



Article

Predicting Health Risks of Adult Asthmatics Susceptible to Indoor Air Quality Using Improved Logistic and Quantile Regression Models

Wan D. Bae ¹, Shayma Alkobaisi ^{2,*}, Matthew Horak ³, Choon-Sik Park ⁴, Sungroul Kim ⁵, Joel Davidson ¹¹ Department of Computer Science, Seattle University, Seattle, WA 98122, USA² College of Information Technology, United Arab Emirates University, Al Ain 15551, United Arab Emirates³ Lockheed Martin Space Systems, Denver, CO 80221, USA⁴ Department of Internal Medicine, Soonchunhyang Bucheon Hospital, Bucheon 420-767, Korea⁵ Department of ICT Environmental Health System, Graduate School, Department of Environmental Sciences, Soonchunhyang University, Asan 336-745, Korea

* Correspondence: shayma.alkobaisi@uaeu.ac.ae

Abstract: The increasing global patterns for asthma disease and its associated fiscal burden to healthcare systems demand a change to healthcare processes and the way asthma risks are managed. Patient-centered health care systems equipped with advanced sensing technologies can empower patients to participate actively in their health risk control, which results in improving health outcomes. Despite having data analytics gradually emerging in health care, the path to well established and successful data driven health care services exhibit some limitations. Low accuracy of existing predictive models causes misclassification and needs improvement. In addition, lack of guidance and explanation of the reasons of a prediction leads to unsuccessful interventions. This paper proposes a modeling framework for an asthma risk management system in which the contributions are three fold: First, the framework uses a deep learning technique to improve the performance of logistic regression classification models. Second, it implements a variable sliding window method considering spatio-temporal properties of the data, which improves the quality of quantile regression models. Lastly, it provides a guidance on how to use the outcomes of the two predictive models in practice. To promote the application of predictive modeling, we present a use case that illustrates the life cycle of the proposed framework. The performance of our proposed framework was extensively evaluated using real datasets in which results showed improvement in the model classification accuracy, approximately 11.5–18.4% in the improved logistic regression classification model and confirmed low relative errors ranging from 0.018 to 0.160 in quantile regression model.



Citation: Bae, W.D.; Alkobaisi, S.; Horak, M.; Park, C.-S.; Kim, S.; Davidson, J. Predicting Health Risks of Adult Asthmatics Susceptible to Indoor Air Quality Using Improved Logistic and Quantile Regression Models. *Life* **2022**, *12*, 1631. <https://doi.org/10.3390/life12101631>

Academic Editors: K. H. Katie Chan, Ka-Chun Wong, Brian Chen and Jie Li

Received: 5 August 2022

Accepted: 4 October 2022

Published: 18 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: personalized asthma care; asthma risk prediction; exposome; indoor air quality; logistic regression; quantile regression; transfer learning; sliding window regression

1. Introduction

1.1. Asthma and Exposome

Asthma is one of the primary care-sensitive conditions that can be controlled and prevented, through effective care management such as the early recognition of high risk and prompt interventions [1]. A small fraction of asthma exacerbations, which are possibly avoidable by predictive risk modelings account for 63% of the annual total asthma cost in the US and significantly contribute to rising socioeconomic burden [2–6]. The increasing global patterns for asthma and its associated fiscal burden on patients and healthcare providers demand a change to care processes and the way asthma risks are managed. Personalized patient-centered care is a shift from the traditional professional-care model to a service-oriented model aimed at reducing the cost of any symptoms that can be treated outside hospitals. Computing technologies and emerging predictive analysis provide great

promise in promoting patient-centered care by empowering patients to participate actively in health risk control [7–12].

While factors responsible for increasing the risks of asthma exacerbation are not completely understood, approximately 70–90% of chronic respiratory diseases are attributed to environmental factors, response to which varies significantly over the population [13–15]. The term “exposome” refers to the assessment of different environmental exposures’ effect on human health [16,17] and exposures are calculated based on a specific time range, at a specific location and under certain environmental circumstances [18]. Exposome analytics seeks to discover effects of environmental factors on the health of individuals by integrating time and location, as well as behavioral patterns to estimate individual exposure, and then predict health risks of individuals. Indoor air quality, particularly at home, has been recognized as a major source of exposure to heightened asthma triggers [19–21]. Most people, elderly people in particular, spend about 80–90% of their time indoors [21–24]. In addition, the modern home is highly thermally insulated to improve energy efficiency, often to be a detriment to indoor air quality [21]. While the list of known or suspected asthma triggers include many variables (e.g., air pollutants, allergens, certain food, stress, etc.), the present study focuses on indoor air quality that can be monitored on an individual patient-specific basis in real-time and over which patients have significant control in terms of the level of exposure to each.

1.2. Asthma Care and Management

Recent health applications appear to positively influence asthma risk management around the globe [25–28]. Several studies indicate that patients using mobile asthma self-management apps (e.g., AsthmaSense, AsthmaMD, Propelle and ADAM) have significantly improved quality of life scores. Subsequently, those patients were less likely to visit emergency departments due to asthma-related complications [29–31]. Additionally, research in [32] proposes a model that incorporates patient’s history of readmission and impacts of patient attribute changes over time on a tree-based classification method to estimate the probability of readmission. In a different direction, several research works present pattern recognition models to find complex interrelations between air pollution, weather, and asthma exacerbation. In [33], authors proposed a method that extracts related features and uses supervised learning approaches such as classification models to detect adverse health events. Overall, however, while systems integrating environment measurement techniques with predictive analytics promise successful implementation of tailoring care to individual patients and thus for transforming the future of healthcare, existing predictive health analytics provide limited help in creating efficient tailored care plans [34]. This continues to be one of the most challenging problems in environmental health research [35–38].

One major challenge in predictive analytics in asthma is that asthma exacerbations resulting in hospitalization and emergency room visits are rare events and current predictive models exhibit unsatisfactory accuracy for risk analysis of such events [39,40]. This is mainly due to imbalanced datasets where the high risk zone is much smaller than the normal zone but also partially due to the small size of individual patients’ datasets. For example, recording one observation per day provides only 180 data points over a 6 month period and is not sufficient to develop a neural network based model, while several over-sampling methods have been proposed to solve the imbalanced dataset problem [41–44] in order to improve the accuracy of classification models, little is known about their effectiveness on the models with small sized training data. On the other hand, most machine learning techniques require a large amount of data to train high quality models. This problem is acute for practical and realistic use of predictive models in health applications where datasets for individuals are frequently very small.

An asthma exacerbation can be the result of a single or a mixture of environmental triggers that are of spatiotemporal nature. In addition, measurements of individual exposures in space and time are affected by many factors and governed by complex interactions and relationships between environmental and human systems [45,46]. However, these

spatio-temporal properties of environmental data and human behavioral changes are not fully captured in the existing predictive models. The review in [47], which assesses the effectiveness and feasibility of using smartphones and tablet apps to facilitate asthma self care and management, highlights a gap in considering the environmental impact and the seasonal nature of asthma. This environmental impact could improve the efficacy of apps as standalone interventions [48]. Another shortcoming of current predictive models is in interpreting the outcomes of predictions and the lack of guidance that helps health professionals utilize to the largest extent possible their domain knowledge in the process of prediction modeling and applying the results to patient interventions [49].

1.3. Our Contributions

In this paper, we focus on the problem of estimating the probability that a patient will experience a critical asthma exacerbation given the patient's exposures to indoor environmental factors. We address the limitations of existing predictive models and propose a framework to improve the predictions and applicability of logistic regression and quantile regression models. In the framework, two regression models collaborate together to make predictions. As the framework continuously collects data, the models evolve with newly updated parameters and hyper-parameters.

The first improvement applied in our framework is the deep learning technique of transfer learning to overcome the performance issues of logistic regression as classifier due to the limited size of training data. Our proposed TL method trains logistic regression models through three phases, the first phase is training a fully connected neural network source model using population data, and it is further tuned for target model using an individual patient's data in the second phase. In the last step, the last layer outputs of the target model are inputted to logistic regressors. The second improvement method is a method of variable sliding window to improve the accuracy of individual patient risk estimation regression models.

In classification modeling, we remark that we studied many other classification models for the target model, which are known to have modest data requirements, such as decision trees, random forests, and support vector machines. We found that logistic regression was the best, so for simplicity of exposition, we restrict our attention to that model. Furthermore, while this paper focuses on asthma risk prediction, it is worth mentioning that the proposed solutions can be transferred to other environmental chronic diseases with adjustments.

2. Materials and Methods

In this section, we present our approach to design and implement an asthma risk management system. Figure 1 presents the conceptual design model for the proposed framework that consists of the following components: (1) real-time data acquisition and management; (2) exposure estimation and risk prediction; (3) daily interventions and feedback, and assessment. The model integrates measurements of environmental and physiological conditions, estimation of exposure, evaluation of current health state and prediction of adverse health events. In the proposed machine learning framework, individuals' exposures to indoor environmental factors are collected in real-time along with their daily routines, which are used as parameters to assess the interactions between asthma risk and indoor air quality. The outcomes of the prediction are then used in the targeted interventions to reduce the probability of high risk.

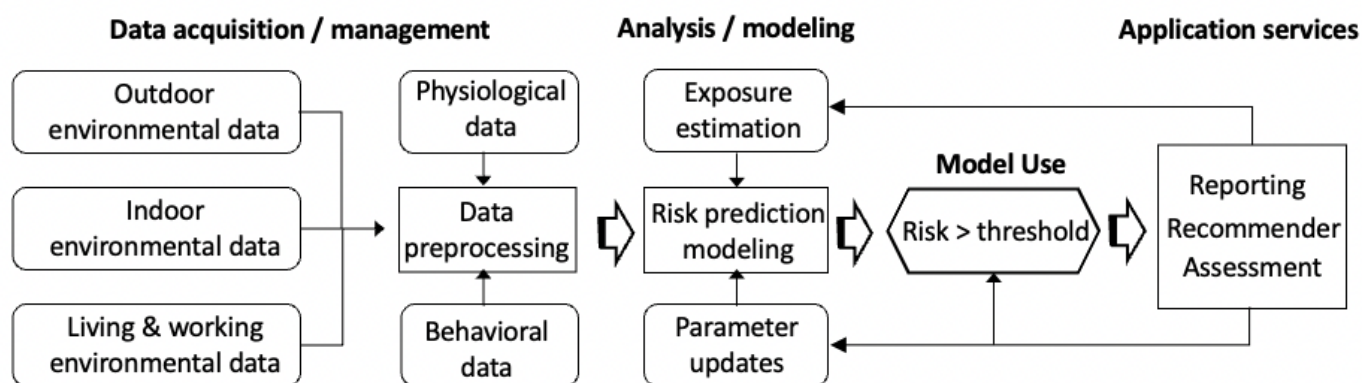


Figure 1. An overview of asthma risk management.

2.1. Study Design

2.1.1. Asthma Risk Measurement

The peak expository flow rate (PEFR) measurement is one of the primary health-level indicators used in asthma care and management [50]. One of the uses of this measurement is to quantify an individual's asthma exacerbation level and provide the medical practitioners with basis for better understanding of the predictions and thus, better decision-making outcomes. The significance of PEFR measurement is categorized into three zones, green, yellow and red, using a standardized "normal" value that is established by the American Lung Association based on population-level data using gender, age and height information [51]. One drawback of estimating asthma severity based on population-level norms in this way is that the high variability of PEFR values within the population makes its applicability on the individual level unrealistic.

In this study, we base our forecasts on a simplified version of the individual-based asthma risk zoning method proposed in [52]. The method allows for classification of a patient's condition into several zones based on the patient's own historical distribution of PEFR values. For the purposes of prediction of high risk days in this paper we use only two zones, a "safe zone" which we nominally take to be PEFR values in the upper 80% of the patient's historical PERF values and a "risk zone" taken to be PEFR values in the lower 20% of the probability density function distribution. The PEFR value dividing the two zones is called the critical PEFR value, $PEFR_C$. The objective of the inference engine will be to predict when a patient is in danger of entering the risk zone, which is understood to be a potential medical emergency where severe airway narrowing is likely to occur and immediate action may be necessary. Therefore, as discussed in [52] it is important for doctors and patients together to analyze the patient's PEFR distribution as it relates to the patient's actual health condition and take care to modify the 80/20 cutoff as necessary. Accordingly, the system is able to evaluate the susceptibility of each individual to an asthma exacerbation, on an individual basis as the value of the exposure to variables changes over time.

The models aim to estimate the probability, $P(y < PEFR_C)$, that a patient's PEFR value today will fall below their or her critical value. Even though $PEFR_C$ may have been established based on the lower, say, 10% or 20% of the patient's PERF values, the probability of falling below $PEFR_C$ depends on many environmental factors and hence use of a various values of $PEFR_C$ can be used to measure a patient's daily health risk. Figure 2 illustrates the distributions of morning PEFR values of the 19 patients who were participants in our study. The blue dotted line represents the average of 20% quantile value of the 19 participants' PEFR data while the red line inside the boxplot represents 20% quantile value of a particular patient's PEFR data.

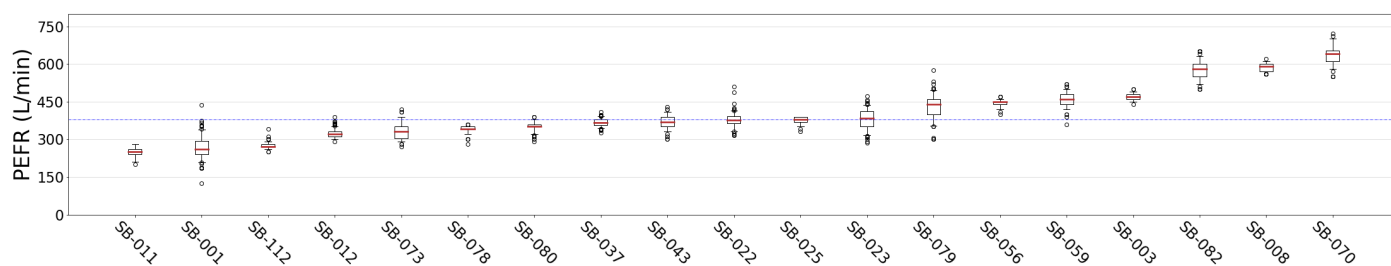


Figure 2. Distributions of participants’ PEFR data.

2.1.2. Indoor Air Quality Measurement

Fine particulate matter of 2.5 μm (*PM*_{2.5}) concentration, carbon dioxide (*CO*₂), temperature, and relative humidity (dampness) are known to be some of the most important asthma risk factors among the list of known or suspected asthma triggers because they have high temporal variability, are strongly affected by participants’ activities and can be monitored in real-time.

In our risk prediction, we study aggregate exposure to those four variables constructed through various environmental factors and aggregation methods. These four air quality data points were obtained through laser-light scattering sensor with 2 min intervals in participants’ main living spaces. Each participant’s exposures to these variables were estimated using 24 h historical air quality data before the participant’s PEFR measurement. The air quality data stamped by time and location were collected in the individual’s daily routine and stored in our database server for data preprocessing of exposure estimation.

2.1.3. Population Data

A total of 19 participants were recruited from the adult asthma patients (aged 34 to 83 years) who had joined our ESCORT (environmental health smart study with connectivity and remote sensing technologies) study [53]. These participants were consulted and monitored by doctors and medical practitioners at Soonchunhyang University Bucheon Hospital, South Korea. In our study, all participants are non-smokers and occupants at their home are all non-smokers.

The patients’ daily PEFR values were collected twice a day (morning and evening) between 1 November 2017 and 31 May 2018 and the resulting dataset sizes vary between 118 days and 212 days with an average of 154 days. Participants agreed to keeping the air quality monitoring unit for monitoring and storing indoor air quality such as *CO*₂, *PM*_{2.5}, temperature and humidity, and writing their daily activities and place visited every 30 min in a diary provided. Several categorical data such as income, living situation and cooking habits were also collected. The comprehensive nature of our framework offers the opportunity to track spatiotemporal exposure patterns for each participant over a period of time and to capture participants’ daily activities. The environmental variables and measurement in this study are summarized in Table 1.

Table 1. Environmental variables and measurement.

Category	Variables	Measurement
Physiological data	yesterday’s PEFRs	twice a day (AM & PM)
Indoor air pollutants & other variables	<i>PM</i> _{2.5} , <i>CO</i> ₂ temperature, humidity	every 60 s interval via remote sensors installed at home
Cooking behavior	the frequency of frying	level 1 (every day)–level 7 (none)
Living environment	distance from home to major roads	level 1 (<1 m)–level 5 (>11 m)
Life style	income level	level 1–level 9

2.2. Exposure Estimation

To simplify the relationship between stochastic, spatio-temporal sequences of pollutant concentration and their physiological consequences, researchers have recently begun to entertain the notion of “exposome” [16], while exposome studies have uncovered many important relationships between environment and human health, the assessment of individuals’ exposure to air quality over time has been confined to population averages, rather than individualized estimates. Measurements of individual patient’s exposure to indoor air pollution is affected by many factors, such as concentration of pollution, location/time of the individual, physical activities (exertion) and behavior, and the human system [45,46]. As part of developing a tailored care plan for a patient utilizing indoor air quality control, historical data of a patient’s exposures to the targeted indoor environmental factors needs to be obtained as well as knowledge of the sources of air pollution and underlying characteristics of the exposure [20]. As more sensor technologies become available, they can be used to monitor real-time indoor air quality and to develop analytical models for asthma care management [16,54].

The proposed system calculates the impact of environmental exposure on individuals biomarkers (e.g., lung function level, PEFr) at any given time. It retrieves information on the concentrations of each air pollutant in the air of identified regions and the timeframe over which the exposure occurs. It then uses the general equation for exposure in (1) and more complex integrative models to quantify exposures. Exposure factors refer to any extra information that is required to calculate the exposure amount such as exposure rates or number of possible spaces, activity-patterns and body weight [53].

Each participant’s exposures to environmental variables can be estimated using a time window on historical air quality data before the participant’s PEFr measurement. The time window can be determined by medical expertise, where approximately 24 h window (between yesterday’s AM measurement time and today’s AM measurement time) was used in our experiments. Then the exposure amount for an activity can be calculated as follows:

$$E_j = \frac{\text{Conc} * \text{Inh.Rate}}{60 \text{ seconds}}, \quad (1)$$

where *Conc* is the environmental concentration (e.g., $PM_{2.5}$) per activity per person measured every 60 s, *Inh.Rate* is inhalation rate of adult patients based on ages, and *j* is type of activities. In the estimation of daily exposures to CO_2 , temperature and relative humidity, E_j is the mean value of the reported values within 24 h time interval.

Equation (1) indicates the accumulated amount of environmental exposure per minute for a particular activity *j*. Then, the accumulated daily exposure to an environmental variable is calculated by:

$$f(x_i, t) = \sum_{j=1}^N E_j T_{i,j}, \quad (2)$$

where x_i is the space (room) on which the individual stayed, $T_{i,j}$ is the time spent for activity *j* and E_j is the exposure amount for activity *j* per minute. The predefined activities and their characteristics can be found in [53].

2.3. Risk Prediction Modeling

Our proposed modeling framework utilizes two commonly used machine learning methods in medical applications, (1) logistic regression (LR) together with a neural network based transfer learning (TR) and (2) quantile regression (QR). The LR method, a popular machine learning method for classification problems, estimates the probability of an event (i.e., a binary response), such as positive or negative, based on a given set of independent variables. It is often used in medical domain to predict the risk of developing a given disease, based on observed variables of the patient [55]. In the context of probability modeling, LR finds an optimal $\Theta = \{\theta_0, \theta_1, \dots, \theta_n\}$, the set of coefficients for the linear combination $z = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$ of the *n* independent variables, which best estimates

the probability, p , of developing the disease (positive) when substituted into the logistic function $p = g(z) = 1/(1 + e^{-z})$. On the other hand, in machine learning contexts, LR is often used for binary classification where $p \geq 0.5$ results in the model outputting the positive class.

In a traditional clinical use, LR has the advantage of having a probability associated with output metrics that are relatively easily understood. An example of typical outputs would be “Based on an estimation of your exposure to air pollutants in last 24 h, you have a 50% chance of falling below your critical PEFr value today”. The patient can be also guided to follow a medical protocol to prevent an asthma exacerbation. One disadvantage is that patients frequently receive warnings for events that are highly critical but nonetheless have low probability. The situation that the critical event does not occur is considered as “false positives”. Therefore, it is important that the system gains more detailed information when LR estimates a non-negligible probability of a critical event. In our proposed two-step system, we use logistic regression only for classification of the patient’s next-day risk state (high-risk or low-risk) and quantile regression to provide more nuanced information to the patient regarding their overall likelihood of experiencing a critical exacerbation event.

The QR method estimates the conditional median or other quantiles of the response variable based on the values of explanatory variables. Linear regression attempts to find the best $\Theta = \{\theta_0, \theta_1, \dots, \theta_n\}$ for the linear equation $y = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$ of the n independent variables, which predicts the average value of the response variable y . The average value is referred to as the “conditional mean” of the distribution of y given explanatory variables (x_1, x_2, \dots, x_n) . QR on the other hand attempts to model the quantile values of the conditional distribution, hence it can approximate the whole conditional distribution of a response variable y [56]. QR has recently found use in many medical applications in which the more extreme values of a patient’s data are of particular interest [57]. In our modeling framework, QR works as the second method.

All machine learning techniques come with a set of advantages and disadvantages: LR models exhibit unsatisfactory accuracy in individual-level health risk prediction application where the size of training datasets is small. Supposing that we collect a patient’s data for 1 year, then the total number of data tuples is 365. Most machine learning algorithms underperform with this small dataset.

On the other hand, one environmental factor may lead to changes in several aspects of the distribution of other environmental variables, including changes in the mean, variability, and severity of extreme cases. Classical quantile regression analyze a single quantile or several quantiles separately [58]. Thus, performance of QR models in this context can be improved by taking into account time trends for each quantile level and time locality (i.e., recent data is used in training/validation).

We propose solutions to these problems in order to improve the performance of predictive models: (1) LR with neural network based transfer learning, and (2) QR with a variable sliding window method. The two improved methods are presented in the following subsections and the results of the improvement are presented in Section 3.

2.3.1. Logistic Regression Classification with a Neural Network Based Transfer Learning

One of the main challenges to improving prediction quality of LR models is the limited availability of large high quality labeled datasets [59]. The TL technique, one of deep learning techniques in machine learning, can help overcome a scarcity of data by focusing on fine tuning a pretrained model with a small amount of specialized training data [60]. This strategy has shown great promise in the medical field in the context of image analysis of MRI or CT scan data and images [61]. Authors in [62] reported results of a preliminary study of the effectiveness of transfer learning for asthma risk forecasting. Still, however, to date little research has been performed in the context of individual-level health risk prediction with limited training data.

We propose an improved TL + LR classification method as a pipeline: it trains a fully connected neural network (NN) (source model) with population data of the 18 asthma

patients (excluding a target patient) and then re-trains the NN with a target patient’s data (target model). The output of the last hidden layer of the target model of the NN model is pipelined into a LR model as input data. Finally, the logistic regression produces a prediction decision (classification) with a probability. The process of transfer learning based logistic regression is shown in Figure 3.

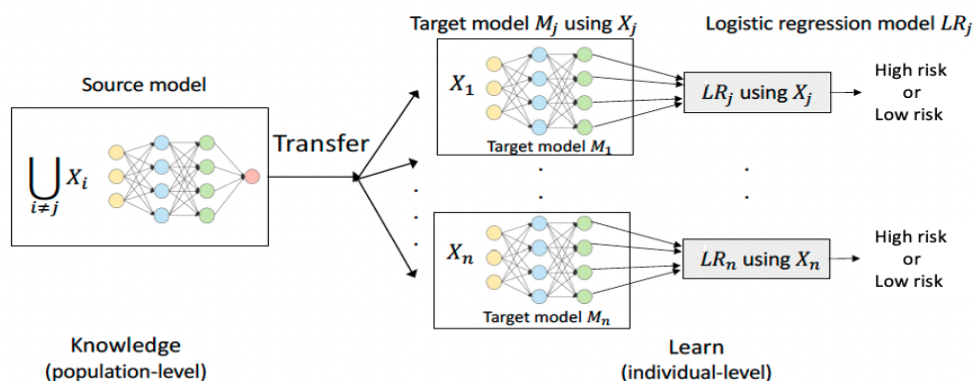


Figure 3. Transfer Learning based Logistic Regression (TL + LR).

2.3.2. Quantile Regression with a Variable Sliding Window Method

Spatio-temporal analysis using QR has known to be one of the successful machine learning techniques for time-series data prediction in business and economics [63]. Recent work in [58] presents a joint model of QR and temporal variability for finding patterns of climate change by taking into consideration the spatio-temporal properties of the data. Asthma risks are known to be associated with an overall increase or decrease in temperature, humidity and other air pollutants in a specifically defined past time period. Many studies consider air quality and other environmental factors in asthma care and management, but the literature is lacking in in-depth analysis of patients’ exposures to these factors in a recent time period.

To improve QR models, we propose a variable sliding window method, where the time window size (the duration of the model construction) and the length of sliding (model usage time) are determined dynamically over time. This method defines two parameters: the window size W of the number of data points (i.e., the number of days) for the model training and validation and the sliding size m^k of the number of days for the current model usage duration at k^{th} iteration, which is also considered as the time period for the next model development. The dataset within a given window consists of a training dataset (D_{train}) and a validation dataset (D_{valid}). Figure 4 illustrates the use of the sliding window method in the QR modeling process.

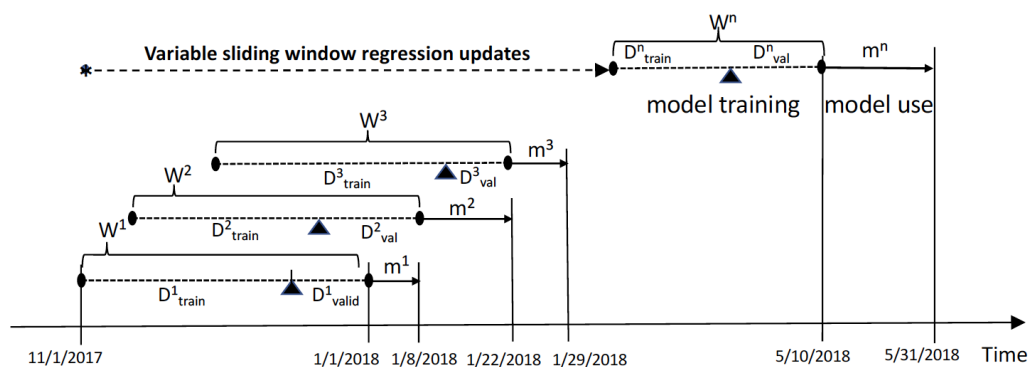


Figure 4. QR with a variable sliding window method.

As an example of the QR modeling, one can build a QR model using the data collected for 45 consecutive days ($W = 45$), 30 days' data points (D_{train}) are used for training the model and 15 days' data points (D_{valid}) are used for validation of the model. Note that using a larger value for W (a larger training/validation dataset) may lose temporal information by the fact that the relationship between health risks and environmental variables is dynamic and typically changes over time. If the quality of the model is acceptable, this model can be used for a certain number of days m to predict the health risk of the patient. The values of m can be determined based on the outcomes of the validation process. With recent advancements in computing hardware and software, updating the model every day is possible, but for utility and practical use, 7 days ($m = 7$) would be a reasonable value for m unless more frequent updates on the model are required to maintain acceptable errors in prediction.

The optimal values of W and m depend on many application-specific factors including the desired model accuracy, specific nature of the given data and available computing resources, and should be searched during the model development phase. In our study, we analyzed $W = 35, 40, 45, 50$, and $m = 30, 20, 10, 5, 1$ to find a good pairing of (W, m) , and the results are presented in Section 3.2.

2.4. A Predictive Modeling Framework and Its Use Case

While various statistical methods exist for evaluating the performance of logistic and quantile regression models [57,64], practical interpretation of their metrics is difficult to convey to medical practitioners and patients. Thus, we propose a new predictive modeling framework that yields understandable information on the model's performance and delivers easy to use the outcomes of predictions.

2.4.1. Training, Validation and Testing

In the QR modeling, training, validation and testing are conducted using a variable sliding window method as described in Section 2.3.2. As m number (moving size) of data are collected and augmented to the dataset, the same number of data points that are the least recently collected are removed from the dataset. On the other hand, the TL + LR modeling uses k -fold cross validation, the most commonly used method for classifiers, for model training, validation and testing. Details of the TL + LR modeling process are below.

Initialization: For each patient, we collect a set of data consisting of the variables listed in Table 1 and integrate them to a dataset. In the improved LR (NN-based TR + LR) classification modeling, each patient's integrated dataset is divided to (a) training/validation data according to an 80%/20% split. We also construct (c) a dataset for the source model by combining all patients' data except the target patient.

Training and Validation: For each patient, we build a source model using the entire dataset (c). For overcoming the class imbalance problem, an oversampling technique is used to generate synthetic data using some samples from dataset (a) and these synthetic data are augmented to dataset (a). We then use k -fold cross validation to build a target model using the augmented data, which is split to k non-overlapping datasets (called as folds): For k rounds of evaluation, $k - 1$ folds are used for training a model and the remaining 1 fold is used for validating the model. In the training/validation phase, we build k models and select the best model by evaluating the models using the standard evaluation metrics. The metrics we used in our study are presented in Section 3.1.1. Model overfitting and underfitting are also tested using learning curves and training loss. The hyperparameters are selected through extended training and k -fold validation processes to avoid over-fitting while to increase the accuracy.

Testing: Once the model is trained and validated, the estimation quality of the model is analyzed through the testing phase on the remaining data (b), which are not used for training. The dataset (b) should keep the same data distribution as the patient's original data. Hence no over-sampling is applied to balance samples among classes. Standard

evaluation metrics for classifiers are used to evaluate the model performance. Averages over all test data represent the quality of the models.

2.4.2. Model Use

In use, the patient’s today’s PEFR value and indoor air quality data values in Table 1 are collected, and the patient’s prediction models are used to estimate the patients’ health risk for tomorrow based on the amount exposure to environmental factors and today’s PEFR value. In our proposed framework, the two methods, TL + LR and QR, collaboratively work to make a risk prediction, influence on parameters and hyper-parameters, and evolve in the life cycle of a prediction framework. Figure 5 illustrates the overview of the proposed predictive modeling framework and a use case of the modeling framework. Steps of the use case are:

Step 1: The framework starts with the development of a TR + LR model through the training and validating process using the patient’s historical data. In model usage, the model predicts the class of the patient’s next-day health risk state in terms of falling below their $PEFR_C$ ($p(PEFR_C) \geq 0.5$).

Step 2: If the model predicts high-risk class, it sends a request to the QR modeling process for prediction for more detailed information. Model parameters and hyper-parameters including τ_C are updated based on the outcomes of the previous step.

Step 3: It uses a QR model to predict the PEFR value $PEFR(\tau_C)$ associated with the critical quantile τ_C .

Note that the model training and validation process by using the QR method is independent from the TL + LR modeling although they both can provide the information for updating parameters and hyper-parameters.

Step 4: If the value of $PEFR(\tau_C)$ estimated in step 3 falls below its previous value and the drop value is larger than a threshold θ , the system outputs a prediction report. Model parameters and hyper-parameters including $PEFR_C$ are updated based on the outcomes of the previous step.

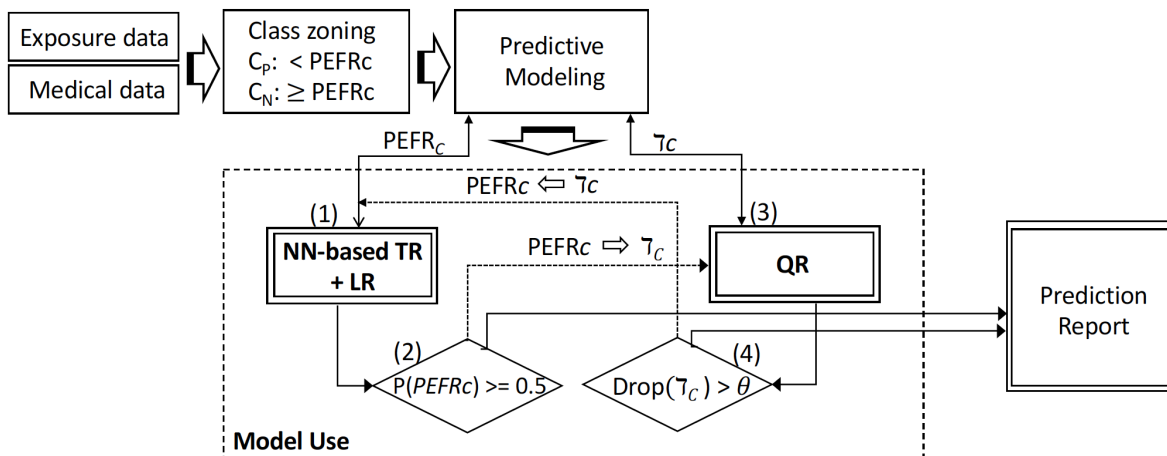


Figure 5. A use case of the predictive modeling framework.

3. Results

The experiments on the quality of the predictive modeling were conducted on the 19 patients’ datasets and Table 2 presents a summary of the datasets.

Table 2. Data distribution.

		Overall (n = 19)			Women (n = 10)	Men (n = 9)
		P25	P50	P75	P50	P50
Data size	per patient (days)	140	188	196	163	203
Age	(years)	56	72	75	65	68
BMI	(Kg/m ²)	23.8	21.9	26.8	23.9	23.6
AM PEFr	(L/min)	313.7	373.3	453.9	350.7	462.3
Daily average exposures (24 h)						
	Temperature (°C)	21.9	22.4	23.7	21.2	22.6
	Relative humidity (%)	37.9	32.7	44.1	40.9	37.3
	PM _{2.5} (µg/m ³)	40.2	35.7	50.6	46.2	35.7
	CO ₂ (ppm)	1005.9	886.9	1241.0	1030.4	918.4

3.1. Performance Evaluation of Classification Models

3.1.1. Evaluation Metrics

The confusion matrix is a commonly used method for evaluating classification models. In a binary confusion matrix, the model performance is evaluated based on the model’s ability to distinguish “positive” data samples from “negative” ones. The confusion matrix is shown in Table 3, where *TP* represents “True Positive”, the number of positive data samples correctly classified as positive, *FN* represents “False Negative”, the number of positive data points incorrectly classified as negative, *FP* represents “False Positive”, the number of negative data points incorrectly classified as positive, and *TN* represents “True Negative”, the number of negative data points correctly classified data as negative. The binary confusion matrix can be generalized to the confusion matrix for multi-class classification. For classification problems with multiple classes, one overall quality metric is arrived at by calculating these numbers for each class independently and averaging the results. In our study, of high-risk prediction, our “positive” samples were the data tuples in which a patient’s PEFr value was below the patient’s critical cutoff (*PEFr_C*) and the class containing these data is called *Class_{Risk}* and the class containing the data above *PEFr_C* is *Class_{noRisk}*.

Table 3. Confusion matrix.

	Predictive <i>class_{Risk}</i>	Predictive <i>class_{noRisk}</i>
Actual <i>class_{Risk}</i>	<i>TP</i>	<i>FN</i>
Actual <i>class_{noRisk}</i>	<i>FP</i>	<i>TN</i>

The following standard metrics that take into account minority classes were used: (1) weighted accuracy = $\frac{TP}{2(TP+FN)} + \frac{TN}{2(TN+FP)}$, (2) sensitivity (also called recall) = $\frac{TP}{TP+FN}$, (3) specificity = $\frac{TN}{FP+FN}$, (4) precision = $\frac{TP}{TP+FP}$, (5) *F*₁-score = $\frac{2*precision*recall}{precision+recall}$, and (6) Receiver Operating Characteristic (ROC), while these metrics are equally important for evaluating classifiers, we emphasize the model’s performance on the target class *Class_{risk}* in the context of risk prediction.

Weighted accuracy is the average of a model’s accuracy rate at classifying positive samples as *Class_{Risk}* and negative samples as *Class_{noRisk}*. Sensitivity is the model’s success rate at classifying positive samples as positive while specificity is the model’s success rate at classifying negative samples as negative. On the other hand, precision measures what percentage of data tuples that the model classifies as positive are actually positive. Typically, precision decreases as recall increases. *F*₁-score is the harmonic mean of sensitivity and precision, which measures the model’s success at both the correct classification of high-risk samples and avoiding the incorrect classification of low-risk samples as high-risk. The area under a ROC curve (denoted as ROC AUC) provides an overall measure of fit of the

model. However, ROC AUC does not account for prevalence or different misclassification costs arising from false-negative and false-positive diagnoses [65]. Change in ROC AUC has little direct clinical meaning for medical practitioners. They proposed an alternative analysis based on the change in sensitivity and specificity at clinically relevant thresholds. This analysis provides full benefits of prediction models by incorporating estimates of prevalence and misclassification costs, and hence it is clinically interpretable since it reflects changes in correct and incorrect risk predictions when a new test is introduced.

3.1.2. Classification Model Performance in Risk Prediction

Our model performance improvement focuses on the metric of sensitivity (correctness of the target high risk zone ($C_p < PEFR_C$) while keeping a good balance in improving all other metrics. Although these model performance metrics assist medical practitioners in integrating them into a care plan, the subtle practical implications of the metrics may be challenging for non machine learning professionals to understand. In fact, metrics such as sensitivity and the F_1 score are sometimes heavily relied upon the model development and on the training and validation phase rather than on the model usage phase.

For external and internal validation of the models, we divided the dataset to training/validation data (80%) and test data (20%). For each patient's model, we conducted 10-fold cross validation for source model training/validation and 5-fold cross validation for target model training/validation. We then evaluated the model on a test data. We describe the models' performance based on the metrics discussed above and present aggregate results and the training loss of the model together with the accuracy.

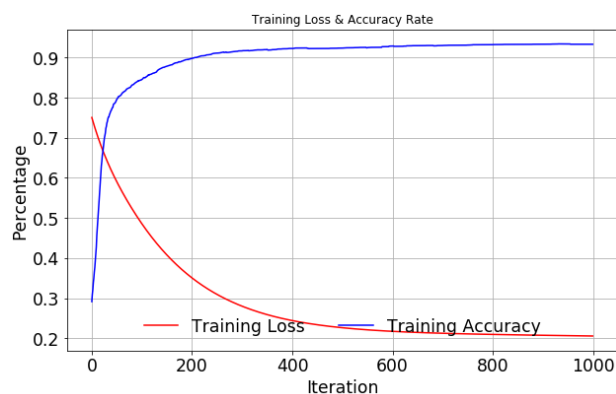
In our experiments, we first applied the synthetic minority oversampling technique (SMOTE) [41] in the training process of the models to overcome the imbalanced class problem. We then implemented a transfer learning paradigm using deep neural networks and LR models. Table 4 summarizes the results of the stand alone LR models and TL-based LR models. The results show the overall performance gains of the TL-based logistic regression models comparing it to that of the stand alone logistic regression, 14.3% in weighted accuracy, 18.4% in sensitivity, 11.5% in specificity, 13.1% in precision, 15.7% in F_1 score, and 18.3% in ROC AUC. In the improved LR models with NN-based TL, the average of sensitivity was 0.727 and the average of specificity was 0.757, while those values are 0.614 and 0.679, respectively, in LR models. This shows that the improved LR models provide a more balanced accuracy between positive class $C_p (<PEFR_C)$ and negative class $C_N (>= PEFR_C)$.

Figure 6 illustrates the loss and accuracy of the source model training and validation and those of target model retraining in the NN based TR + LR models. Figure (a) shows training loss and accuracy and figure (c) shows validating loss and accuracy in the training phases for the source model using 24 patients' datasets (except the target patient SB-078). The loss and accuracy of retraining for the target model for SB-078 are shown in figure (f). Similarly, figures in (b), (d) and (f) show the loss and accuracy of source model training-validation-target model retraining for SB-083.

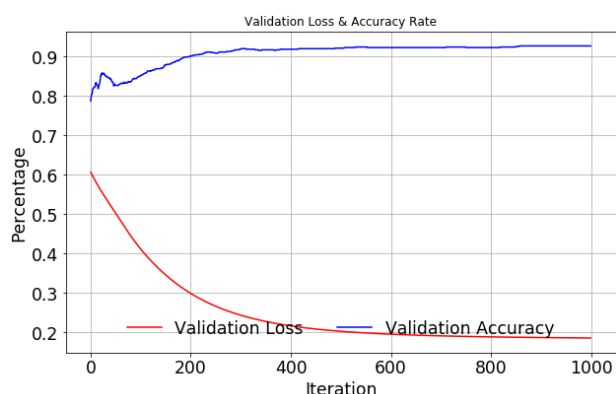
The TL + LR classification models perform with reasonable accuracy rates for use in health risk prediction as compared to the accuracy of commonly used models in health domains. In Table 4, the average sensitivity of logistic regression models was 63% and this was increased to 70% when the transfer learning technique was applied, which resulted in 11% overall improvement rate. At the same time, the average specificity of the models was also improved from 66% to 74%.



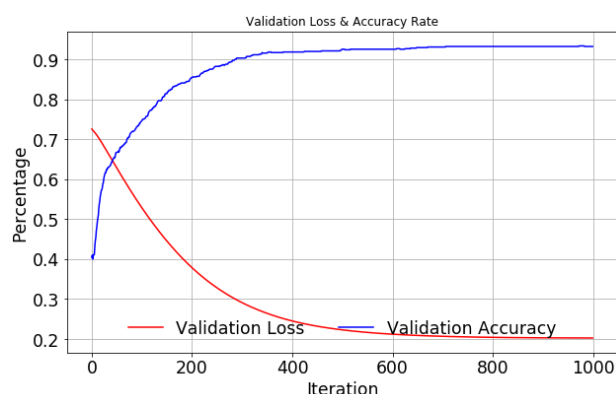
(a) SB-078: source model training



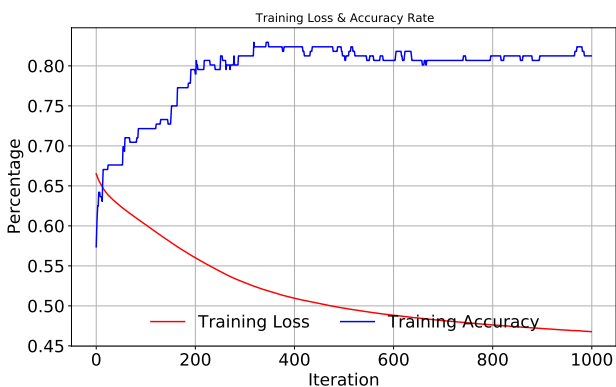
(b) SB-083: Source model training



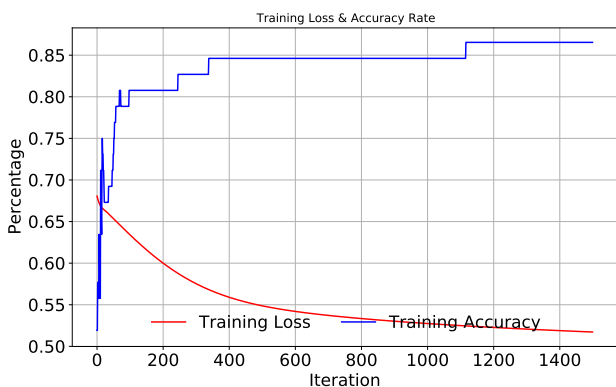
(c) SB-078: Source model validation



(d) SB-083: Source model validation



(e) SB-078: target model retraining



(f) SB-083: target model retraining

Figure 6. Examples of training loss and accuracy rate of NN-based TL in source model training, validation and target model retraining phases.

Table 4. Average model performance of 19 individuals.

Method	Weighted Accuracy	Sensitivity	Specificity	Precision	F ₁ Score	ROC AUC
LR with SMOTE *	0.645	0.614	0.679	0.607	0.596	0.618
NN-based TL + LR with SMOTE *	0.738	0.727	0.757	0.687	0.689	0.741

* SMOTE: the synthetic minority over-sampling technique [41].

Similar improvement in TL + LR was found in other metrics, such as weighted accuracy, precision, and F₁ score. The similar improvement trends in model quality can be seen in the results of 19 individual patients’ models as shown in Figure 7. The performance summary

of TL + LR for 19 individuals are shown in (a) and (b), respectively. The results also show that TL + LR results in tighter bound in the performance measures.

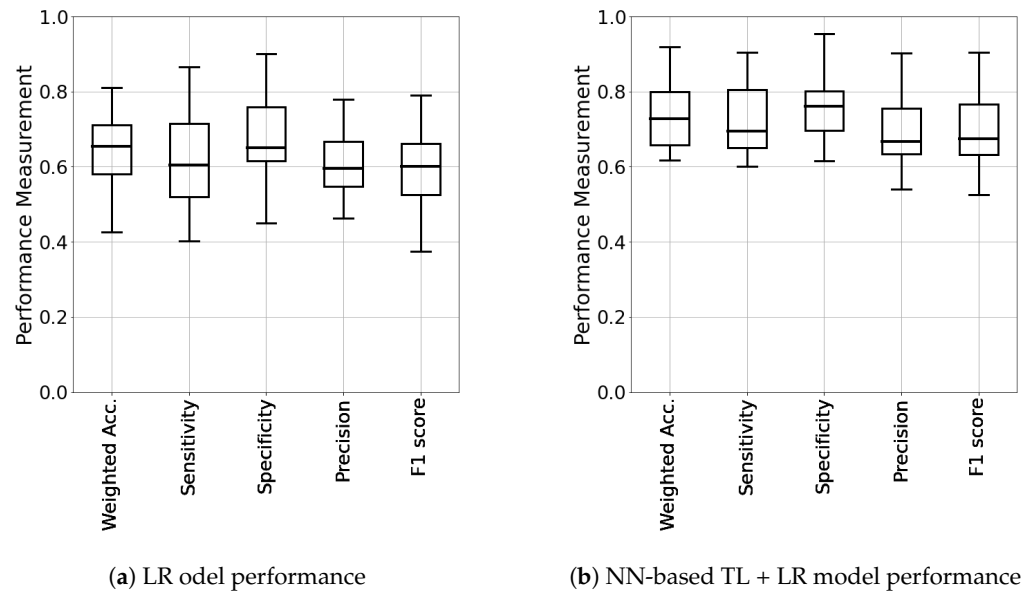


Figure 7. Model performance summary of 19 individuals: LR vs. NN-based TL + LR.

3.2. Performance Evaluation of Quantile Regression Models

To evaluate the quality of the QR model, we used a uniform measure of the relative error for each quantile τ proposed in [66] and evaluated the model through extensive experiments on real patients’ datasets. The uniform measure of the errors is calculated as follows:

$$Err_{\tau} = \left| \frac{N_{\tau}}{N} - \tau \right|, \tag{3}$$

where N_{τ} is the number of data points (days having observed PEFR values) under that day’s predicted τ PEFR quantile value and N is the total number of test days.

Table 5 shows the mean and standard deviation of the values of the relative error, Err_{τ} , for individuals QR analysis with varying the sizes of training window W (T_{train}) with 7 days of the model use time m ($T_{use} = 7$). Our analysis shows that the average relative errors of the 19 patients’ models are very low for all τ values ranging from 0.018 to 0.16 on average. A general trend is that small values of τ and large values of τ result in higher errors. Figure 8 shows the relative error in each of 19 patients’ models with 45 days for training window ($T_{train} = 45$) and 7 days of the model use time ($T_{use} = 7$).

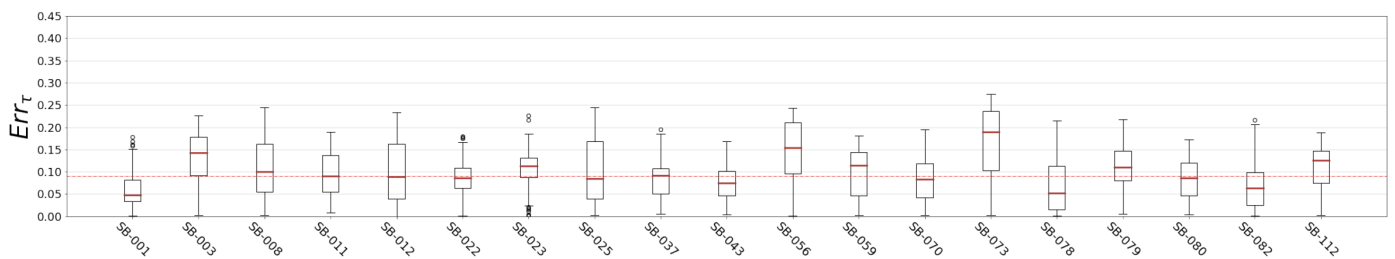


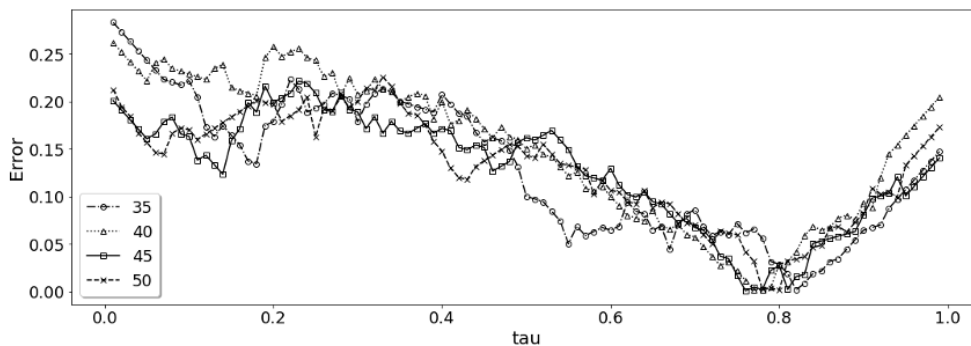
Figure 8. QR relative error analysis of 19 individual models $T_{use} = 7$.

Table 5. Quantile regression analysis with $T_{use} = 7$.

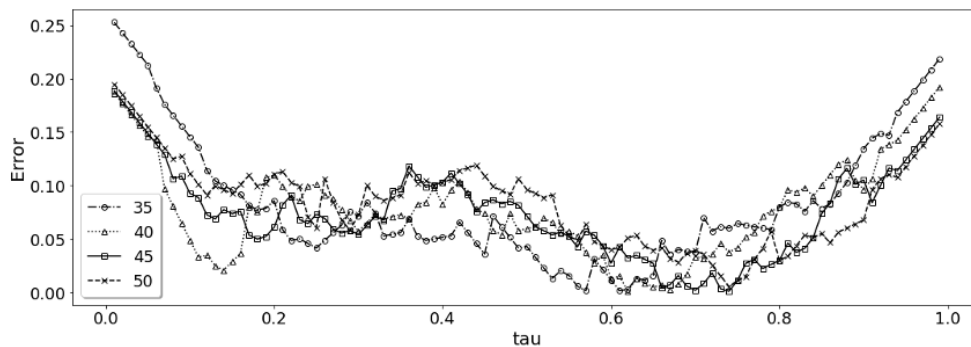
T_{train}		30		35		45		50		Average	
		Err_{τ}^*	std	Err_{τ}	std	Err_{τ}	std	Err_{τ}	std	Err_{τ}	std
Tau(τ)	0.01–0.10	0.092	0.020	0.103	0.016	0.118	0.016	0.037	0.011	0.087	0.016
	0.11–0.20	0.057	0.005	0.112	0.014	0.068	0.018	0.042	0.009	0.070	0.011
	0.21–0.30	0.019	0.020	0.053	0.034	0.021	0.009	0.032	0.006	0.031	0.017
	0.31–0.40	0.030	0.009	0.015	0.010	0.016	0.006	0.012	0.014	0.018	0.010
	0.41–0.50	0.030	0.010	0.009	0.005	0.025	0.015	0.045	0.010	0.027	0.010
	0.51–0.60	0.033	0.015	0.050	0.026	0.048	0.010	0.078	0.021	0.052	0.018
	0.61–0.70	0.046	0.015	0.101	0.006	0.071	0.015	0.055	0.022	0.068	0.014
	0.71–0.80	0.111	0.020	0.105	0.027	0.100	0.012	0.056	0.019	0.093	0.020
	0.81–0.90	0.143	0.020	0.157	0.009	0.152	0.011	0.113	0.023	0.141	0.016
0.91–0.99	0.173	0.026	0.169	0.023	0.145	0.014	0.153	0.009	0.160	0.018	
average		0.0734	0.016	0.0874	0.017	0.0764	0.0126	0.0623	0.0144	0.0747	0.015

* Err_{τ} = a measure of the error for τ , T_{train} = # of days of model training, T_{use} = # of days of model use.

A general trend found is that increasing the sliding window size ($T_{train} = 35, 45, 50$) reduces the errors but the results also show that increasing the training window size T_{train} does not always reduce the errors for some τ values. We also see that the average Err_{τ} of the models using 30 day is little lower than that of the model using 45 day window. This means that a model can be developed for a patient in a relatively shorter period (i.e., in 1 month) and the model can be refined further while the system serves the patient risk management. With an optimal window for the QR model for each individual, we show the average relative errors of each individual’s quantile regression analysis for different window sizes in Figure 9.

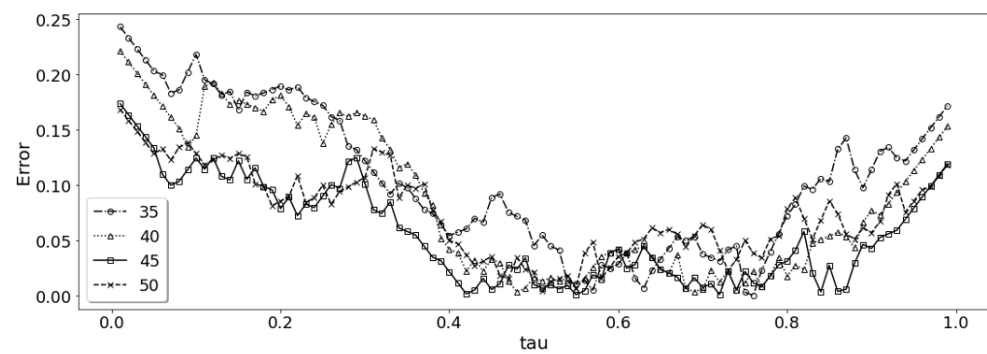


(a) QR relative errors of SB-003

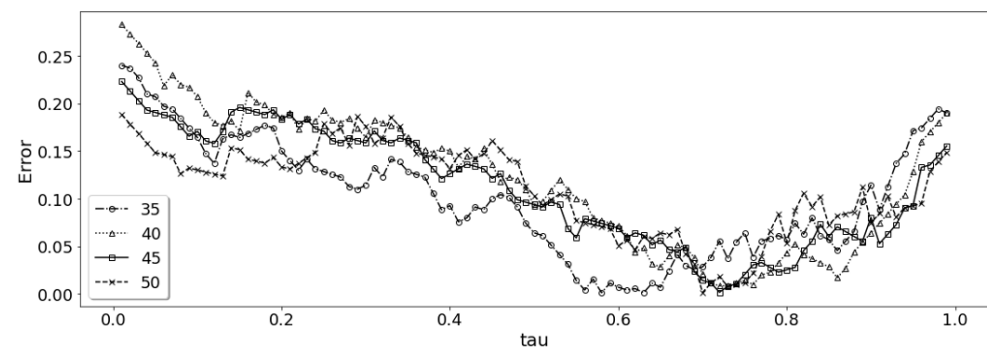


(b) QR relative errors of SB-037

Figure 9. Cont.



(c) QR relative errors of SB-043



(d) QR relative errors of SB-112

Figure 9. QR relative errors in 4 selected individual models.

4. Discussion

The ability to control individual asthma attacks caused by environmental triggers contributes to asthma aggravation reduction, and therefore decreases mortality and treatment cost as well. It is well established that machine learning techniques can contribute significantly to the management of asthma exacerbations and the reduction of its risks but the efforts are still minimal. One major challenge in individual-level health risk modeling is that the performance of commonly used machine learning methods is degraded with the small sized training data, which is frequently found in health applications. Moreover, many of these methods often ignore spatio-temporal properties existing in the data.

Another shortcoming is that doctors and patients often can have significant difficulty understanding the outputs from the models and hence arriving at a practical and useful interpretation of the risk prediction with a probability that is associated with the patient's critical $PEFR_C$ value $PEFR_C$ [49]. Suppose that the system generates a prediction report that the patient's $PEFR$ value will fall below their $PEFR_C$ with a relatively low probability, let us say 20%. A formal meaning of this report is $P(y < PEFR_C) = 0.20$, which can be interpreted to a message that the patient's falling into the risk zone is unlikely so the patient will struggle only slightly with asthma condition. Does this message deliver a sensitive and useful information that can help the patient?

A different challenge in the use of the probability associated with a health risk is to deal with many false alarms. If the system warns the patient with the probability of 20%, then 80% of the warnings are false warnings. Still most people would like to receive a report when the change of having an asthma exacerbation is 20% to avoid hospital admissions or emergency room visits. Therefore, machine learning techniques to automatically explain the results of risk prediction and provide guidance on how to use the outcomes in an asthma care are critical. This opens the possibility of real-time intervention to minimize asthma risk at home.

In this paper, we propose a modeling framework that incorporates two well-known machine learning algorithms to deliver more accurate predictions hence more effective

solutions for progressive, individualized and preventive asthma risk management. As it is one of the major challenges in any individual-level health modeling, the dataset size of each patient is small (mean = 172 days) in our application and needs to be addressed. Training on such a small data set results in relative low accuracy. Our approach is to use a “transfer learning” strategy that incorporates population data as a base model and then refines the model using an individual patient’s data. The results of transfer learning based logistic regression show success of performing transfer learning on all patients’ data for individual’s risk prediction. Our study demonstrates the promise of transfer learning in the development of high quality predictive models based on small dataset.

Author Contributions: Conceptualization, W.D.B., S.A. and M.H.; Data curation, W.D.B., C.-S.P. and S.K.; Formal analysis, W.D.B. and M.H.; Investigation, W.D.B. and S.A.; Methodology, W.D.B. and M.H.; Project administration, W.D.B. and S.K.; Resources, S.A., C.-S.P. and S.K.; Software, J.D. and W.D.B.; Supervision, W.D.B.; Validation, M.H., C.-S.P. and S.K.; Visualization, J.D.; Writing: original draft, W.D.B. and M.H.; Writing: review & editing, W.D.B., S.A., M.H., C.-S.P. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study received support from Seattle University (Grant No. CSE-01-2021), Korea Disease Control and Prevention Agency, South Korea (Grant No. 2016-ER7402-00 and 2017-NE-740200), the Korean Environmental Industry & Technology Institute, Ministry of Environment, South Korea (Grant No. 2016001360002), Soonchunhyang University Brain Korea 21, and Seattle University.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the research ethics committee of the Soonchunhyang University (IRB No. 202001-BR-001-03, 6 January 2022)

Informed Consent Statement: The original study protocols related to our study were approved by the research ethics committee of the Soonchunhyang University (IRB No. 202001-BR-001-03); written informed consent was obtained from all participants.

Data Availability Statement: Data cannot be shared publicly because of personal data protection guideline. Data are available from the Soonchunhyang Risk Assessment Center of Data Access/Ethics Committee (contact via leesr@sch.ac.kr or phil.cjs@gmail.com) for researchers who meet the criteria for access to confidential data.

Acknowledgments: The authors would like to thank the medical staff in the Division of Allergy and Respiratory Medicine at the Soonchunhyang University Bucheon Hospital for the provision of the data. All demographic and biodata were obtained from Soonchunhyang University Bucheon Hospital, a member of the Korea Biobank Network (KBN4-A06).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Purdy, S.; Griffin, T.; Salisbury, C.; Sharp, D. Ambulatory care sensitive conditions: Terminology and disease coding need to be more specific to aid policy makers and clinicians. *Public Health* **2009**, *123*, 169–173. [[CrossRef](#)] [[PubMed](#)]
2. Loftus, P.A.; Wise, S.K. Epidemiology and economic burden of asthma. *Int. Forum Allergy Rhinol.* **2015**, *5*, S7–S10. [[CrossRef](#)] [[PubMed](#)]
3. Johnson, N.B.; Hayes, L.D.; Brown, K.; Hoo, E.C.; Ethier, K.A. CDC National Health Report: Leading causes of morbidity and mortality and associated behavioral risk and protective factors—United States, 2005–2013. *MMWR Suppl.* **2014**, *63*, 3–27.
4. Nunes, C.; Pereira, A.M.; Morais-Almeida, M. Asthma costs and social impact. *Asthma Res. and Pract.* **2017**, *3*, 1. [[CrossRef](#)] [[PubMed](#)]
5. Kelly, C.S.; Morrow, A.L.; Shults, J.; Nakas, N.; Strope, G.L.; Adelman, R.D. Outcomes evaluation of a comprehensive intervention program for asthmatic children enrolled in Medicaid. *Pediatrics* **2000**, *105*, 1029–1035. [[CrossRef](#)] [[PubMed](#)]
6. Forno, E.; Celedón, J.C. Predicting asthma exacerbations in children. *Curr. Opin. Pulm. Med.* **2012**, *18*, 63. [[CrossRef](#)]
7. Hermann, M.; Pentek, T.; Otto, B. Design principles for industrie 4.0 scenarios. In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA, 5–8 January 2016; pp. 3928–3937.
8. Thuemmler, C.; Bai, C. *Health 4.0: How Virtualization and Big Data Are Revolutionizing Healthcare*; Springer: Berlin/Heidelberg, Germany, 2017.
9. Adams, J.; Dorr, D.A.; Leung, M.; Popescu, B.; Rich, J. *Predicting the Financial Risks of Seriously Ill Patients*; CHCF: Oakland, CA, USA, 2011.

10. Greineder, D.K.; Loane, K.C.; Parks, P. A randomized controlled trial of a pediatric asthma outreach program. *J. Allergy Clin. Immunol.* **1999**, *103*, 436–440. [[CrossRef](#)]
11. Dorr, D.A.; Wilcox, A.B.; Brunner, C.P.; Burdon, R.E.; Donnelly, S.M. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. *J. Am. Geriatr. Soc.* **2008**, *56*, 2195–2202. [[CrossRef](#)]
12. Ainsworth, V.J. A Disease Management Program Utilizing “Life Coaches” for Children with Asthma. *JCOM* **2001**, *8*, S7–S10.
13. Asher, M.I.; Ellwood, P. *The Global Asthma Report 2014*; Global Asthma Network: Auckland, New Zealand, 2014.
14. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J. P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [[CrossRef](#)]
15. Morawska, L.; Salthammer, T. *Fundamentals of Indoor Particles and Settled Dust*; Wiley Online Library: Hoboken, NJ, USA, 2003
16. Wild, C.P. The exposome: From concept to utility. *Int. J. Epidemiol.* **2012**, *41*, 24–32. [[CrossRef](#)]
17. Adler, T.; Sawyer, K.; Shelton-Davenport, M. *The Exposome: A Powerful Approach for Evaluating Environmental Exposures and Their Influences on Human Disease*. ESEH-Committee Newsletter; The National Academies Press: Washington, DC, USA, 2010.
18. Schima, R.; Goblirsch, T.; Salbach, C.; Franczyk, B.; Aleithe, M.; Bumberger, J.; Dietrich, P. Research in Progress: Implementation of an Integrated Data Model for an Improved Monitoring of Environmental Processes. In *Business Information Systems Workshops; Lecture Notes in Business Information Processing*; Springer: Cham, Switzerland, 2017; pp. 332–339. [[CrossRef](#)]
19. Jie, Y.; Ismail, N.H.; Isa, Z.M.; others. Do indoor environments influence asthma and asthma-related symptoms among adults in homes? A review of the literature. *J. Formos. Med. Assoc.* **2011**, *110*, 555–563. [[CrossRef](#)] [[PubMed](#)]
20. Matsui, E.C.; Abramson, S.L.; Sandel, M.T.; others. Indoor environmental control practices and asthma management. *Pediatrics* **2016**, *138*, e20162589. [[CrossRef](#)] [[PubMed](#)]
21. Richardson, G.; Eick, S.; Jones, R. How is the indoor environment related to asthma?: Literature review. *J. Adv. Nurs.* **2005**, *52*, 328–339. [[CrossRef](#)]
22. Brooks, B.O.; Utter, G.M.; DeBroy, J.A.; Schimke, R.D. Indoor air pollution: An edifice complex. *J. Toxicol. Clin. Toxicol.* **1991**, *29*, 315–374. [[CrossRef](#)] [[PubMed](#)]
23. Simoni, M.; Jaakkola, M.; Carrozzi, L.; Baldacci, S.; Di Pede, F.; Viegi, G. Indoor air pollution and respiratory health in the elderly. *Eur. Respir. J.* **2003**, *21*, 15s–20s. [[CrossRef](#)] [[PubMed](#)]
24. Bentayeb, M.; Simoni, M.; Norback, D.; Baldacci, S.; Maio, S.; Viegi, G.; Annesi-Maesano, I. Indoor air pollution and respiratory health in the elderly. *J. Environ. Sci. Health Part A* **2013**, *48*, 1783–1789. [[CrossRef](#)] [[PubMed](#)]
25. Asthana, S.; Strong, R.; Megahed, A. HealthAdvisor: Recommendation System for Wearable Technologies enabling Proactive Health Monitoring. *arXiv* **2016**, arXiv:1612.00800.
26. Kothamasu, R.; Huang, S.H.; VerDuin, W.H. System health monitoring and prognostics—A review of current paradigms and practices. *Int. J. Adv. Manuf. Technol.* **2006**, *28*, 1012–1024. [[CrossRef](#)]
27. Gay, V.; Leijdekkers, P. A health monitoring system using smart phones and wearable sensors. *Int. J. Adv. Manuf. Technol.* **2007**, *8*, 29–35.
28. Shen, Y.; Giurgiutiu, V. Predictive modeling of nonlinear wave propagation for structural health monitoring with piezoelectric wafer active sensors. *Int. J. Adv. Manuf. Technol.* **2014**, *25*, 506–520. [[CrossRef](#)]
29. Lv, Y.; Zhao, H.; Liang, Z.; Dong, H.; Liu, L.; Zhang, D.; Cai, S. A Mobile Phone Short Message Service Improves Perceived Control of Asthma: A Randomized Controlled Trial. *Telemed. E-Health* **2012**, *18*, 420–426. [[CrossRef](#)] [[PubMed](#)]
30. Liciskai, C.J.; Sands, T.W.; Ferrone, M. Development and pilot testing of a mobile health solution for asthma self-management: Asthma action plan smartphone application pilot study. *Can. Respir. J.* **2013**, *20*, 301–306. [[CrossRef](#)] [[PubMed](#)]
31. Wu, A.C. The Promise of Improving Asthma Control Using Mobile Health. *J. Allergy Clin. Immunol. Pract.* **2016**, *4*, 738–739. [[CrossRef](#)] [[PubMed](#)]
32. Finkelstein, J.; others. Machine learning approaches to personalize early prediction of asthma exacerbations. *Ann. N. Y. Acad. Sci.* **2017**, *1387*, 153–165. [[CrossRef](#)]
33. Jalali, L.; Dao, M.S.; Jain, R.; Zetts, K. Complex asthma risk factor recognition from heterogeneous data streams. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops, Turin, Italy, 29 June–3 July 2015; pp. 1–6.
34. Luo, G.; Sward, K. A roadmap for optimizing asthma care management via computational approaches. *JMIR Med. Inform.* **2017**, *5*, e32. [[CrossRef](#)]
35. Yang, Q.; Wu, X. 10 challenging problems in data mining research. *Int. J. Inf. Technol. Decis. Mak.* **2006**, *5*, 597–604. [[CrossRef](#)]
36. Jee, K.; Kim, G.H. Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthc. Inform. Res.* **2013**, *19*, 79–85. [[CrossRef](#)]
37. Wills, M.J. Decisions through data: Analytics in healthcare. *J. Healthc. Manag.* **2014**, *59*, 254–262. [[CrossRef](#)]
38. Raghupathi, W.; Raghupathi, V. An overview of health analytics. *J. Health Med. Informat.* **2013**, *4*, 2. [[CrossRef](#)]
39. Belle, A.; Thiagarajan, R.; Soroushmehr, S.; Navidi, F.; Beard, D.A.; Najarian, K. Big data analytics in healthcare. *BioMed Res. Int.* **2015**, *2015*, 370194. [[CrossRef](#)] [[PubMed](#)]
40. Chen, R.; Su, H.; Khalilia, M.; Lin, S.; Peng, Y.; Davis, T.; Hirsh, D.A.; Searles, E.; Tejedor-Sojo, J.; Thompson, M.; et al. Cloud-based predictive modeling system and its application to asthma readmission prediction. *AMIA Annu. Symp. Proc.* **2015**, *2015*, 406. [[PubMed](#)]

41. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
42. Hoens, T.R.; Chawla, N.V. Imbalanced datasets: From sampling to classifiers. In *Imbalanced Learning: Foundations, Algorithms, and Applications*; Wiley: Hoboken, NJ, USA, 2013.
43. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [[CrossRef](#)]
44. Wan, Q.; Deng, X.; Li, M.; Yang, H. SDDSMOTE: Synthetic Minority Oversampling Technique based on Sample Density Distribution for Enhanced Classification on Imbalanced Microarray Data. In Proceedings of the 6th International Conference on Compute and Data Analysis, Shanghai, China, 25–27 February 2022; pp. 35–42.
45. Miller, G.W.; Jones, D.P. The Nature of Nurture: Refining the Definition of the Exposome. *Toxicol. Sci.* **2013**, *137*, 1–2. [[CrossRef](#)]
46. Betts, K.S. Characterizing Exposomes: Tools for Measuring Personal Environmental Exposures. *Environ. Health Perspect.* **2012**, *120*. [[CrossRef](#)]
47. Belisario, J.S.M.; Huckvale, K.; Greenfield, G.; Car, J.; Gunn, L.H. Smartphone and tablet self management apps for asthma. *Cochrane Database Syst. Rev.* **2013** *2013*, CD010013. [[CrossRef](#)]
48. Wu, A.C.; Carpenter, J.F.; Himes, B.E. Mobile health applications for asthma. *J. Allergy Clin. Immunol. Pract.* **2015**, *3*, 446–448.e16. [[CrossRef](#)]
49. Bellazzi, R.; Zupan, B. Predictive data mining in clinical medicine: Current issues and guidelines. *Int. J. Med. Inform.* **2008**, *77*, 81–97. [[CrossRef](#)]
50. Clement Clarke International, Predictive Normal Values (Nomogram, EU scale). Available online: http://www.peakflow.com/top_nav/normal_values/index.html (accessed on 19 August 2018).
51. American Lung Association. Measuring Your Peak Flow Rate. Available online: <https://www.lungusa.org> (accessed on 3 October 2022).
52. Alkobaisi, S.; Bae, W.D.; Horak, M.; Narayanappa, S.; Lee, J.; AbuKhoua, E.; Park, C.S.; Bae, D.J. Predictive and exposome analytics: A case study of asthma exacerbation management. *J. Ambient. Intell. Smart Environ.* **2019**, 1–26. [[CrossRef](#)]
53. Woo, J.; Rudasingwa, G.; Kim, S. Assessment of Daily Personal PM_{2.5} Exposure Level According to Four Major Activities among Children. *Appl. Sci.* **2020**, *10*, 159. [[CrossRef](#)]
54. Vineis, P.; Chadeau-Hyam, M.; Gmuender, H.; Gulliver, J.; Herceg, Z.; Kleinjans, J.; Kogevinas, M.; Kyrtopoulos, S.; Nieuwenhuijsen, M.; Phillips, D.H.; et al. The exposome in practice: Design of the EXPOsOMICS project. *Int. J. Hyg. Environ. Health* **2017**, *220*, 142–151. [[CrossRef](#)] [[PubMed](#)]
55. Freedman, D.A. *Statistical Models: Theory and Practice*; Cambridge University Press: Cambridge, UK, 2009.
56. Koenker, R.; Hallock, K.F. Quantile regression. *J. Econ. Perspect.* **2001**, *15*, 143–156. [[CrossRef](#)]
57. He, X.; Zhu, L.X. A lack-of-fit test for quantile regression. *J. Am. Stat. Assoc.* **2003**, *98*, 1013–1022. [[CrossRef](#)]
58. Duan, Q.; McGrory, C.A.; Brown, G.; Mengersen, K.; Wang, Y.G. Spatio-temporal quantile regression analysis revealing more nuanced patterns of climate change: A study of long-term daily temperature in Australia. *arXiv* **2021**, arXiv:2103.05791.
59. Battineni, G.; Sagaro, G.G.; Chinatalapudi, N.; Amenta, F. Applications of machine learning predictive models in the chronic disease diagnosis. *J. Pers. Med.* **2020**, *10*, 21. [[CrossRef](#)]
60. Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global: Hershey, PA, USA, 2010; pp. 242–264.
61. Liao, X.; Qian, Y.; Chen, Y.; Xiong, X.; Wang, Q.; Heng, P.A. MMTLNet: Multi-Modality Transfer Learning Network with adversarial training for 3D whole heart segmentation. *Comput. Med. Imaging Graph.* **2020**, *85*, 101785. [[CrossRef](#)]
62. Bae, W.D.; Kim, S.; Park, C.S.; Alkobaisi, S.; Lee, J.; Seo, W.; Park, J.S.; Park, S.; Lee, S.; Lee, J.W. Performance improvement of machine learning techniques predicting the association of exacerbation of peak expiratory flow ratio with short term exposure level to indoor air quality using adult asthmatics clustered data. *PLoS ONE* **2021**, *16*, e0244233. [[CrossRef](#)]
63. Kihm, A.; Ritter, N.; Vance, C. Is the German retail gasoline market competitive? A spatial-temporal analysis using quantile regression. *Land Econom.* **2016**, *92*, 718–736. [[CrossRef](#)]
64. Koenker, R.; Machado, J.A. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **1999**, *94*, 1296–1310. [[CrossRef](#)]
65. Halligan, S.; Altman, D.G.; Mallett, S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *Eur. Radiol.* **2015**, *25*, 932–939. [[CrossRef](#)] [[PubMed](#)]
66. Bae, W.D.; Horak, M.; Alkobaisi, S.; Kim, S.; Narayanappa, S.; Park, C.S.; Bae, D.J. A two-step approach to predictive modeling of individual-based environmental health risks. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, 8–12 April 2019; pp. 729–738.