Diagnostic Modeling to Identify Unrecognized Inpatient Hypercapnia Using Health Record Data

Brian W. Locke[1] [0000-0002-3588-5238], W. Wayne Richards[1,2], Jeanette P. Brown[1], Wanting Cui[2] [0000-0001-7341-363X], Joseph Finkelstein[2] [0000-0002-8084-7441], Krishna M. Sundar[1] [0000-0001-7220-5767], and Ramkiran Gouripeddi[2,3] [0000-0002-4345-9669]

[1] Division of Pulmonary, Critical Care, and Occupational Pulmonary Medicine, University of Utah, Salt Lake City, UT.
[2] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT.
[3] Clinical and Translational Science Institute, University of Utah, Salt Lake City, UT.
brian.locke@hsc.utah.edu

Abstract. Hypercapnic respiratory failure (an accumulation of carbon dioxide, CO_2, in the blood) is often missed in clinical practice. Arterial blood gas is the standard diagnostic test, but it is painful and not routine. When clinicians fail to make the diagnosis, it is often because an arterial blood gas was not obtained. This 'partial verification' of CO_2 levels presents a challenge for machine learning algorithms. We assessed the accuracy of two machine learning methods using demographics and routine lab work to estimate the likelihood that a patient has hypercapnic respiratory failure at hospital admission. Hospitalized patients who received an arterial blood gas sample consituted the training (n=111,015) and geographic validation (n=20,834) sets. Acceptance of "silver standard" diagnostic criteria and weighting observations by their modeled likelihood of receiving arterial blood gas sampling were used to assess the stability of findings in the presence of partial verification. Both regularized logistic regression and random-forest-based models resulted in acceptable performance (area under the curve: 0.763 and 0.758 respectively), with minimal changes in the auxiliary analyses. This work suggests that routinely available health record data can stratify the likelihood of hypercapnic respiratory failure among hospitalized adults, and findings may generalize to patients who have not received arterial blood gas sampling in clinical practice.

Keywords: Hypercapnic Respiratory Failure, Diagnostic Model, Machine Learning, Partial Verification

1 Introduction

Hypercapnic respiratory failure is a condition where the amount of inspired air that participates in gas exchange is insufficient to match the metabolic production of carbon dioxide (CO_2) in the body, leading to a buildup of CO_2 in the blood called hypercapnia. Hypercapnic respiratory failure is common among patients presenting to hospitals [1], is associated with high rates of readmission [2], and indicates a high risk of death in the months after recognition [3, 4].

Clinicians often fail to recognize patients who have hypercapnic respiratory failure [5]. Organ failure or death occurs when respiratory failure leads to low oxygen levels, but high blood CO_2 levels are acutely better tolerated. Thus, unlike oxygen, blood CO_2 monitoring is not routine in hospitalized patients. Arterial blood gas sampling is the reference standard diagnostic test to confirm hypercapnia. It is painful [6] and can lead to complications [7], so clinicians only order the test when their suspicion for hypercapnia is particularly high.

Methods to reliably predict which patients have hypercapnic respiratory failure could improve patient outcomes by helping to identify high-risk patients that are missed in clinical practice. For many of these patients, evidence-based treatments exist to improve their symptoms, lessen their risk of hospitalization, and improve their mortality[8]. Some routinely obtained lab values, like serum potassium and bicarbonate, change the likelihood of hypercapnia substantially [9, 10]. Using additional data elements to model the likelihood of hypercapnia being present, termed diagnostic modeling, might improve diagnostic reliability. However, diagnostic models for hypercapnic respiratory failure have not been previously reported.

A key challenge to developing such models is partial verification [11]: blood CO_2 levels are known only for the patients who underwent arterial blood sampling. The patients who would benefit most from more reliable diagnosis are less likely to have received arterial blood gas testing, and are therefore under-represented among the cases available for training a diagnostic model. In the literature evaluating medical tests, two commonly used approaches to address partial verification bias are analyses accepting "silver standard" diagnostic criteria and weighting observations by the likelihood that they'd receive the definitive test. In this work, we aim to assess the accuracy of two approaches to diagnostic modeling of hypercapnic respiratory failure in the presence of partial verification blood CO_2 levels.

2 Methods

This was a retrospective analysis of de-identified electronic health record data from the TriNetX research network database (TriNetX, LLC. Cambridge, Massachusetts) and was exempted by the University of Utah Institutional Review Board (#00152089).

2.1 Data Source

The TriNetX research network is a federated network of electronic health record data from 76 medical centers across the US, serving roughly 115 million patients [12]. All adult, inpatient encounters occurring during the calendar year 2022 that met any of the following criteria were requested: received a diagnostic code for any respiratory failure, had a condition known to cause hypercapnia (including severe obesity), received a procedure code for the treatment of respiratory failure (non-invasive or invasive ventilation), or had an arterial (ABG) or venous blood gas (VBG) obtained on the first day of the encounter. These criteria indicate that consideration of hypercapnia was warranted, which is the spectrum of patients in whom diagnostic modeling would be used [13]. First encounters for each patient were used. Data was cleaned to remove

physiologically impossible data and encounters with evidence of incomplete data submission to TriNetX (for example: missing categories of data such as no procedure codes or diagnoses for the encounter of interest).

2.2 Model Creation and Feature Selection

Predictors (age, sex, body mass index [BMI], components of basic blood chemistry testing [sodium, potassium, chloride, bicarbonate, urea, and creatinine], and hemoglobin) were selected a priori on clinical relevance. Additionally, these data elements are routinely present in all hospital admissions and are clinically ascertained independent of (and, generally, before) knowledge of the presence of hypercapnic respiratory failure. No imputation was performed as data elements were selected for low missingness.

To focus on the influence of partial verification, two standard machine learning approaches were used: logistic regression with L1-regularization (termed logistic least absolute shrinkage and selection operator, or LASSO regression, for short) and random forest modeling. LASSO and random forest models were selected for their ubiquity and to balance understandble model outputs compared to possible performance gains from the handling of non-linear relationships. For the LASSO regression, continuous predictors were represented as restricted cubic splines with 4 knots, and 10-fold cross-validation was used to select the minimum prediction error λ. For random forest, hyperparameter (tree depth, number of splitting features, and number of bootstrapped trees) tuning was performed using grid search and 10-fold cross-validation.

2.3 Performance Analysis

Patients who had any ABG on the calendar day of admission showing a partial pressure of CO_2 ($PaCO_2$) over 45 mm Hg were considered to have hypercapnia. In the primary analysis, diagnostic model predictions of the likelihood of hypercapnia were generated for patients in whom ABG sampling was performed. Model discrimination (ability to separate patients with hypercapnia from those without) was assessed using receiver operating characteristic (ROC) curves and summarized as the area under the ROC curve (AUC). Model calibration (how closely the predicted likelihood of hypercapnia correlates with the true likelihood) was assessed by the full sample expected to observed event ratio (E:O), calibration in the large (CITL; the relation of the mean predicted risk to the mean observed risk), calibration slope (CS, slope; whether risks are too extreme for high- and low-risk patients) and calibration plots visualized by decile of predicted risk. Overall performance was summarized using Brier scores. The importance of individual predictors was assessed using standardized regression coefficients (LASSO) and mean decrease in impurity (random forest). Models were trained on data submitted from the Western, Southeastern, and Northeastern US regions, and evaluated in hospitals from the Midwestern region. The evaluation region was chosen arbitrarily by two coin flips and sensitivity analyses holding out the other regions for evaluation show similar results but are omitted for brevity. Geographic validation, rather than a random hold-out set, was performed for a more severe test of the distribution shift associated with regional variations in care practices.

2.4 Evaluation of the Influence of Partial Verification

Two additional analyses were performed to assess the applicability of diagnostic modeling to patients who did not receive reference-standard (ABG) blood CO_2 level assessment. First, VBG sampling is considered a "silver standard" method of diagnosis, with high venous CO_2 levels serving as an imperfect but usable surrogate for arterial CO_2 levels [14]. The same performance metrics were calculated on patients who had either an ABG or VBG that showed hypercapnia (an arterial $PaCO_2$ over 45 mm Hg or a venous PCO_2 over 50 mm Hg) obtained on the day of admission.

Second, inverse probability weighting was used to create a *pseudopopulation* that represents the sample where all patients had an equal probability of receiving ABG sampling, conditional on the covariables used to model the propensity of receiving an ABG sample [11]. Logistic regression was used to generate propensity scores, using prior diagnoses, demographics, lab values, and outpatient medications clinically suspected to relate to the propensity to receive an ABG. The missing indicator method was used to account for missing data [14]. The model was then evaluated in the weighted population to estimate the model's performance if it had been applied across the entire population, as opposed to only those who actually received an ABG.

Statistical analysis used Stata 18 (StataCorp, College Station, TX) with pmcalplot [17] and scikit-learn (1.4.1) [15] via c_ml_stata_cv packages [16].

3 Results

Of 401,079 potentially eligible adult, inpatient encounters, n=32,987 were excluded for missing data, and n=68,190 were repeat encounters (Figure 1). For the primary analysis, an admission-day ABG sample was obtained in 131,849 of 299,902 patients. An additional n=44,600 had a VBG and were included in the supplementary analysis.

Demographics of the included groups are given in Table 1. Among the patients who underwent ABG sampling on the day of hospital admission, 30% (n=39,676) had hypercapnia. When VBG verification was accepted to determine hypercapnia status, 33% (n=57,349) of patients who received either an arterial or venous blood gas showed evidence of hypercapnia. All predictors had <10% missingness rates.

LASSO logistic regression modeling achieved an AUC of 0.763 and a Brier score of 0.180 in the test set (Midwestern US Hospitals), with adequate calibration (Figure 2). Serum bicarbonate, potassium, and BMI were the strongest individual predictors. Performance dropped to an AUC of 0.749 and a Brier score of 0.190 in the silver standard analysis and increased to an AUC of 0.792 and a Brier Score of 0.178 when inverse propensity weighting was performed (Table 2).

Random forest-based predictions achieved clinically indistinguishable performance in the primary (AUC 0.758, Brier score 0.184), silver standard (AUC 0.745, Brier Score 0.193), and inverse probability-weighted analysis (AUC 0.782, Brier score 0.184). Particularly in the silver standard analysis, there was a mild underestimation of the likelihood of hypercapnia across all risk categories in both regression and random forest-based models. Serum bicarbonate had the highest feature importance, followed by serum creatinine, hemoglobin, and potassium.

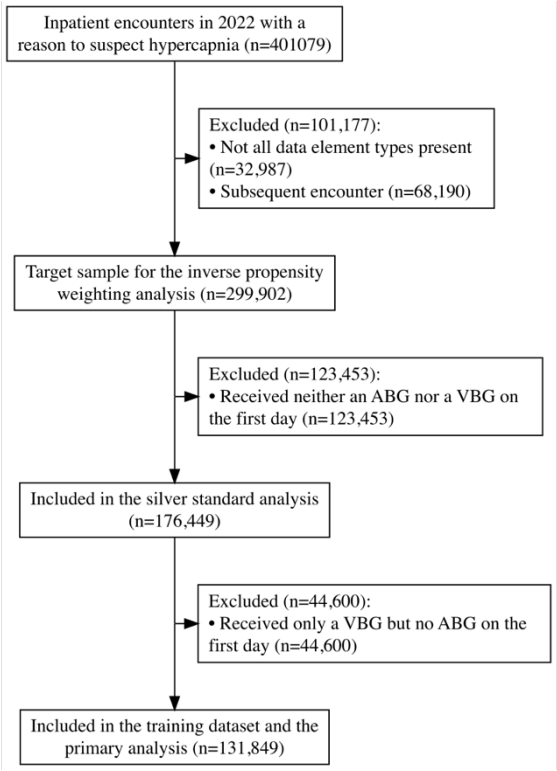

Fig. 1. Enrollment Flowchart. Only patients who received arterial blood gas sampling to verify their hypercapnia status were included in the primary analysis. ABG = arterial blood gas. VBG = venous blood gas

Table 1. Characteristics of included patients. Western, Northeastern, and Southeastern US hospitals constituted the training set, while Midwestern hospitals were the testing set. Diagnoses were based on diagnosis codes rendered up to the admission of interest. BMI = body mass index, COPD = chronic obstructive pulmonary disease, ABG = arterial blood gas (threshold 45 mmHg), VBG = venous blood gas (threshold 50 mmHg). EHR-recorded death occurred with a median follow-up of 11 months.

	Entire Cohort	First-day ABG or VBG obtained	First-day ABG obtained
		Silver standard analysis	Training, primary analyses
	N=299,902	N=176,449	N=131,849
Training dataset	Not applicable	Not applicable	84% (111,015)
Testing dataset	Not applicable	15% (27,138)	16% (20,834)
Age (years)	62 (±17)	62 (±17)	62 (±17)
Female	47% (141,032)	45% (79,805)	45% (58,792)

Black or African American	18% (53,933)	17% (30,483)	17% (22,006)
Asian	2% (5,866)	2% (3,860)	2% (2,922)
White	68% (204,471)	67% (118,890)	68% (89,414)
Hispanic or Latino Ethnicity	6% (18,366)	6% (10,777)	6% (7,574)
BMI (kg/m^2)	30 (±9)	29 (±8)	29 (±8)
Heart failure	17% (50,053)	16% (27,761)	16% (20,675)
Chronic kidney disease	15% (45,599)	15% (25,670)	14% (18,521)
COPD	15% (45,301)	15% (26,137)	15% (19,309)
Neuromuscular disease	3% (10,338)	3% (6,015)	4% (4,686)
Obstructive sleep apnea	14% (42,029)	11% (20,052)	11% (14,512)
Hypercapnia on admission-day ABG			
No ABG	56% (168,053)	25% (44,600)	0% (0)
All PCO_2 < threshold	31% (92,173)	52% (92,173)	70% (92,173)
$PCO_2 \geq$ threshold	13% (39,676)	22% (39,676)	30% (39,676)
Hypercapnia on admission-day ABG or VBG			
No VBG or ABG	41% (123,439)	0% (0)	0% (0)
All PCO_2 < threshold	40% (119,104)	67% (119,104)	66% (87,129)
Any $PCO_2 \geq$ threshold	19% (57,349)	33% (57,349)	34% (44,722)
Critical care services	26% (78,937)	35% (60,943)	38% (49,917)
Death	16% (47,012)	18% (32,560)	19% (25,646)

Table 2. Performance Metrics. Only test set results are shown. Discrim. = discrimination, ABG = arterial blood gas, VBG = venous blood gas, LASSO = least absolute shrinkage and selection operator, RF = random forest, E:O = expected to observed ratio, CITL = calibration in the large, CS = calibration slope, AUC = area under the receiver operating characteristic curve. A score of 1 is perfect for the AUC, E:O, and CS. For CITL and Brier Score, a score of 0 is perfect.

Analysis	Patients	Model	Discrim.	Calibration			Overall
			AUC	E:O	CITL	CS	Brier Score
Primary	ABG	LASSO	0.763	0.986	0.028	1.110	0.180
		RF	0.758	0.969	0.062	1.127	0.184
Silver Standard	ABG or VBG	LASSO	0.749	0.925	0.156	1.098	0.190
		RF	0.745	0.912	0.180	1.190	0.193
Inverse Proba-bility Weighted	ABG (weighted to full sample)	LASSO	0.792	0.982	0.041	1.227	0.178
		RF	0.782	0.957	0.094	1.251	0.184

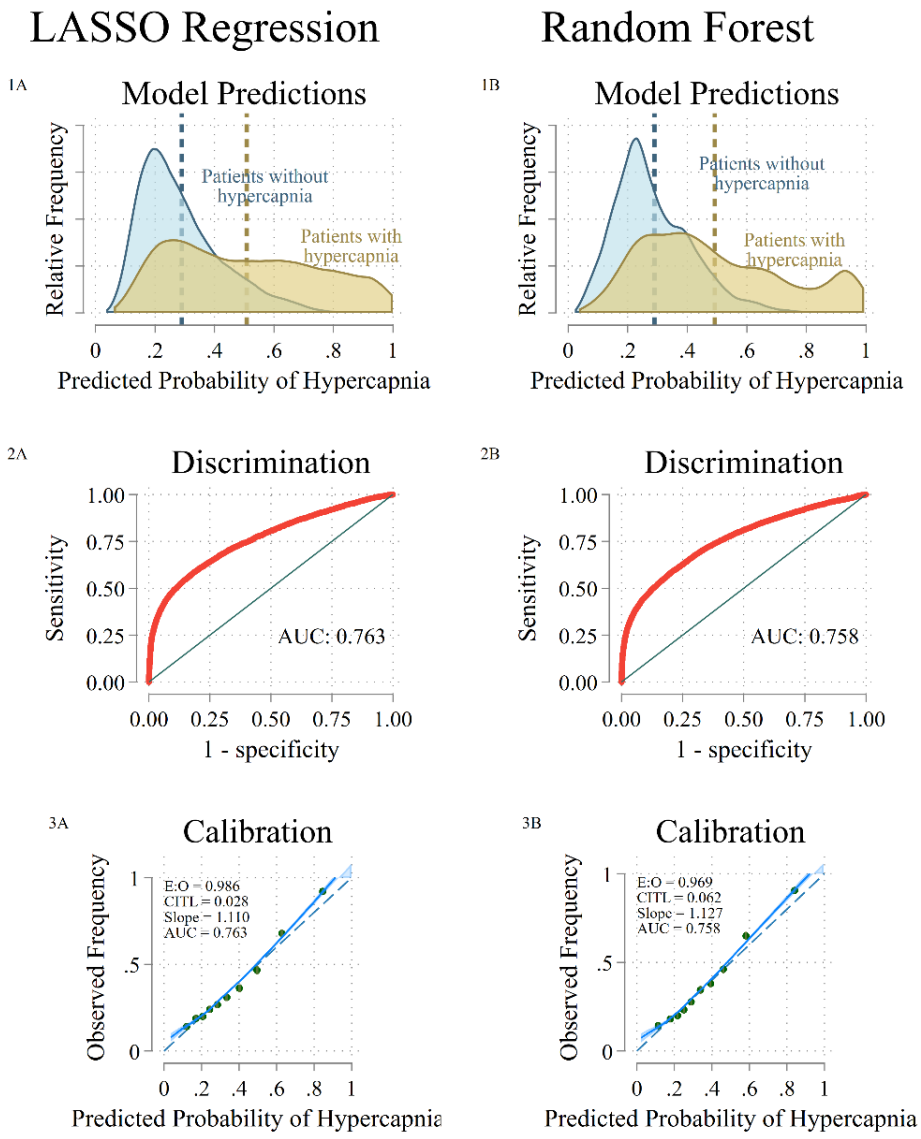

Fig. 2. Performance of L1-regularized (LASSO) logistic regression and random forest-based diagnostic models of the likelihood of hypercapnia at hospital admission. Panel 1A and 1B show the distribution of model predictions by hypercapnia status. Mean predictions are indicated by dashed lines. In panel 3A and 3B, calibration is presented by decile of predicted risk. E:O = expected to observed ratio , CITL = calibration in the large (intercept), slope = calibration slope, AUC = area under the receiver operating characteristic curve.

4 Discussion

Both LASSO logistic regression and random forest-based diagnostic models achieved acceptable discrimination and calibration for identifying patients with hypercapnic respiratory failure at the time of hospital admission. Both models' performance was maintained when two methods assessing the impact of partial outcome ($PaCO_2$) verification were applied. This suggests that either modeling approach may help identify patients with hypercapnia that are currently not recognized in clinical care.

Both models exclusively rely on predictors collected in nearly all acutely hospitalized patients (routine lab work and demographics). This modeling approach could be used to risk-stratify patients that plausibly have hypercapnia. As can be seen from Figure 2, panels 1A and 1B, both models were able to render stronger "rule-in" predictions (raising the likelihood of hypercapnia) as compared to "rule-out" predictions, which suggests utility flagging additional patients for further workup as opposed to identifying patients in whom further testing is not indicated..

A key barrier to developing a diagnostic model for hypercapnic respiratory failure is that not all patients receive an assessment of their blood CO_2 levels. To address this, two approaches could be considered. Theoretically, a model could be trained on a prospectively constructed research cohort where all patients undergo ABG assessment. However, the representativeness of patients consenting to participate in this research and the difficulty of enrolling enough patients to train robust machine-learning models limit the feasibility of this approach. Alternatively, models can be trained on large, existing databases, with the acknowledgment that the patients who receive arterial blood gases (and thus are available for model training) may be different from those who do not (and thus stand to benefit most from diagnostic modeling). The primary purpose of the current analyses was to assess how problematic this difference might be.

Performance was relatively maintained when the models were applied to a broader set of patients (those receiving either an ABG or VBG) and the re-weighted population approximating if all eligible patients had been equally likely to receive an ABG. In fact, the model's performance improved in the re-weighted sample, likely because it is easier to detect compensated hypercapnia (defined as hypercapnia where the kidneys have been able to retain bicarbonate) while those patients are also less likely to undergo ABG sampling due to subtler symptoms.

The findings of this study provide some preliminary reassurance that much of the variability in the type of patients who receive ABG verification of hypercapnia status does not importantly confound the relationship between model predictions and the true likelihood of hypercapnia. Ultimately, however, prospective validation of the model predictions by assessing CO_2 levels in patients who have not received ABG sampling will be required before clinical or research use is advisable.

A notable strength of the study is the large (over 100,000 patients in the training set) and geographically diverse sample, which guards against overfitting and modeling of local, idiosyncratic practice patterns. Only near-universally available predictors from

the same hospitalization were used in the model to minimize the influence of informed presence bias and the dependence on inter-institution data linkages.

Several additional limitations exist. First, we used relatively simple machine learning approaches to estimate the generalizability of diagnostic predictions to patients who did not receive arterial blood gas outcome sampling, but more advanced methods of imputation, feature selection, and model choice may improve performance. Though no benchmark for comparison exists, neither model is likely sufficient for stand-alone diagnosis or labeling at the current accuracy. The inclusion of unstructured elements (e.g. signs and symptoms) might improve performance enough for this use. The validity of inverse probability weighting analyses depends on several assumptions that do not strictly hold, though the approximation of performance may still be useful. Similarly, we cannot quantify how much of the performance drop when including venous CO_2 as an outcome occurs due to the imperfect relationship of venous to arterial CO_2. Lastly, predictors and outcomes were matched only to the calendar day, and thus transient changes in blood CO_2 levels may be misclassified. However, actionable long-term treatments for hypercapnic respiratory failure require the persistence of hypercapnia, so the performance in patients with stable elevations may be more clinically relevant.

In summary, we show that diagnostic modeling of the likelihood of hypercapnia using routine lab and demographic data is likely sufficient for risk stratification and may perform well on patients who do not currently receive definitive ABG diagnosis.

Acknowledgments. This research was supported by the National Institutes of Health under Ruth L. Kirschstein National Research Service Award 5T32HL105321 from the NIH. (B.W.L.), the American Thoracic Society Academic Sleep Pulmonary Integrated Research/Clinical Fellowship (ASPIRE) Fellowship, UL1TR002538, and UM1TR004409 awards from NIH NCATS (R.G.). The data used in this study was collected on June 2, 2023 from the TriNetX Research Network.

Disclosure of Interests. K.M.S. Sundar is co-founder of Hypnoscure LLC—a software application for population management of sleep apnea through the University of Utah Technology Commercialization Office. J.F. is a general chair for the AIME 2024 conference. All other authors report no conflict of interest.

References

1. Chung, Y., Garden, F.L., Marks, G.B., Vedam, H.: Population Prevalence of Hypercapnic Respiratory Failure from Any Cause. Am. J. Respir. Crit. Care Med. (2022). https://doi.org/10.1164/rccm.202108-1912le.
2. Meservey, A.J., Burton, M.C., Priest, J.S., Teneback, C.C., Dixon, A.E.: Risk of Readmission and Mortality Following Hospitalization with Hypercapnic Respiratory Failure. Lung. (2020). https://doi.org/10.1007/s00408-019-00300-w.
3. Wilson, M.W., Labaki, W.W., Choi, P.J.: Mortality and Healthcare Utilization of Patients with Compensated Hypercapnia. Ann. Am. Thorac. Soc. (2021). https://doi.org/10.1513/annalsats.202009-1197oc.

4. Vonderbank, S., Gibis, N., Schulz, A., Boyko, M., Erbuth, A., Gürleyen, H., Bastian, A.: Hypercapnia at Hospital Admission as a Predictor of Mortality. Open Access Emerg. Med. OAEM. 12, 173–180 (2020). https://doi.org/10.2147/OAEM.S242075.
5. Nowbar, S., Burkart, K.M., Gonzales, R., Fedorowicz, A., Gozansky, W.S., Gaudio, J.C., Taylor, M.R.G., Zwillich, C.W.: Obesity-associated hypoventilation in hospitalized patients: prevalence, effects, and outcome. Am. J. Med. (2004). https://doi.org/10.1016/j.amjmed.2003.08.022.
6. Gonella, S., Clari, M., Conti, A., Simionato, L., Tassone, C., Berchialla, P., Campagna, S.: Interventions to reduce arterial puncture-related pain: A systematic review and meta-analysis. Int. J. Nurs. Stud. (2021). https://doi.org/10.1016/j.ijnurstu.2021.104131.
7. Rowling, S.C., Fløjstrup, M., Henriksen, D.P., Viberg, B., Hallenberg, C., Lindholt, J.S., Alberg-Fløjborg, A., Nanayakkara, P.W., Brabrand, M.: Arterial blood gas analysis: as safe as we think? A multicentre historical cohort study. ERJ Open Res. 8, (2022).
8. Gay, P.C., Owens, R.L.: Executive Summary: Optimal NIV Medicare Access Promotion: A Technical Expert Panel Report From the American College of Chest Physicians, the American Association for Respiratory Care, the American Academy of Sleep Medicine, and the American Thoracic Society. Chest. 160, 1808–1821 (2021). https://doi.org/10.1016/j.chest.2021.05.074.
9. Mokhlesi, B., Masa, J.F., Brozek, J., Gurubhagavatula, I., Murphy, P.B., Piper, A.J., Tulaimat, A., Afshar, M., Balachandran, J.S., Dweik, R.A., Grunstein, R.R., Hart, N., Kaw, R., Lorenzi-Filho, G., Pamidi, S., Patel, B.K., Patil, S.P., Pépin, J.-L., Soghier, I., Kakazu, M.T., Teodorescu, M.: Evaluation and Management of Obesity Hypoventilation Syndrome. An Official American Thoracic Society Clinical Practice Guideline. Am. J. Respir. Crit. Care Med. (2019). https://doi.org/10.1164/rccm.201905-1071st.
10. Locke, B., Gouripeddi, R., Richards, W., Brown, J., Sundar, K.: Test Performance of Serum Bicarbonate in Identifying Hypercapnia Across Settings and Diseases. In: D30. INTEGRATING OSA AND COMORBIDITIES FOR EFFECTIVE THERAPIES. pp. A6495–A6495. American Thoracic Society (2023).
11. Pepe, M.S.: The Statistical Evaluation of Medical Tests for Classification and Prediction. (2003). https://doi.org/10.1198/tech.2005.s278.
12. Palchuk, M., London, J., Pérez-Rey, D., Drebert, Z., Winer-Jones, J., Thompson, C., Esposito, J., Claerhout, B.: A global federated real-world data and analytics platform for research. JAMIA Open. (2023). https://doi.org/10.1093/jamiaopen/ooad035.
13. Usher-Smith, J.A., Sharp, S.J., Griffin, S.J.: The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ. 353, 0 (2016). https://doi.org/10.1136/bmj.i3139.
14. Groenwold, R.H., White, I.R., Donders, A.R.T., Carpenter, J.R., Altman, D.G., Moons, K.G.: Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Cmaj. 184, 1265–1269 (2012).
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.: Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830 (2011).
16. Cerulli, G.: Machine learning using stata/python. Stata J. 22, 772–810 (2022).