From code to clinic: theory and practice for artificial intelligence prediction algorithms

Hond, A.A.H. de

5

Machine learning did not beat logistic regression in time series prediction for severe asthma exacerbations

Anne A. H. de Hond, Ilse M. J. Kant, Persijn J. Honkoop, Andrew D. Smith, Ewout W. Steyerberg, and Jacob K. Sont

5.1 ABSTRACT

Early detection of severe asthma exacerbations through home monitoring data in patients with stable mild-to-moderate chronic asthma could help to timely adjust medication. We evaluated the potential of machine learning methods compared to a clinical rule and logistic regression to predict severe exacerbations.

We used daily home monitoring data from two studies in asthma patients (development: n=165 and validation: n=101 patients). Two ML models (XGBoost, one class SVM) and a logistic regression model provided predictions based on peak expiratory flow and asthma symptoms. These models were compared with an asthma action plan rule.

Severe exacerbations occurred in 0.2% of all daily measurements in the development (154/92,787 days) and validation cohorts (94/40,185 days). The AUC of the best performing XGBoost was 0.85 (0.82-0.87) and 0.88 (0.86-0.90) for logistic regression in the validation cohort. The XGBoost model provided overly extreme risk estimates, whereas the logistic regression underestimated predicted risks. Sensitivity and specificity were better overall for XGBoost and logistic regression compared to one class SVM and the clinical rule.

We conclude that ML models did not beat logistic regression in predicting short-term severe asthma exacerbations based on home monitoring data. Clinical application remains challenging in settings with low event incidence and high false alarm rates with high sensitivity.

5.2 BACKGROUND

The collection of home monitoring data via mobile applications, online surveys and wearables is becoming increasingly popular to remotely monitor patients. Monitoring has the potential to aid in detecting clinical deterioration earlier, which is associated with better clinical outcomes [1]. For many applications, simple clinical rules have been developed to predict short-term events such as severe clinical deterioration [2-5].

The advent of machine learning (ML) means we can develop highly flexible models with the ability to automatically learn from data, capture complex patterns, and incorporate time-series trends. ML models might overtake some of the moderately effective clinical rules [2-5]. ML has shown great results in application areas such as image recognition [6-8]. Its utility for home monitoring time-series data remains to be determined. Home monitoring time series data present a distinctive set of challenges for the application of ML predictive algorithms. A large effective sample size is important [9, 10], which is challenging with a low incidence of the outcome of interest. For example, severe asthma exacerbations occur in less than 0.5% of days. All the other days are normal asthma control days [9, 11]. Moreover, fair external validation of ML predictive algorithms on a truly independent data is rare, commonly leading to an overoptimistic impression of predictive performance [12, 13]. Due to these challenges, only few models have been developed for home monitoring data [14], and even fewer have been externally validated.

We aim to develop and validate prediction models for short-term prediction of severe asthma exacerbations in patients with stable mild-to-moderate chronic asthma based on home monitoring data. We compare the performance of two machine learning algorithms, a statistical model, and a simple asthma action plan rule [5].

5.3 METHODS

5.3.1 Development and validation cohorts

We analyzed two previous studies which had as the primary aim to study adjustments in asthma treatment [15, 16]. The development cohort was a random-

ized controlled trial comparing different inhaler medications with follow up of approximately 84 weeks [16]. The validation cohort was a single-blind placebo-controlled trial examining alternative treatment pathways with follow up of approximately 60 weeks [17]. All patients had stable mild-to-moderate chronic asthma. Both studies were conducted in an asthma clinic in New Zealand on patients referred by their general practitioners. For both studies, patients recorded their peak expiratory flow and use of β2-reliever (yes/no) in the morning and evening of every trial day in diaries. Nocturnal awakening (yes/no) was recorded in the morning (see below).

5.3.2 Outcome

The outcome variable was measured daily and was defined as the occurrence of a severe asthma exacerbation within 2 days (the day of the measurement or the following day). Table 5.1 provides a visualization of this 2-day window outcome. Severe asthma exacerbations were defined as the need for a course of oral corticosteroids (prednisone) for a minimum of three days, as documented in medical records [15, 16] .

Table 5.1 Definition of the outcome variable

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Exacerbation															
2-day window															
4-day window															
8-day window															

This is a hypothetical example of the definition of the outcome variable over 15 days of measurement. The patient experiences an exacerbation at day 9 and day 15. The outcome variable corresponding to a severe asthma exacerbation within 2 days is displayed on the 2-day window row. For example, at day 8 an exacerbation will occur within 2 days – it occurs the next day – and day 8 is therefore part of the 2-day window outcome. Similarly, the outcome variable definitions corresponding to exacerbations within 4 and 8 days are displayed on the 4- and 8-day window rows.

5.3.3 Predictors

All predictors were measured or calculated daily. Nocturnal awakening (yes/no), the average of morning and evening peak expiratory flow (PEF, measured in liters per minute) and the use of β2-reliever in morning and evening (used in both morning and evening/used in morning or evening/not used in morning

and evening) were considered as potential predictors. For a rolling window of 7 days, we also calculated the PEF average, standard deviation, maximum and minimum and added these as predictors. This rolling window consisted of the current day and all 6 preceding days. The PEF personal best was determined per patient during a run-in period of four weeks and added to the models. Lastly, we constructed and added first differences (the difference in today's measurement with respect to yesterday's measurement) and lags (yesterday's measurement) for PEF, nocturnal awakening, and use of β2-reliever.

5.3.4 Model development

Demographics and descriptive statistics of predictors (i.e., age, sex, mean PEF, PEF % personal best, nocturnal awakening, and use of β2-reliever) were calculated for each individual patient over their respective observational periods.

Missing values were interpolated based on previous and succeeding values and the data was normalized. The first ML model developed through supervised learning was a gradient boosted decision trees (XGBoost) model. This model was chosen as it is one of the most popular ML techniques, and it performs well for a wide selection of problems, including time series prediction [18]. The XGBoost model estimates many decision-trees sequentially. This is also called boosting. These decision tree predictions are combined into an ensemble model to arrive at the final predictions. The sequential training makes the XGBoost model faster and more efficient than other tree-based algorithms, such as random forest. A downside of this model is that, due to its complexity, it becomes hard to interpret. Moreover, when the missingness is high, tuning an XGBoost model may become increasingly difficult, which is less of an issue with other tree-based models like random forest.

Second, we trained an outlier detection model (one class SVM with Radial Basis Kernel)[19]. The one class SVM aims to find a frontier that delimits the contours of the original distribution. By estimating this frontier, it can identify whether a new data point falls outside of the original distribution and should therefore be classified as 'irregular'. An advantage of this model is that it is particularly apt at dealing with the low event rate in the asthma data. A downside of this model is that it does not provide probability estimates like a regular support

vector machine and we therefore must base its predictive performance on its classification metrics only (see below).

Additionally, we developed a prediction model using logistic regression as the popular classical prediction counterpart of these two ML models. Logistic regression assumes a probability distribution for the outcome variable and models the log-odds of each patient experiencing the outcome linearly. The log-odds are converted into probabilities via the logistic function. Logistic regression is an inherently interpretable technique and a hallmark of classical prediction modelling [20, 21]. Due to its linearity restriction, it may however not provide the level of complexity needed to adequately model certain prediction problems. Machine learning methods, like XGBoost and one class SVM, provide more flexibility, which comes at a cost of the interpretability of these methods.

The hyperparameters of the XGBoost, one class SVM, and logistic regression models (see supplementary Table S1) were set using a full grid search and 5x5-fold cross-validation (stratified by patient) on the development cohort. We trained the final models using all data with optimized hyperparameters. We compared these model outcomes with a clinical rule that is currently proposed as action point in an asthma action plan by the British Thoracic Society: start oral corticosteroids treatment if PEF < 60% of personal best [2, 5].

5.3.5 Model performance

After completing model development on the development cohort, all models and the clinical rule were applied to the validation cohort. The discriminative performance of the models producing probabilities (XGBoost and logistic regression) was measured via the area under the receiver operating characteristic curve (AUC) and histograms of the probability distributions were plotted. We applied the DeLong test to compare the AUCs from these two models. Calibration was assessed graphically and quantified through the calibration slope and intercept [22]. Confidence intervals were obtained through bootstrapping (based on a 1000 iterations). Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for all models at the following probability thresholds (the cut-off point at which probabilities are converted into binary outcomes): 0.1% and 0.2%. These were chosen as they circle the prevalence rate of the outcome in our data. For a fair comparison

with the clinical rule, we also calculated these performance metrics (sensitivity, specificity, etc.) for the XGBoost and logistic regression models at the probability thresholds producing the same number of positive predictions as produced by the one class SVM and the clinical rule.

5.3.6 Sensitivity analysis

We performed a sensitivity analysis for predicting exacerbations within 4 and 8 days as opposed to 2 days (Table 5.1). This enabled us to study the effect of a variation in the length of the outcome window on the models' discrimination and calibration capacities.

Second, we performed a sensitivity analysis to assess the effect of the number of lags on model performance. For this analysis, we varied the number of lags from 1 to 5 for the models predicting exacerbations within 2 days. For the XGBoost and logistic regression model, the AUC was compared. For the one class SVM model, the sensitivity, specificity, PPV, and NPV were compared.

5.3.7 Software

All analyses were performed in Python 3.8.0. with R 3.6.3 plug-ins to obtain calibration results. The complete code is available on request.

5.4 RESULTS

The development and validation cohorts consisted of 165 and 101 asthma patients respectively (Table 5.2). Patients were followed for a median period of 610 days in the development and 417 days in the validation cohort. Among the development data patients, 49 had one or more exacerbations (30%). This amounted to a total of 154 exacerbations across all patients (0.2% of total 92,787 daily measurements). For the validation data this was 38 patients (38%) and a total of 94 exacerbations (also 0.2% of total 40,185 daily measurements). The percentage of missing daily measurements was below 1% for the development and below 5% for the validation cohort for all candidate predictors (Table 5.2). Figure 5.1 provides an illustration of the time series for PEF, nocturnal awakening, and use of β2-reliever for three representative patients with various degrees of asthma exacerbations.

Table 5.2 Descriptive statistics of the development and validation cohorts

	Development cohort	Validation cohort
Demographics		
Patient, N	165	101
Total daily measurements, N	92787	40185
Observational period, median (25-75)	610 (580-640)	417 (376-473)
Age, median (25-75)	38 (28-47)	46.5 (34-56)
Sex (female), N (%)	92 (56%)	62 (61%)
Predictors		
Peak expiratory flow, mean (std)	438 (98)	404 (104)
Missing (%)	477 (0.5%)	1171 (2.9%)
Peak expiratory flow personal best*, mean (std)	467 (100)	437 (103)
Nocturnal awakening, mean % per patient patient	6.3%	4.7%
Missing (%)	876 (0.9%)	1665 (4.1%)
Use of β2 reliever, mean % per patient	7.2%	8.9%
Missing (%)	302 (0.3%)	1188 (3.0%)
Outcome		
Exacerbations per patient, N (%)		
0 exacerbations	116 (70%)	63 (62%)
1 exacerbation	25 (15%)	20 (20%)
2 or more exacerbations	24 (15%)	18 (18%)
Total exacerbations, N (%)	154 (0.2%)	94 (0.2%)

Statistics were calculated for each individual patient over their respective observational periods. Then these statistics were pooled across patients.
* No % missing is reported for maximum peak expiratory flow as this is a summary statistic calculated per patient over a run-in period of 4 weeks.

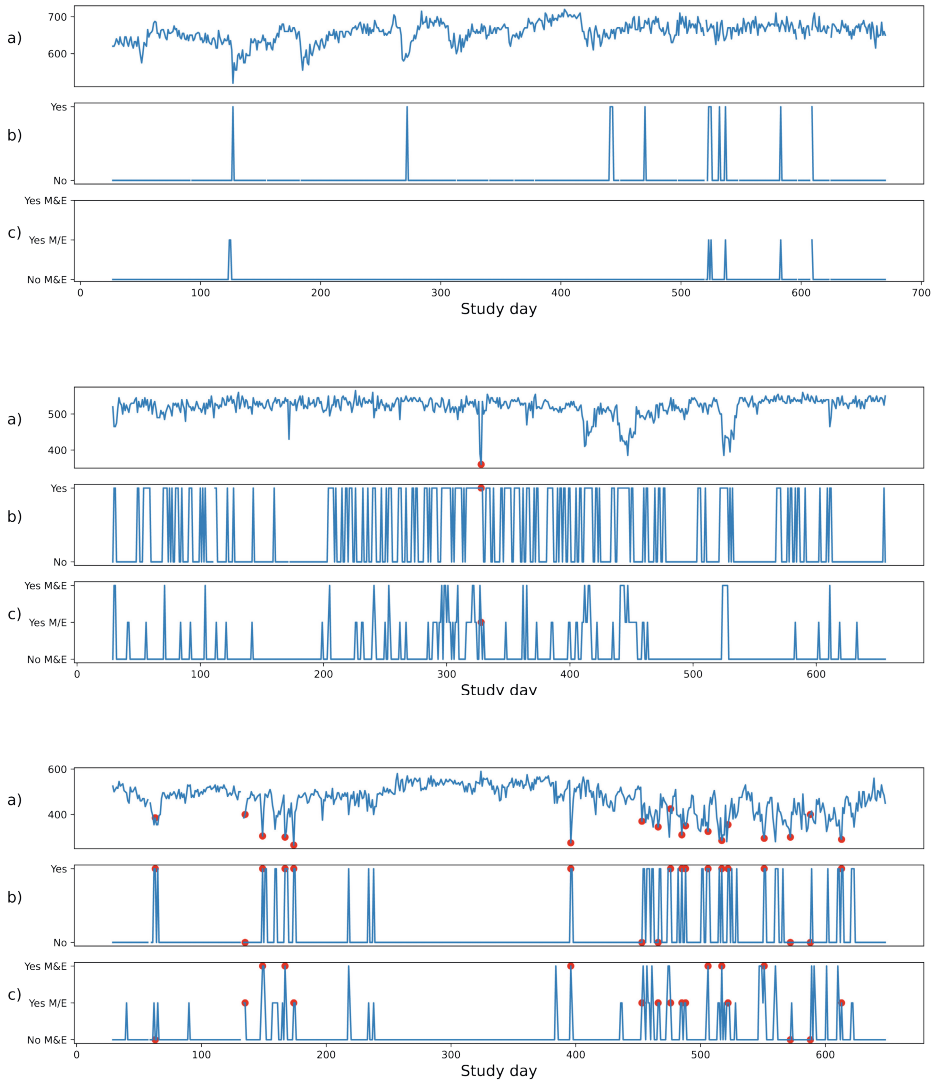

Figure 5.1 Time series for patients with no, one and many exacerbations

a) Peak expiratory flow, b) nocturnal awakening (yes/no), and c) use of β2 reliever (No M&E = No Morning & Evening, Yes M/E = Yes morning or evening, Yes M&E = Yes morning and evening) over time for three patients with no, one and many exacerbations respectively. The case of no exacerbations (top figure) is most prevalent in the data. Exacerbations are marked with red dots.

XGBoost included PEF, nocturnal awakening, and use of β2-reliever and their corresponding statistics as predictors with first differences and first lags. At validation, the algorithm obtained an AUC of 0.81 (95% CI 0.78-0.84, Table 5.3, Figure 5.2). The logistic regression model had a higher validated AUC of 0.88 (95% CI 0.86-0.90, p=0.00, DeLong test). The probability distributions of the two models were heavily skewed (supplementary Figure S8). Poor calibration with too extreme risk estimates was noted for the XGBoost model (calibration slope 0.56, 95% CI 0.50-0.61, Table 5.3, supplementary Figure S9). It also underestimated the risks (calibration intercept 0.32 (95% CI 0.15-0.48). Near perfect calibration was found for the logistic regression model (slope 1.02, 95% CI 0.93-1.10, Table 5.3, supplementary Figure S9), with some underestimation of the risk of exacerbations (intercept 0.75, 95% CI 0.60-0.90).

Table 5.3 Discrimination and calibration for predicting exacerbation within 2 days (validation cohort)

	AUC	Calibration intercept	Calibration slope
XGBoost	0.81 (0.78, 0.84)	0.32 (0.15, 0.48)	0.56 (0.5, 0.61)
Logistic regression	0.88 (0.86, 0.90)	0.75 (0.6, 0.90)	1.02 (0.93, 1.10)

Abbreviations: *XGBoost* gradient boosted decision trees, *AUC* Area Under the Receiver Operating Characteristics Curve

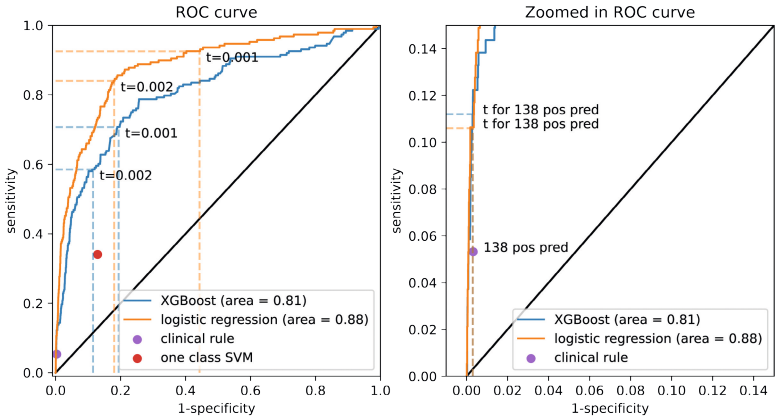

Figure 5.2 ROC-curve for predictions from XGBoost and the logistic regression model

The sensitivity and specificity of the one class SVM an clinical prediction rule are also plotted on the left curve. On the left the points corresponding to the 0.001 ('t=0.001') and 0.002 ('t=0.002') probability thresholds are plotted for the XGBoost and logistic regression model. On the right the points corresponding to the thresholds resulting in 138 positive predictions ('t for 138 pos pred', equaling the clinical rule positive predictions) are plotted for the XGBoost and logistic regression model.

For the 0.2% threshold, the XGBoost model obtained a sensitivity of 0.59, a specificity of 0.89, a positive predictive value (PPV) of 0.02, and a negative predictive value (NPV) of 1 (Table 5.4). For the logistic regression model, this was 0.84, 0.82, 0.02, and 1 respectively.

The one class SVM obtained a sensitivity of 0.34, specificity of 0.87, PPV of 0.01 and NPV of 1 (Table 5.4). At the probability thresholds leading to the same number of positive predictions as produced by the one class SVM (5217 positive predictions), the XGBoost and logistic regression models had a higher sensitivity and PPV, and an equal specificity and NPV. The clinical prediction rule had a sensitivity of 0.05, specificity of 1, PPV of 0.07 and NPV of 1 (Table 5.4). With 138 positive predictions as for the clinical rule, the XGBoost and logistic regression models again had a higher sensitivity and PPV, and equal specificity and NPV.

Table 5.4 Threshold specific performance metrics for predicting exacerbation within 2 days (validation cohort)

Probability threshold	Model	Sensitivity	Specificity	PPV	NPV
0.001	XGBoost	0.71 (133/188)	0.81 (32178/39904)	0.02 (133/7859)	1.0 (32178/32233)
	Logistic regression	0.93 (174/188)	0.56 (22227/39904)	0.01 (174/17851)	1.0 (22227/22241)
0.002	XGBoost	0.59 (110/188)	0.89 (35326/39904)	0.02 (110/4688)	1.0 (35326/35404)
	Logistic regression	0.84 (158/188)	0.82 (32720/39904)	0.02 (158/7342)	1.0 (32720/32750)
Resulting in 5217 positive predictions**	One class SVM	0.34 (64/188)	0.87 (34751/39904)	0.01 (64/5217)	1.0 (34751/34875)
	XGBoost	0.6 (112/188)	0.87 (34800/39904)	0.02 (112/5216)	1.0 (34800/34876)
	Logistic regression	0.73 (137/188)	0.87 (34823/39904)	0.03 (137/5218)	1.0 (34823/34874)
Resulting in 138 positive predictions**	Clinical rule*	0.05 (10/188)	1.0 (39776/39904)	0.07 (10/138)	1.0 (39776/39954)
	XGBoost	0.11 (21/188)	1.0 (39787/39904)	0.15 (21/138)	1.0 (39787/39954)
	Logistic regression	0.11 (20/188)	1.0 (39787/39904)	0.15 (20/137)	1.0 (39787/39955)

*Peak Expiratory Flow < 60% personal best
**This threshold is set so that the XGBoost and logistic regression models produce the same number of positive predictions as the one class SVM or clinical rule.
Abbreviations: *SVM* Support Vector Machine, *XGBoost* gradient boosted decision trees, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value

Similar results were found for the prediction of exacerbations within 4 and 8 days as the 2-days models (supplementary Tables S2-S5). The AUC of the XGBoost model increased for the 5-lag model (0.85, 95% CI 0.82-0.87, supplementary Table S6). No such improvement for a higher number of lags was found for the logistic regression model (based on AUC, supplementary Table S6). The one class SVM model showed a higher sensitivity, but lower specificity for the 2-lag and 3-lag models, and a sensitivity of (almost) 1 and specificity of almost 0 for the 4-lag and 5-lag models (supplementary Table S7). The differences between the AUCs of the best performing logistic regression model with one lag and XGBoost model with five lags were still significant (p=0.02, DeLong test).

5.5 DISCUSSION

In this study, we aimed to assess the performance of ML techniques and classic models for short-term prediction of severe asthma exacerbations based on home monitoring data. ML and logistic regression both reached higher discriminative performance than a previously proposed simple clinical rule. Logistic regression provided slightly better discriminative performance than the XGBoost algorithm. However, logistic regression still produced many false positives at high levels of sensitivity.

Our finding that ML models do not outperform classical prediction methods is in line with other recent studies [14, 23-25]. This finding may be explained by the (lack of) complexity of the data that was studied. An advantage of ML techniques is the natural flexibility they offer to model complex (e.g. highly nonlinear) relationships, versus logistic regression techniques that have the advantage of being easily interpretable. Our findings illustrate that the flexibility provided by ML models may not always be needed to arrive at the best performing prediction model for medical data. The benefits of ML methods may differ between settings and should be further investigated.

Second, we found a substantial number of false positive predictions at high levels of sensitivity. The false positive rate (reflected by the low PPV) can be linked directly to the low incidence rate. Similar results can be found in the literature [2, 26-29]. The potential implications of the high false positive rate are alarm fatigue, loss of model acceptance and trust, and ultimately disuse of the prediction model [30]. Improvement in discriminative ability may be achieved by reducing the noise in the exacerbation event at the time of data collection. For example, the recording of severe exacerbations in our dataset might have been incomplete or there might have been a delay between the recording of the exacerbations and their true onset. Moreover, better predicting variables of exacerbations may be needed, which need evaluation in large data sets.

Another insight based on our findings is that the interpretability of a prediction algorithm does not always have to come at the cost of model performance. An argument in favor of black-box ML and its broader field of artificial intelligence (AI) techniques is their potentially superior predictive performance.

For this superior performance, it is deemed acceptable to not exactly know how a prediction is made: the accuracy-interpretability trade-off [31, 32]. Our findings form a counterexample by showing that inherently interpretable techniques such as logistic regression may outperform ML for certain application types and clinical settings. Interpretability is especially relevant for clinical settings, as physicians often prefer interpretable models to assist in clinical decision making.

Strengths of our study include that we performed a comparison of ML models with a statistical model and a clinical prediction rule, which to our knowledge has not, or only partly been performed for this type of home monitoring data [14]. Our findings therefore contribute to answering the question when and how to apply ML methods safely and effectively, thereby putting ML in perspective. Moreover, the data used in this study contained few missing values, possibly due to the trial setting. The quality of the data was therefore high.

The current investigation also had limitations. First, by opting to predict exacerbation in the short-term (exacerbation within two days), the exacerbation window became small. Such a small window was chosen to keep the predictions clinically meaningful and relevant. This resulted in a very low incidence rate. We performed a sensitivity analysis in which we expanded the window to four and eight days without noticeable differences in model performance. We therefore recommend investigating the best way to operationalize and capture the clinical definition of a severe asthma exacerbation in home monitoring data. Second, the low event rate may have caused the (best performing) logistic regression model to consistently underestimate the predicted risks [33]. Low event rates are common for the home monitoring setting. We therefore advise future researchers to investigate techniques that address any associated calibration issues. Poor calibration forms an obstacle for the implementation of any algorithm in clinical practice, since reliability of the predicted probabilities is required to be clinically meaningful [22]. Lastly, home monitoring patients based on daily diary entries can be perceived as old fashioned. Clinicians nowadays will often opt for digital telemonitoring approaches. Yet, the monitored parameters have remained largely the same across different registration modes (on paper or digitally) [26, 34-36]. This implies that the registration method is unlikely to affect our conclusions.

5.6 CONCLUSION

ML models may not outperform classical regression prediction model in predicting short-term asthma exacerbations based on home monitoring data. A simple regression model outperforms a simple rule. Clinical application may be challenging, due to the high false alarm rate associated with the low probability thresholds required for high sensitivity.

REFERENCES

1. Malasinghe, L.P., N. Ramzan, and K. Dahal, *Remote patient monitoring: a comprehensive study.* Journal of Ambient Intelligence and Humanized Computing, 2019. **10**(1): p. 57-76.
2. Honkoop, P.J., D.R. Taylor, A.D. Smith, J.B. Snoeck-Stroband, and J.K. Sont, *Early detection of asthma exacerbations by using action points in self-management plans.* Eur Respir J, 2013. **41**(1): p. 53-9.
3. Fine, M.J., T.E. Auble, D.M. Yealy, et al., *A Prediction Rule to Identify Low-Risk Patients with Community-Acquired Pneumonia.* New England Journal of Medicine, 1997. **336**(4): p. 243-250.
4. Wells, P.S., D.R. Anderson, M. Rodger, et al., *Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer.* Thromb Haemost, 2000. **83**(3): p. 416-20.
5. British Thoraic Society, *British Guideline on the Management of Asthma.* 2019.
6. Mak, R.H., M.G. Endres, J.H. Paik, et al., *Use of Crowd Innovation to Develop an Artificial Intelligence–Based Solution for Radiation Therapy Targeting.* JAMA Oncology, 2019. **5**(5): p. 654-661.
7. Esteva, A., B. Kuprel, R.A. Novoa, et al., *Dermatologist-level classification of skin cancer with deep neural networks.* Nature, 2017. **542**(7639): p. 115-118.
8. McKinney, S.M., M. Sieniek, V. Godbole, et al., *International evaluation of an AI system for breast cancer screening.* Nature, 2020. **577**(7788): p. 89-94.
9. Cearns, M., T. Hahn, and B.T. Baune, *Recommendations and future directions for supervised machine learning in psychiatry.* Translational Psychiatry, 2019. **9**(1): p. 271.
10. Neuhaus, A.H. and F.C. Popescu, *Sample Size, Model Robustness, and Classification Accuracy in Diagnostic Multivariate Neuroimaging Analyses.* Biological Psychiatry, 2018. **84**(11): p. e81-e82.
11. Chen, P.-H.C., Y. Liu, and L. Peng, *How to develop machine learning models for healthcare.* Nature Materials, 2019. **18**(5): p. 410-414.
12. Altman, D.G., Y. Vergouwe, P. Royston, and K.G.M. Moons, *Prognosis and prognostic research: validating a prognostic model.* BMJ, 2009. **338**: p. b605.
13. Wynants, L., L.J.M. Smits, and B. Van Calster, *Demystifying AI in healthcare.* BMJ, 2020. **370**: p. m3505.
14. Tsang, K.C.H., H. Pinnock, A.M. Wilson, and S.A. Shah. *Application of Machine Learning to Support Self-Management of Asthma with mHealth*. in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020.
15. Smith, A.D., J.O. Cowan, K.P. Brassett, G.P. Herbison, and D.R. Taylor, *Use of exhaled nitric oxide measurements to guide treatment in chronic asthma.* N Engl J Med, 2005. **352**(21): p. 2163-73.
16. Taylor, D.R., G.I. Town, G.P. Herbison, et al., *Asthma control during long-term treatment with regular inhaled salbutamol and salmeterol.* Thorax, 1998. **53**(9): p. 744-52.

17. Smith, A.E., C.D. Nugent, and S.I. McClean, *Evaluation of inherent performance of intelligent medical decision support systems: utilising neural networks as an example.* Artif Intell Med, 2003. **27**(1): p. 1-27.
18. Nielsen, D., *Tree boosting with xgboost-why does xgboost win" every" machine learning competition?* 2016, NTNU.
19. Ma, J. and S. Perkins. *Time-series novelty detection using one-class support vector machines*. in *Proceedings of the International Joint Conference on Neural Networks, 2003.* 2003. IEEE.
20. Schober, P. and T.R. Vetter, *Logistic Regression in Medical Research.* Anesthesia and analgesia, 2021. **132**(2): p. 365-366.
21. Steyerberg, E.W., *Clinical Prediction Models*, ed. M. Gail, M.S. Jonathan, and B. Singer. 2009, Cham, Switzerland: Springer Nature.
22. Van Calster, B., D.J. McLernon, M. van Smeden, et al., *Calibration: the Achilles heel of predictive analytics.* BMC Medicine, 2019. **17**(1): p. 230.
23. Christodoulou, E., J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, and B. Van Calster, *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* Journal of Clinical Epidemiology, 2019. **110**: p. 12-22.
24. Gravesteijn, B.Y., D. Nieboer, A. Ercole, et al., *Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury.* J Clin Epidemiol, 2020. **122**: p. 95-107.
25. Nusinovici, S., Y.C. Tham, M.Y. Chak Yan, et al., *Logistic regression was as good as machine learning for predicting major chronic diseases.* Journal of Clinical Epidemiology, 2020. **122**: p. 56-69.
26. Martin, A., V. Bauer, A. Datta, et al., *Development and validation of an asthma exacerbation prediction model using electronic health record (EHR) data.* Journal of Asthma, 2020. **57**(12): p. 1339-1346.
27. Sanders, S., J. Doust, and P. Glasziou, *A Systematic Review of Studies Comparing Diagnostic Clinical Prediction Rules with Clinical Judgment.* PLOS ONE, 2015. **10**(6): p. e0128233.
28. Satici, C., M.A. Demirkol, E. Sargin Altunok, et al., *Performance of pneumonia severity index and CURB-65 in predicting 30-day mortality in patients with COVID-19.* Int J Infect Dis, 2020. **98**: p. 84-89.
29. Obradović, D., B. Joveš, S. Pena Karan, S. Stefanović, I. Ivanov, and M. Vukoja, *Correlation between the Wells score and the Quanadli index in patients with pulmonary embolism.* Clin Respir J, 2016. **10**(6): p. 784-790.
30. Winters, B.D., M.M. Cvach, C.P. Bonafide, et al., *Technological Distractions (Part 2): A Summary of Approaches to Manage Clinical Alarms With Intent to Reduce Alarm Fatigue.* Crit Care Med, 2018. **46**(1): p. 130-137.
31. Mori, T. and N. Uchihira, *Balancing the trade-off between accuracy and interpretability in software defect prediction.* Empirical Software Engineering, 2019. **24**(2): p. 779-825.
32. Johansson, U., C. Sönströd, U. Norinder, and H. Boström, *Trade-off between accuracy and interpretability for predictive in silico modeling.* Future Med Chem, 2011. **3**(6): p. 647-63.

33. Wallace, B.C. and I.J. Dahabreh, *Improving class probability estimates for imbalanced data.* Knowledge and Information Systems, 2014. **41**(1): p. 33-52.
34. Honkoop, P.J., A. Simpson, M. Bonini, et al., *MyAirCoach: the use of home-monitoring and mHealth systems to predict deterioration in asthma control and the occurrence of asthma exacerbations; study protocol of an observational study.* BMJ Open, 2017. **7**(1): p. e013935.
35. Finkelstein, J. and I.C. Jeong, *Machine learning approaches to personalize early prediction of asthma exacerbations.* Ann N Y Acad Sci, 2017. **1387**(1): p. 153-165.
36. Sanchez-Morillo, D., M.A. Fernandez-Granero, and A. Leon-Jimenez, *Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: A systematic review.* Chronic Respiratory Disease, 2016. **13**(3): p. 264-283.

SUPPLEMENTARY MATERIAL

Table S1 Hyper parameters machine learning and logistic regression models

Hyperparameters	Values	Parameters for exacerbations within 2 days	Parameters for exacerbations within 4 days	Parameters for exacerbations within 8 days
XGBoost				
Number of trees	[50, 100, 200]	25	25	25
Maximum depth	[1, 3, 5, 7, 8, 9]	3	3	3
Learning rate	[0.1, 0.3, 0.5, 0.7, 0.9]	0.7	0.9	0.7
One class SVM				
nu	[0.001, 0.0015, 0.002, 0.004, 0.006, 0.008, 0.01]	0.001	0.001	0.001
gamma	[0.001, 0.01, 0.1, 1]	0.001	0.001	0.001
Logistic regression				
Penalty	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]	0.9	0.7	0.9

Table S2 Discrimination and calibration for predicting exacerbation within 4 days

	AUC	Calibration intercept	Calibration slope
XGBoost	0.77 (0.74, 0.79)	0.57 (0.45, 0.68)	0.48 (0.43, 0.52)
Logistic regression	0.85 (0.83, 0.86)	0.65 (0.54, 0.75)	1.07 (1.0, 1.14)

Abbreviations: *XGBoost* gradient boosted decision trees, *AUC* Area Under the Receiver Operating Characteristics Curve

Table S3 Threshold specific performance metrics for predicting exacerbation within 4 days

Probability threshold	Model	Sensitivity	Specificity	PPV	NPV
0.001	XGBoost	0.77 (285/372)	0.6 (23638/39720)	0.02 (285/16367)	1.0 (23638/23725)
	Logistic regression	0.99 (370/372)	0.06 (2564/39720)	0.01 (370/37526)	1.0(2564/2566)
0.002	XGBoost	0.69 (258/372)	0.76 (30234/39720)	0.03 (258/9744)	1.0 (30234/30348)
	Logistic regression	0.96 (356/372)	0.36 (14385/39720)	0.01 (356/25691)	1.0 (14385/14401)
Resulting in 5217 positive predictions**	One class SVM	0.3 (113/372)	0.87 (34616/39720)	0.02 (113/5217)	0.99 (34616/34875)
	XGBoost	0.58 (215/372)	0.87 (34543/39720)	0.04 (215/5392)	1.0 (34543/34700)
	Logistic regression	0.62 (229/372)	0.88 (34810/39720)	0.04 (229/5139)	1.0 (34810/34953)
Resulting in 138 positive predictions**	Clinical rule*	0.03 (13/372)	1.0 (39595/39720)	0.09(13/138)	0.99 (39595/39954)
	XGBoost	0.03 (11/372)	1.0 (39593/39720)	0.08(11/138)	0.99 (39593/39954)
	Logistic regression	0.06 (22/372)	1.0 (39604/39720)	0.16(22/138)	0.99 (39604/39954)

*Peak Expiratory Flow < 60% personal best

**This threshold is set so that the XGBoost and logistic regression models produce the same number of positive predictions as the one class SVM or clinical rule.

Abbreviations: *SVM* Support Vector Machine, *XGBoost* gradient boosted decision trees, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value

Table S4 Discrimination and calibration for predicting exacerbation within 8 days

	AUC	Calibration intercept	Calibration slope
XGBoost	0.7 (0.68, 0.72)	0.58 (0.5, 0.66)	0.43 (0.39, 0.47)
Logistic regression	0.81 (0.79, 0.82)	0.59 (0.52, 0.67)	1.11 (1.04, 1.17)

Abbreviations: *XGBoost* gradient boosted decision trees, *AUC* Area Under the Receiver Operating Characteristics Curve

Table S5 Threshold specific performance metrics for predicting exacerbation within 8 days

Probability threshold	Model	Sensitivity	Specificity	PPV	NPV
0.001	XGBoost	0.86 (615/712)	0.25 (9817/39380)	0.02 (615/30178)	0.99 (9817/9914)
	Logistic regression	1.0 (712/712)	0.0 (149/39380)	0.02 (712/39943)	1.0 (149/149)
0.002	XGBoost	0.79 (563/712)	0.37 (14592/39380)	0.02 (563/25351)	0.99 (14592/14741)
	Logistic regression	1.0 (711/712)	0.03 (1224/39380)	0.02 (711/38867)	1.0 (1224/1225)
Resulting in 5217 positive predictions**	One class SVM	0.3 (211/712)	0.87 (34374/39380)	0.04 (211/5217)	0.99 (34374/34875)
	XGBoost	0.41 (293/712)	0.87 (34385/39380)	0.06 (293/5288)	0.99 (34385/34804)
	Logistic regression	0.5 (357/712)	0.88 (34512/39380)	0.07 (357/5225)	0.99 (34512/34867)
Resulting in 138 positive predictions**	Clinical rule*	0.03 (21/712)	1.0 (39263/39380)	0.15 (21/138)	0.98 (39263/39954)
	XGBoost	0.02 (17/712)	1.0 (39259/39380)	0.12 (17/138)	0.98 (39259/39954)
	Logistic regression	0.05 (35/712)	1.0 (39277/39380)	0.25 (35/138)	0.98 (39277/39954)

*Peak Expiratory Flow < 60% personal best

**This threshold is set so that the XGBoost and logistic regression models produce the same number of positive predictions as the one class SVM or clinical rule.

Abbreviations: *SVM* Support Vector Machine, *XGBoost* gradient boosted decision trees, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value

Table S6 Discrimination for predicting exacerbation within 2 days with varying number of lags

	AUC XGBoost	AUC Logistic regression
1 lag	0.81 (0.78, 0.84)	0.88 (0.86, 0.90)
2 lags	0.81 (0.78, 0.84)	0.88 (0.86, 0.90)
3 lags	0.82 (0.79, 0.85)	0.88 (0.86, 0.90)
4 lags	0.82 (0.80, 0.85)	0.88 (0.85, 0.90)
5 lags	0.85 (0.82, 0.87)	0.88 (0.85, 0.90)

Abbreviations: *XGBoost* gradient boosted decision trees, *AUC* Area Under the Receiver Operating Characteristics Curve

Table S7 Classification of one class SVM for predicting exacerbation within 2 days with varying number of lags

	Sensitivity	Specificity	PPV	NPV
1 lag	0.34 (64/188)	0.87 (34751/39904)	0.01 (64/5217)	1.0 (34751/34875)
2 lags	0.45 (85/188)	0.85 (33985/39904)	0.01 (85/6004)	1.0 (33985/34088)
3 lags	0.46 (86/188)	0.83 (33296/39904)	0.01 (86/6694)	1.0 (33296/33398)
4 lags	1.0 (188/188)	0.02 (840/39904)	0.0 (188/39252)	1.0 (840/840)
5 lags	0.99 (186/188)	0.03 (1022/39904)	0.0 (186/39068)	1.0 (1022/1024)

Abbreviations: *SVM* Support Vector Machine, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value

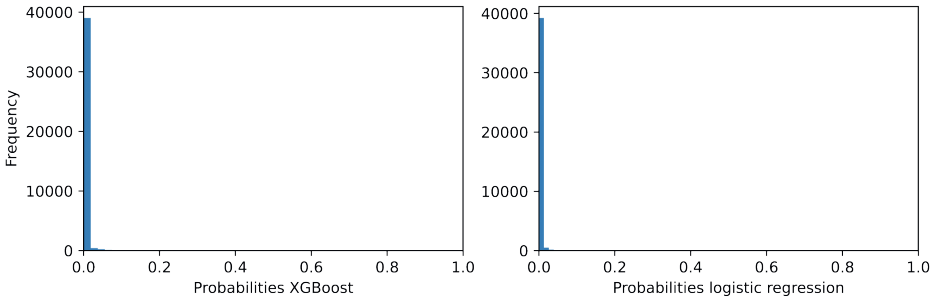

Figure S8 Histogram of probability predictions for a) XGBoost model and b) logistic regression model

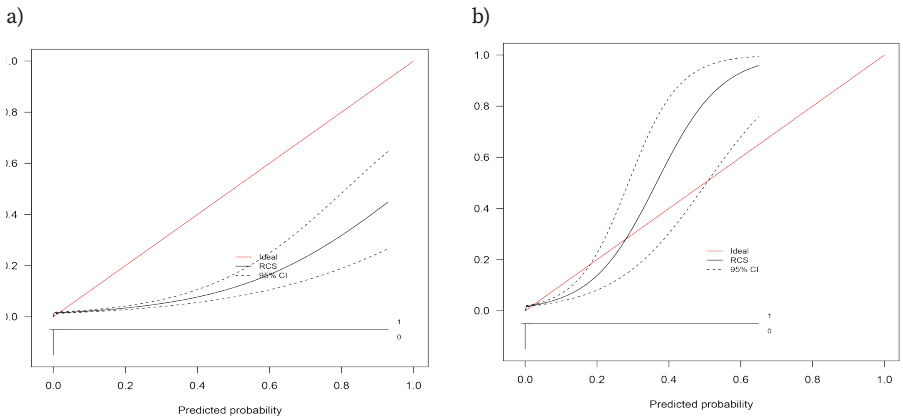

Figure S9 Calibration curves for a) XGBoost model and b) logistic regression model