehensively explored.

The issues discussed above are quite global in many respects. The hypothetical flow diagram (Figure 1) illustrates the relationship between asthma symptoms and immediate or underlying causes. Asthma is manifested by an inflammation and/or subsequent obstruction of air flow within the respiratory system.64,65 It is known that inflammation of the airways in an individual can result in asthmatic symptoms. Alternatively, inflammation may lead to obstruction of air flow directly or indirectly by causing hyperresponsiveness of the airways, ie, a state characterized by easily triggered contraction of the small airways (spasm), which may then cause obstruction of the airways.

Predicting asthma episodes in an ideal situation would require that we account for all the “known” potential predictors/indicators, which of course includes all immediate and underlying mediating factors. However, the availability
of data that is usable is a common limitation. These factors and interrelationships, as illustrated in Figure 1, may present some clues to data sources that can be mined and used for forecasting asthma.

### Semistructured black-box modeling

Among the many health issues that can be forecast, the need for emergency care is the commonest form of health forecasting.66–71 This is particularly related to hospital bed occupancy or number of visits to the emergency room. Although popular, there have been some challenges associated with forecasting the demand for emergency department services. However, this paper does not necessarily focus on these lapses, given that they have been discussed more elaborately elsewhere.72

The number of daily asthma admissions or routine measures for similar health conditions can be presented as integer value indicators (also referred to as rate data). This type of count data is common in many disciplines, can form a time series, and thus be used for causal or predictive modeling and forecasting.

Causal models are constructed to provide an explanation of model parameters. Hence, for a given outcome variable, Y will be defined by a function of the variables (X) known to have causal links with the dependent (outcome) variable Y, plus random noise and then the parameter error (Equation 1). In relation to asthma (Figure 1), Y could be represented as “primary care provider visit” for affected individuals, whilst X could be any one or more underlying causes (eg, air quality, weather, and temporal factors) associated with the affected individuals. The autoregressive integrated moving average (ARIMA) is the commonest technique used in this kind of health forecasting.

\[
\text{Outcome (Y)} = \text{Function of} \ (Xs + \text{random noise} + \text{parameters error}) \quad (1)
\]
In the black-box approach to modeling, formulation of the predictive model does not require prior knowledge of causal links. As illustrated in Figure 2, the process of predicting an outcome involves generation of suitable predictors and models, which are then validated before use in predictions.

There are more studies on causal (structural) modeling than on predictive (black-box) modeling, but the focus of this paper is on semistructured black-box modeling. For the latter approach, even though selection of variables is based originally on prior knowledge, it may also include important predictive factors/variables that have no causal relationship merely because those that end up being used in forecasting are based on their predictive capacity and not just on conformance to a particular theoretical relationship(s). This means that the approach is data-driven. Although data-driven approaches have sometimes created quite tense disagreements between causal modelers and predictive modelers, both approaches have their roles, and in the empirical forecasting and data mining areas, data-driven approaches are generally regarded as superior for the purposes of forecasting and out-of-sample prediction. In some specific context (as in the use of negative binomial models which is described subsequently), their outputs could still have an extended or additional use for describing relationships in past data.

In modeling, count regression models, such as the Poisson or negative binomial model, are most suitable. This is because they have the advantages of being able to handle time series data and their autocorrelations, whilst adjusting for any potential intercorrelation between dependent and independent predictors. Generally, count models are estimated using the maximum likelihood, which computationally proceeds iteratively until there is a convergence of the log likelihood. The exact choice of an appropriate count model depends on many factors which are directly related to the properties of the primary variable, such as the skewness of the distribution (kernel density) and the proportion and distribution of “0s” within the dataset. Figure 3 illustrates two major pathways for selecting a suitable count model, ie, a step-by-step approach and a one-stop test selection criterion, which involves the likelihood approach suggested by Long and Freese, and is also available as an application in Stata statistical software.

Using a sample of hospital admission data on asthma, two statistical methods are proposed, ie, a negative binomial model and a quantile regression model, for the development of predictive forecasting models that are aimed at predicting a future/anticipated event(s) and at predicting peak events. The asthma dataset has already been described elsewhere in the nationally recorded hospital episode statistics maintained by the National Health Service in the United Kingdom, and has also been used in some preliminary studies. Other data sources from which potential predictors could be extracted or derived include environmental data containing routine measures of weather and air quality indicators. Such data are accessible from the databases of the United Kingdom Meteorological Office. Additional variables (eg, temporal effects like day of the week or month of the year) can also be generated in addition to these data, to help in further investigations.
Predictive modeling with negative binomial models

Negative binomial models are applicable in developing both univariate and multivariate forecasting models. Given an expected number of daily admissions for asthma, the negative binomial regression can be presented in the form:

\[
Pr(Y = y | \lambda, \alpha) = \frac{[\Gamma(y + \alpha^{-1})/y!\Gamma(\alpha^{-1})]}{[\alpha^{-1}/(\alpha^{-1} + \lambda)]y^{\alpha^{-1}} \cdot [\lambda/\alpha^{-1} + \lambda]^{\alpha^{-1}}}
\]

(2)

where \(\lambda\) is the mean of the distribution, \(\alpha\) is the over dispersion parameter, \(y\) is the number of daily asthma admissions, and \(\Gamma\) is the gamma function.

A positive coefficient in the regression output indicates that a factor will increase the number of daily asthma admissions relative to its reference category. Conversely, a negative coefficient will decrease the number of daily asthma admissions relative to its reference category. The exponent of the coefficient can be interpreted, all other things being equal, as the proportionate increase (for values > 1) or decrease (for values between 0 and 1) in number of daily asthma admissions associated with a one unit increase in the explanatory variable.\(^9,79\)

Obtaining an improved fit for a model in this situation can be established by inspecting the Akaike information criterion.\(^9,14,80–82\)

Figure 4 Asthma daily admissions and predictive model based on month and week day. (A) Model development sample (hold-in dataset). (B) Model validation sample (hold-out dataset).
To exemplify this proposed approach, we present an analysis and outputs based on hospital episode statistics data on asthma in London, which have been described elsewhere. The dataset, which contains nonidentifiable individual records of asthma patients visiting their general practitioner or the emergency department, was transposed by daily sums to provide a time series dataset of total daily attendance. For the purpose of convenience, and also because of the lack of compatible and adequate data on other potential predictors of asthma already mentioned in the literature above (Figure 1), we only illustrate with a temporal predictive model (ie, multivariable model based on temporal factors). The temporal factors we considered included seasons (spring, summer, autumn, and winter), month of the year (January, February through December) and day of the week (Sunday, Monday through Saturday), which were synthetically generated from the date of attendance record using statistical software.

A hold-in sample of the data was used in model development, and validation was done with a hold-out sample.

**Figure 5** Asthma daily admissions and predictive model based on season month and week day. (A) Model development sample (hold-in dataset). (B) Model validation sample (hold-out dataset).
of the data from January 1, 2006 to December 31, 2006. Three bivariate and four multivariate predictive models were generated (Figures 4 and 5, Appendix 1) and compared based on their Akaike information criterion (Table 3). These models were cross-validated with the hold-out dataset. For instance, in Figure 4A, the light blue (or gray) time series plot shows total daily asthma admissions in the London area, and the red (or dark) plot shows the predicted multivariate model based on month and week day. However, Figure 4B shows the validation output. The predicted plot reasonably tracks the real distribution of asthma admissions, although it misses out on extreme variations. Even though the focus of the work is on predictive modeling and not causal modeling, the negative binomial regression outputs for these models are also presented (Tables 1 and 2).

Findings from the Akaike information criterion model fitness tests (Table 3) show that the “day of the week model” (III) was the least performing model, whilst the VI and VII multivariate models outperformed all the others, and by as much as 2.4% of III in the hold-in data and over 6.6% in the holdout dataset. Model V was slightly better than both II and IV in the hold-in dataset, and conversely less fit than the latter in the hold-out dataset. Temporal factors have been used to predict exacerbations of chronic respiratory diseases, such as asthma and chronic obstructive pulmonary disease. In addition to temporal factors, other studies have used allergens, weather, and air quality factors in specific areas to predict asthma hospital admissions. Hence our findings should be interpreted with caution because the dataset used represents a large and diverse population for London, and does not account for other important predictors of asthma, such as weather and air quality. Nonetheless, predicting asthma events can have policy implications for planning and executing health care delivery. This may then produce some indirect benefits pertaining to health budgets and resource allocation.

### Forecasting extreme events with quantile regression models

This paper also proposes the use of quantile regression models in forecasting peak events as a more ideal approach compared with others. Quantile regression is an extension of the linear model, and is better equipped to characterize the relationship between a response distribution and explanatory variable(s) for selective quintiles. It has been used more extensively in other areas of forecasting, but has yet to be fully tapped in health forecasting. Quantile regression models can be considered as fitting a linear model to a cross-section of the data/distribution within the anticipated range.

Using the asthma data as an example, for a peak number of daily admissions, the quantile regression model is presented in the form:

$$ Y_i = \beta_0^{(p)} + \beta_1^{(p)} x_i + \epsilon_i^{(p)} $$

where $Y_i$ is asthma hospital admissions for a given day, $\beta_0^{(p)}$ is a constant term, $\beta_1^{(p)}$ is a coefficient of exposure term, $x_i$ is the exposure term, $\epsilon_i^{(p)}$ is the error term, and $p$ is the quantile. Further illustration of the quantile regression model equation above is available elsewhere.

The pseudo $R^2$ (comparable with the $R^2$ for least-squares procedures) is the coefficient of determination for quantile regression, and it represents the goodness-of-fit statistic, which is most appropriate for comparing models of specific quantiles. It is based on change in the deviance statistic, and ranges between 0 and 1. The pseudo $R^2$ is thus estimated as:

$$ \text{Pseudo } R^2 = 1 - \frac{\text{Sum of deviations about the estimated quantile}}{\text{Sum of deviations about the raw quantile}} $$

### Table 1 Bivariate temporal models of daily asthma admissions in London for 2001–2005

<table>
<thead>
<tr>
<th>Temporal factors</th>
<th>Coefficient</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Seasonal model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring*</td>
<td>–0.0458**</td>
<td>–0.0872 to –0.0043</td>
</tr>
<tr>
<td>Autumn</td>
<td>0.2378***</td>
<td>0.1974 to 0.2783</td>
</tr>
<tr>
<td>Winter</td>
<td>0.0758***</td>
<td>0.0347 to 0.1169</td>
</tr>
<tr>
<td><strong>Month of year model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>January*</td>
<td>0.0358</td>
<td>–0.0352 to 0.1069</td>
</tr>
<tr>
<td>February</td>
<td>0.0065</td>
<td>–0.0631 to 0.0761</td>
</tr>
<tr>
<td>March</td>
<td>0.0168</td>
<td>–0.0533 to 0.0869</td>
</tr>
<tr>
<td>April</td>
<td>0.0733*</td>
<td>0.0042 to 0.1424</td>
</tr>
<tr>
<td>May</td>
<td>0.0865**</td>
<td>0.0169 to 0.1561</td>
</tr>
<tr>
<td>June</td>
<td>–0.0252</td>
<td>–0.0951 to 0.0447</td>
</tr>
<tr>
<td>July</td>
<td>–0.1061***</td>
<td>–0.1767 to –0.0356</td>
</tr>
<tr>
<td>August</td>
<td>0.3112***</td>
<td>0.2431 to 0.3792</td>
</tr>
<tr>
<td>September</td>
<td>0.2451***</td>
<td>0.1771 to 0.3130</td>
</tr>
<tr>
<td>October</td>
<td>0.2553***</td>
<td>0.1868 to 0.3237</td>
</tr>
<tr>
<td>November</td>
<td>0.2628***</td>
<td>0.1950 to 0.3307</td>
</tr>
<tr>
<td><strong>Day of week model</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sunday*</td>
<td>0.1376***</td>
<td>0.0831 to 0.1922</td>
</tr>
<tr>
<td>Monday</td>
<td>0.0488</td>
<td>–0.0062 to 0.1038</td>
</tr>
<tr>
<td>Tuesday</td>
<td>–0.0364</td>
<td>–0.0918 to 0.0190</td>
</tr>
<tr>
<td>Wednesday</td>
<td>–0.0702***</td>
<td>–0.1258 to –0.0146</td>
</tr>
<tr>
<td>Thursday</td>
<td>–0.0705***</td>
<td>–0.1261 to –0.0149</td>
</tr>
<tr>
<td>Friday</td>
<td>–0.1275***</td>
<td>–0.1834 to –0.0716</td>
</tr>
</tbody>
</table>

**Notes:** *Reference category; **p < 0.05; ***p < 0.01; ****p < 0.001.**

**Abbreviations:** Coefficient, coefficient of the negative binomial regression; CI, confidence interval.
### Table 2 Multivariable temporal models of asthma daily admissions in London, 2001–2005

<table>
<thead>
<tr>
<th>Temporal factors</th>
<th>Season and month model</th>
<th>Season and day of week model</th>
<th>Month and day of week model</th>
<th>Season, month, day of week model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>95% CI</td>
<td>Coefficient</td>
<td>95% CI</td>
</tr>
<tr>
<td><strong>Season</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summer</td>
<td>-0.1794***</td>
<td>-0.2494</td>
<td>-0.1094</td>
<td>-0.0847</td>
</tr>
<tr>
<td>Autumn</td>
<td>0.1819***</td>
<td>0.1141</td>
<td>0.2498</td>
<td>0.1979</td>
</tr>
<tr>
<td>Winter</td>
<td>-0.0733*</td>
<td>-0.1424</td>
<td>-0.0042</td>
<td>0.0758***</td>
</tr>
<tr>
<td><strong>Month</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>January</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>0.0358</td>
<td>-0.0152</td>
<td>0.1069</td>
<td>0.0351</td>
</tr>
<tr>
<td>March</td>
<td>-0.0668</td>
<td>-0.1359</td>
<td>0.0022</td>
<td>0.0077</td>
</tr>
<tr>
<td>April</td>
<td>-0.0565</td>
<td>-0.1261</td>
<td>0.0131</td>
<td>0.0159</td>
</tr>
<tr>
<td>May</td>
<td>(omitted)</td>
<td></td>
<td>(omitted)</td>
<td>0.0731*</td>
</tr>
<tr>
<td>June</td>
<td>0.1926***</td>
<td>0.1221</td>
<td>0.2630</td>
<td>0.0897**</td>
</tr>
<tr>
<td>July</td>
<td>0.0809*</td>
<td>0.0102</td>
<td>0.1517</td>
<td>-0.0261</td>
</tr>
<tr>
<td>August</td>
<td>(omitted)</td>
<td></td>
<td>(omitted)</td>
<td>-0.1050**</td>
</tr>
<tr>
<td>September</td>
<td>0.0559</td>
<td>-0.0109</td>
<td>0.1227</td>
<td>0.3092***</td>
</tr>
<tr>
<td>October</td>
<td>-0.0102</td>
<td>-0.0769</td>
<td>0.0565</td>
<td>0.2445***</td>
</tr>
<tr>
<td>November</td>
<td>(omitted)</td>
<td></td>
<td>(omitted)</td>
<td>0.2554***</td>
</tr>
<tr>
<td>December</td>
<td>0.2628***</td>
<td>0.1950</td>
<td>0.3307</td>
<td>0.2635***</td>
</tr>
<tr>
<td><strong>Week day</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sunday</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>0.1382***</td>
<td>0.0868</td>
<td>0.1895</td>
<td>0.1385***</td>
</tr>
<tr>
<td>Tuesday</td>
<td>0.0494</td>
<td>-0.0024</td>
<td>0.1012</td>
<td>0.0513*</td>
</tr>
<tr>
<td>Wednesday</td>
<td>-0.0339</td>
<td>-0.0862</td>
<td>0.0184</td>
<td>-0.0335</td>
</tr>
<tr>
<td>Thursday</td>
<td>-0.0685***</td>
<td>-0.1210</td>
<td>-0.0161</td>
<td>-0.0674**</td>
</tr>
<tr>
<td>Friday</td>
<td>-0.0682***</td>
<td>-0.1207</td>
<td>-0.0157</td>
<td>-0.0669***</td>
</tr>
<tr>
<td>Saturday</td>
<td>-0.1243***</td>
<td>-0.1771</td>
<td>-0.0715</td>
<td>-0.1259***</td>
</tr>
</tbody>
</table>

**Notes:** #Reference category; *P < 0.05; **P < 0.01; ***P < 0.001.

**Abbreviations:** Coefficient, coefficient of the negative binomial regression; CI, confidence interval; omitted, omitted because of collinearity.
Table 3 Comparison of model fitness using the Akaike information criterion

<table>
<thead>
<tr>
<th>Number</th>
<th>Temporal models</th>
<th>Hold-in model</th>
<th>Hold-out model</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Seasonal model</td>
<td>12647.92</td>
<td>2578.40</td>
</tr>
<tr>
<td>II</td>
<td>Month of year model</td>
<td>12560.37</td>
<td>2518.14</td>
</tr>
<tr>
<td>III</td>
<td>Day of week model</td>
<td>12746.41</td>
<td>2681.11</td>
</tr>
<tr>
<td>IV</td>
<td>Season and month model</td>
<td>12560.37</td>
<td>2518.14</td>
</tr>
<tr>
<td>V</td>
<td>Season and day of week model</td>
<td>12532.29</td>
<td>2568.72</td>
</tr>
<tr>
<td>VI</td>
<td>Month and day of week model</td>
<td>12436.21</td>
<td>2504.38</td>
</tr>
<tr>
<td>VII</td>
<td>Season, month, and day of week model</td>
<td>12436.21</td>
<td>2504.38</td>
</tr>
</tbody>
</table>

Note: %, percentage improvement of model fit over the least performing model (III).
Abbreviation: AIC, Akaike information criterion.

Variable selection approaches
The variable selection approach is critical in modeling because it determines the final functional form, which is subsequently used for forecasting. As discussed earlier, there is a wide range of environment-related variables known to influence the incidence and/or exacerbation of asthma and other respiratory illnesses. However, common limitations in using environmental measures to forecast a wide population health issue like asthma include a reliable data source and its quality, as well as the extent to which these measures can represent an individual’s level of exposure. The inclusion of temporal components as independent factors in a predictive model (eg, day of the week, month, season), accounts for any temporal kinetics and also allows for the identification of lag times, which may improve predictions.78,92

There are equally unlimited approaches that can be adopted for the selection of variables in predictive modeling.27 These approaches range from selection by convenience to computationally exhaustive searches for the best combination of predictors.27,28 The selection of potential predictors can also be biased by our common understanding of the mechanisms by which environmental agents cause diseases, particularly for respiratory illnesses. However, the ultimate goal is that one obtains a reliable and parsimonious forecasting model.

Backward elimination is one of the commonest approaches utilized to select an appropriate model. This method involves a systematic and/or automatic procedure of reducing a base model, ie, a multivariable model consisting of all possible predictors that are either independently strongly correlated with the dependent variable, or are of already known importance (ie, removing variables that are not statistically significant), while ensuring that the fit (ie, with either AIC or pseudo-R²) is improved or at least maintained.

Another strategy of variable selection in modeling is to conduct an exhaustive search of the best fitting model using all possible combinations of the available predictors. This approach is however computationally intensive and has a very high chance of over fitting the model. Nonetheless, such computationally exhaustive approaches are most adapted to the novel idea of semistructured black-box models.28

Conclusion
Asthma poses a great burden to populations. Discrete measures of the incidence of the disease can be used for forecasting. Though environmental factors have specific effects on the disease, and these effects often vary by location, they may still provide supplementary variables for developing a forecast.

Two methods for developing predictive models as well as variable selection strategies, which have potential roles in semistructured black-box forecasting models, have been discussed. Both negative binomial models and quantile regression models are applicable to integer value health indicators (eg, total daily hospital admissions records). The negative binomial models predict anticipated events, whilst quantile regression models are designed to predict peculiar events. These kinds of forecast models vary in their complexity and methods, depending on the specific health condition and population data. The idea of semistructured black-box predictive modeling may stimulate further research on asthma and, possibly, health forecasting.

Disclosure
The authors declare they have no competing interests in this work.

References


75. Long JS, Freese J. Regression Models for Categorical Dependent Variables Using Stata. College Station, TX: Stata Press; 2000.

76. Stata SE. Release 11. Statistical Software. College Station, TX: StataCorp LP; 2011.


Semistructured black-box prediction of asthma admissions