Journal of Intensive Medicine 5 (2025) 167–175

Contents lists available at ScienceDirect

Journal of Intensive Medicine

journal homepage: www.elsevier.com/locate/jointm

Original Article

Preparing for future pandemics: Automated intensive care electronic health record data extraction to accelerate clinical insights

Lada Lijović[1,2,*], Harm Jan de Grooth[1], Patrick Thoral[1], Lieuwe Bos[1,2], Zheng Feng[3], Tomislav Radočaj[2], Paul Elbers[1]

[1] *Department of Intensive Care Medicine, Laboratory for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam Public Health, Amsterdam Cardiovascular Science, Amsterdam Institute for Infection and Immunity, Amsterdam UMC, University of Amsterdam, Vrije Universiteit, Amsterdam, The Netherlands*

[2] *Department of Anesthesiology, Intensive Care and Pain Management, Sestre Milosrdnice University Hospital Center, Zagreb, Croatia*

[3] *Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands*

ARTICLE INFO

Managing Editor: Jingling Bao/Zhiyu Wang

Keywords:
Data extraction
Electronic health records
COVID-19
Acute respiratory distress syndrome
Intensive care units

ABSTRACT

Background: Manual data abstraction from electronic health records (EHRs) for research on intensive care patients is time-intensive and challenging, especially during high-pressure periods such as pandemics. Automated data extraction is a potential alternative but may raise quality concerns. This study assessed the feasibility and credibility of automated data extraction during the coronavirus disease 2019 (COVID-19) pandemic.

Methods: We retrieved routinely collected data from the COVID-Predict Dutch Data Warehouse, a multicenter database containing the following data on intensive care patients with COVID-19: demographic, medication, laboratory results, and data from monitoring and life support devices. These data were sourced from EHRs using automated data extraction. We used these data to determine indices of wasted ventilation and their prognostic value and compared our findings to a previously published original study that relied on manual data abstraction largely from the same hospitals.

Results: Using automatically extracted data, we replicated the original study. Among 1515 patients intubated for over 2 days, Harris–Benedict (HB) estimates of dead space fraction increased over time and were higher in non-survivors at each time point: at the start of ventilation (0.70±0.13 *vs.* 0.67±0.15, P <0.001), day 1 (0.74±0.10 *vs.* 0.71±0.11, P<0.001), day 2 (0.77±0.09 *vs.* 0.73±0.11, P<0.001), and day 3 (0.78±0.09 *vs.* 0.74±0.10, P<0.001). Patients with HB dead space fraction above the median had an increased mortality rate of 13.5%, compared to 10.1% in those with values below the median (P<0.005). Ventilatory ratio showed similar trends, with mortality increasing from 10.8% to 12.9% (P=0.040). Conversely, the end-tidal-to-arterial partial pressure of carbon dioxide ($PaCO_2$) ratio was inversely related to mortality, with a lower 28-day mortality in the higher than median group (8.5% *vs.* 15.1%, P<0.001). After adjusting for base risk, impaired ventilation markers showed no significant association with 28-day mortality.

Conclusion: Manual data abstraction from EHRs may be unnecessary for reliable research on intensive care patients, highlighting the feasibility and credibility of automated data extraction as a trustworthy and scalable solution to accelerate clinical insights, especially during future pandemics.

Introduction

Manual data extraction of health records using case forms continues to be recognized as a valid data collection method for retrospective observational research.[1] However, the process typically requires extensive review of patient documentation by highly trained personnel to minimize inaccuracies and inconsistencies.[2,3] It requires significant time commitment and is not always feasible or practical, especially when the data of interest is granular and hospital systems are under pressure, such as during the coronavirus disease 2019 (COVID-19) pandemic.

The advent of electronic health records (EHRs) has enabled the secondary use of routinely collected data for administrative

* Corresponding author: Lada Lijović, Department of Anesthesiology, Intensive Care and Pain Management, Sestre Milosrdnice University Hospital Center, Vinogradska 29, Zagreb 10000, Croatia.
E-mail address: l.lijovic@amsterdamumc.nl (L. Lijović).

https://doi.org/10.1016/j.jointm.2024.10.003
Received 28 July 2024; Received in revised form 5 October 2024; Accepted 14 October 2024
Available online 30 November 2024

and research purposes, potentially reducing the need for manual abstraction. However, automated data extraction also comes with challenges. Significant resources related to information technology and medical expertise may be required to ensure adequate data retrieval and transformation, especially if data from multiple hospitals must be combined. In addition, concerns persist regarding the quality and suitability of data from EHRs for research, as these data may not have been recorded or curated with the same care as research data. Therefore, it remains unclear if automated data extraction can replace manual data extraction in this setting.

During the COVID-19 pandemic, large collaborative groups of investigators in the Netherlands adopted both approaches to study the disease course of patients with severe COVID-19, resulting in acute respiratory distress syndrome (ARDS). The PRactice of VENTilation in patients with COVID–19 (PRoVENT-COVID) collaboration used manual data abstraction, while the COVID-Predict collaboration used automated data collection. Both cohorts included very similar, and sometimes the same, patients receiving comparable care, as they recruited from partly overlapping hospitals within the same country and timeframe. As such, this represents a unique opportunity to evaluate the suitability of automated data collected for observational research in intensive care.

The investigators of PRoVENT-COVID previously reported on the association of estimated ventilatory dead space with 28-day mortality rates in COVID-19 patients. We set out to replicate these results using automatically extracted EHR data from COVID-Predict.[4] In ARDS patients, microvascular changes are the main determinants for increased dead space, which has been consistently associated with poor outcomes.[5–7] In COVID-19-related ARDS, increased ventilatory dead space, linked to widespread thrombosis in the pulmonary circulation, has also been associated with increased mortality.[8–10] The PRoVENT-COVID paper demonstrated associations between dead space and mortality but found that this association attenuated to null after accounting for more conventional risk factors: a clinically counterintuitive finding.

We hypothesized that successfully replicating the results of the PRoVENT-COVID—specifically, the absence of an association between ventilatory dead space (V_d/V_t) and 28-day mortality—using only automated data extraction from EHRs would support the feasibility and credibility of this approach as a valid and safe alternative to manual data abstraction for intensive care research.

Methods

PRoVENT-COVID data

PRoVENT-COVID was a multicenter observational study of COVID-19 patients requiring invasive mechanical ventilation in intensive care units (ICUs) in the Netherlands. Data were collected from 24 ICUs, covering 1000 patients over 10 weeks. The study collected comprehensive data on ventilation duration, mortality, and clinical variables, including for patients transferred between hospitals. Data were transcribed from local electronic patient management systems into an anonymized electronic case report form, with access protected by personalized login credentials. Collected data included baseline demographics, ventilation parameters (mode, tidal volume, airway pressures, and oxygen levels), hemodynamics (blood pressure, heart rate), and daily outcomes (sequential organ failure assessment [SOFA] scores, fluid balance, and use of prone positioning or extracorporeal membrane oxygenation). These data were captured in real-time at specific intervals during the first 3 days after ICU admission, with follow-up data collected on days 7, 28, and 90. The primary outcome focused on ventilatory management during the first 3 days of ICU admission, while secondary outcomes included various ventilation variables, organ function, and the use of rescue therapies. Mortality rates, duration of ICU/hospital stay, and ventilator-free days were also recorded.[11]

COVID-Predict Dutch Data Warehouse

Data for the present study were extracted from the COVID-Predict Dutch Data Warehouse (DDW), a multicenter EHR database with full-admission data from critically ill COVID-19 patients from 25 hospitals. The data were passed through the extract-transform load pipeline, and Structured Query Language (SQL) queries were used to automatically extract data from each major EHR system in the Netherlands. The final database contains 200 million clinical data points from 3464 patients from the first two waves of the COVID-19 pandemic in The Netherlands.[12]

The extraction phase involved creating customized SQL queries to retrieve data from several local EHR systems. These queries were designed to automatically extract a wide range of data, including patient demographics, clinical observations, medications, and vital signs, ensuring consistency across hospitals. To protect patient privacy, the extracted data were pseudonymized at the source using Secure Hash Algorithm (SHA-256) encryption before being transferred in comma-separated value format with end-to-end encryption.

The transformation phase focused on harmonizing and standardizing the data. This involved mapping raw parameters from the different EHR systems to a common concept vocabulary, resolving discrepancies in nomenclature, and ensuring unit standardization across all hospitals. The DDW team created a vocabulary of 942 clinically relevant parameters, supplemented by international standards like Logical Observation Identifiers Names and Codes and Systematized Nomenclature of Medicine Clinical Terms. Hospitals used different terminologies and formats for the same clinical measurements, requiring manual mapping and aggregation into higher-level concepts (e.g., aggregating various temperature measurements into a single “temperature” variable). In addition to basic transformations, the pipeline derived additional clinical parameters, such as ventilatory ratios (VRs) and respiratory system compliance, using predefined algorithms. Complex clinical events like intubation were identified through combinations of multiple data points.

After transformation, the data were loaded into the final database. The structured data were organized into various tables, each corresponding to specific domains such as patient demographics, clinical observations, medications, and outcomes. Data enrichment was also performed during this phase, where derived clinical concepts and scores like the SOFA and Acute Physiology and Chronic Health Evaluation II were calculated from the transformed data.

Validation was a continuous process throughout the extract, transform, and load pipeline. Data quality was assessed at multiple points, including verifying the completeness of the extracted data, ensuring that parameter mappings were correct, and comparing clinical scores to national benchmarks. Distribution plots for each mapped parameter helped identify anomalies, and validation checks were performed to ensure consistency and integrity across all hospitals. In cases of discrepancies, the data were cross-referenced with the original hospital to clarify and resolve issues.

For this study, data were extracted from processed admissions of patients over 18 years old intubated for longer than 2 days with complete information on sex, age, comorbidities, medication, and urine output. Only patients with one episode of invasive ventilation were included. Supplementary Figure S1 shows the cohort extraction process. Demographic data and data regarding comorbidities and medication were extracted at baseline. Following the protocol of the replicated study,[4] we extracted and analyzed ventilator settings and parameter means for the first hour of invasive ventilation and, from continuous data, the means for the first four calendar days.

Exposure variables

The Harris–Benedict (HB) dead space fraction was calculated using the alveolar ventilation Eq. (1):

$$\frac{V_d}{V_t} = 1 - \frac{\left(0.863 \times \dot{V}CO_2\right)}{\left(RR \times V_t \times PaCO_2\right)} \quad (1)$$

where RR is the respiratory rate (breaths per minute), V_d is the dead space volume in liters, V_t is the tidal volume in liters, $PaCO_2$ is the partial pressure of carbon dioxide in millimeter of mercury, and $\dot{V}CO_2$ is the CO_2 production in milliliter per minute calculated from resting energy expenditure (REE) using the rearranged Weir Eq. (2)[13]:

$$\dot{V}CO_2 = \frac{REE_{HB}}{\left(\frac{5.616}{RQ} + 1.584\right)} \quad (2)$$

RQ is the respiratory quotient, assumed to be 0.8, and REE_{HB} is the rest energy expenditure calculated by the unadjusted HB estimate using Eq. (3)[14]:

$$\begin{aligned} \text{Males}: REE_{HB} &= 66.473 + (13.752 \times \text{weight}) + (5.003 \times \text{height}) \\ &\quad - (6.755 \times \text{age}) \\ \text{Females}: REE_{HB} &= 655.096 + (9.563 \times \text{weight}) \\ &\quad + (1.850 \times \text{height}) - (4.676 \times \text{age}) \end{aligned} \quad (3)$$

Weight is the actual body weight in kilograms, height is in centimeters, and age is in years.

Additional estimations of V_d/V_t included a direct estimation and end-tidal-to-arterial $PaCO_2$ ratio. Direct estimation of dead space fraction based on a prediction model derived using least angle regression was calculated using Eq. (4)[15]:

$$\frac{V_d}{V_t} = 0.1726 + (0.0059 \times RR) + 0.0054 \times PEEP + (0.0293 \times LIS) + \left(0.0036 \times PaCO_2 \times V_E\right) + \left(0.000057 \times PaCO_2 \times \text{age}\right) \quad (4)$$

where RR is the respiratory rate in breaths per minute, PEEP is the positive end-expiratory pressure in centimeter of water, LIS is the Murray lung injury score (due to lack of chest X-ray data in the database, calculated without the chest X-ray score), $PaCO_2$ is the partial pressure of carbon dioxide in millimeter of mercury, and V_E is the minute ventilation in liter per minute.[16]

End-tidal-to-arterial $PaCO_2$ ratio was calculated using the Eq. (5):

$$\frac{P_{ET}CO_2}{PaCO_2} \quad (5)$$

VR was calculated using Eq. (6).[17]

$$VR = \frac{\dot{V}_{Emeasured} \times PaCO_{2measured}}{\dot{V}_{Epredicted} \times PaCO_{2predicted}} \quad (6)$$

VR is the ventilatory ratio, $V_{Emeasured}$ is the measured minute ventilation in milliliter per minute, $PaCO_{2measured}$ is the measured $PaCO_2$ in millimeter of mercury, $V_{Epredicted}$ was taken to be 100 mL/(kg·min) extracted based on population nomograms from anesthetic practice multiplied by predicted body weight from ARDSnet predicted body weight (PBW) calculator.[18] $PaCO_{2predicted}$ was taken to be 35 mmHg. All exposure variables used in the analysis were aggregated as the mean for each respective day.

Outcomes

The primary outcome was death at 28 days, defined as the mortality within 28 days after the start of invasive ventilation.

Statistical analysis

The normality of continuous variables was assessed using the Shapiro–Wilk test. Continuous variables were summarized as median (interquartile range) and compared using the Mann–Whitney *U* test. Qualitative variables were summarized using frequencies and percentages and compared using the χ^2 or Fisher's exact test (for frequencies <5).

Following the original study protocol, exposure variables were presented using boxplots for survivors and non-survivors over the first four calendar days. The direction of effect over time of the variables was assessed with mixed-effect linear models with hospital center and patients as a random effect and 28-day vital status (alive/dead), time (as a continuous variable), and the interaction between 28-day mortality and time as a fixed effect. All daily measurements of variables were aggregated as mean per day. To compare variables across days, the variable for each day was entered as a categorical variable in the model, and the *P*-value for the daily differences was obtained by pairwise comparisons with Bonferroni correction.

The risk of death for each tertile of lung-specific physiological variables was used to evaluate whether the predictive ability of each variable varied by levels of the variable. A simple stratification of variables into two groups based on the median of each variable was also assessed. The two groups were compared using Kaplan–Meier curves and log-rank tests.

Univariable mixed-effect generalized linear models, assuming a binomial distribution and with hospital center as a random effect, were used to estimate the unadjusted effect of each variable on 28-day mortality. A multivariable mixed-effect generalized linear model, also assuming a binomial distribution and with hospital center as a random effect, was used to evaluate

the association of each exposure described above with 28-day mortality. The list of candidate confounders was the same as the original publication, and baseline values used were age, sex, body mass index, ratio of arterial oxygen partial pressure (PaO_2 in mmHg) to fractional inspired oxygen (PaO_2/FiO_2), plasma creatinine, hypertension, diabetes, use of angiotensin-converting enzyme (ACE) inhibitors, use of angiotensin II receptor blockers (ARBs), use of vasopressor or inotropic drugs, fluid balance, pH, mean arterial pressure, heart rate, respiratory system compliance, and positive end-expiratory pressure. Multicollinearity was assessed using the variance inflation factors, and the discrimination and calibration of the final model were assessed using c-statistic and Brier scores, respectively.

In addition to the odds ratio and its 95% confidence interval, the discriminative accuracy of the lung-specific physiological variables was measured using the area under the receiver operating characteristics curve (AUC-ROC). The net reclassification improvement and integrated discrimination index were used to assess whether these variables improved predictive accuracy beyond the base model described above.

Missing values for covariates were imputed using multivariate imputation by chained equations (MICE). All analyses were conducted using Python and its scikit-learn and SciPy packages.

Results

Study population

A total of 3203 patients from processed admissions were screened. After excluding non-intubated patients, patients lacking age, sex, and complete baseline characteristics information, patients intubated for <2 days, and those with multiple episodes, the final cohort included 1515 patients: 1157 (76.4%) survivors and 358 (23.6%) non-survivors (Supplementary Figure S1). Missing respiratory data ranged from 1.5% for tidal volume to 17.7% for mechanical power and were imputed using MICE imputation.

Table 1 shows the baseline characteristics of the study participants. Non-survivors and survivors differed by age, sex, body mass index, severity of ARDS, presence of diabetes and immunosuppression, use of ACE inhibitors, baseline creatinine, treatment with continuous sedation, ionotropic or vasopressor support, and mean arterial pressure (all $P<0.05$).

Baseline ventilatory variables

Table 2 presents the baseline ventilatory variables. Survivors and non-survivors differed in baseline PaO_2/FiO_2 values ($P=0.045$) but not in baseline end-tidal carbon dioxide value (EtCO2) ($P=0.119$).

Temporal changes in estimated ventilatory dead space

The calculated dead space values over the first four days of ventilation are shown in Table 3 and Figure 1. Dead space fraction calculated using the HB formula was consistently higher in non-survivors and increased over time (all $P<0.001$). Direct dead space fraction and VR did not differ between survivors and non-survivors early during mechanical ventilation but increased and diverged over time. The end-tidal-to-arterial $PaCO_2$ ratio

Table 1
Baseline characteristics of patients by 28-day mortality.

Variables	Non-survivors (n=358)	Survivors (n=1157)	P-value
Age (years)	70 (64–75)	63 (56–71)	< 0.001
Sex (male)	281 (78.5)	834 (72.1)	0.020
Body mass index (kg/m^2)	28 (26–30)	29 (27–31)	0.001
Severity of ARDS	0.020		
Mild	42 (11.7)	208 (18.0)	
Moderate	243 (67.8)	735 (63.5)	
Severe	73 (20.4)	214 (18.5)	
Co-existing disorders			
Hypertension	13 (3.6)	45 (3.9)	0.948
Heart failure	9 (2.5)	16 (1.4)	0.227
Diabetes	106 (29.6)	264 (22.8)	0.010
Chronic kidney disease	22 (6.1)	41 (3.5)	0.053
Baseline creatinine (μmol/L)	91 (74–129)	80 (65–103)	< 0.001
Liver cirrhosis	0 (0.0)	52 (0.9)	0.596
Chronic obstructive pulmonary disease	38 (10.6)	94 (8.1)	0.192
Active hematological neoplasia	12 (3.3)	20 (1.7)	0.105
Immunosuppression	43 (12.0)	91 (7.8)	0.025
Previous medication			
Systemic steroids	196 (54.7)	600 (51.8)	0.370
Inhalation steroids	2 (0.0)	20 (1.7)	0.172
ACE inhibitor	56 (15.6)	286 (24.7)	< 0.001
ARB	8 (0.2)	35 (3.0)	0.545
Vital signs			
Heart rate (bpm)	87 (76–102)	90 (77–102)	0.485
Mean arterial pressure (mmHg)	83 (76–95)	87 (77–100)	0.005
Organ support			
Continuous sedation	324 (90.5)	991 (85.6)	0.023
Ionotropic or vasopressor	344 (96.1)	1032 (89.2)	< 0.001
Vasopressor	343 (95.8)	1031 (89.1)	<0.001
Ionotropic	43 (12.0)	40 (3.5)	< 0.001
Fluid balance (mL)	523 (76–806)	840 (77–428)	0.022
Urine output (mL)	1357 (819–3245)	1489 (905–3430)	0.047

Data are presented as median (interquartile range) or n (%).
ACE: Angiotensin-converting enzyme; ARBs: Angiotensin II receptor blockers; ARDS: Acute respiratory distress syndrome.

Table 2
Respiratory variables at the start of ventilation by 28-day mortality.

Variables	All patients (*n*=1515)	Non-survivors (*n*=358)	Survivors (*n*=1157)	*P*-value
Tidal volume (mL/kg PBW)	6.4 (5.7–7.4)	6.4 (5.7–7.2)	6.4 (5.7–7.4)	0.253
PEEP (cmH_2O)	12 (10–14)	12 (10–14)	12 (10–14)	0.185
Driving pressure (cmH_2O)	13 (11–15)	13 (10–16)	13 (11–15)	0.721
PaO_2/FiO_2	141 (109–176)	139 (106–162)	142 (110–183)	0.045
$EtCO_2$ (mmHg)	35 (30–41)	34 (30–41)	35 (30–41)	0.119
Mechanical power (J/min)	27 (19–37)	28 (19–39)	27 (18–37)	0.149
Compliance (mL/cmH_2O)	36 (27–44)	34 (26–43)	36 (28–44)	0.087

Data are presented as median (interquartile range).
$EtCO_2$: End-tidal carbon dioxide value; PaO_2/FiO_2: Ratio of arterial oxygen partial pressure (PaO_2 in mmHg) to fractional inspired oxygen; PBW: Predicted body weight; PEEP: Positive end-expiratory pressure.

Table 3
Lung-specific physiological variables in the first four days of ventilation by 28-day mortality.

Variables	All patients (*n*=1515)	Non-survivors (*n*=358)	Survivors (*n*=1157)	*P*-value
Dead space fraction by HB				
At start of ventilation	0.68 (0.60-0.74)	0.70 (0.64-0.77)	0.67 (0.59-0.74)	< 0.001
Day 01	0.72 (0.66-0.77)	0.74 (0.69-0.78)	0.71 (0.65-0.76)	< 0.001
Day 02	0.74 (0.68-0.79)	0.77 (0.71-0.80)	0.73 (0.67-0.78)	< 0.001
Day 03	0.75 (0.70-0.79)	0.78 (0.73-0.81)	0.74 (0.68-0.79)	< 0.001
P-value (interaction survival × day)	0.223			
Dead space fraction direct				
At start of ventilation	2.49 (2.10-2.96)	2.54 (2.10-3.11)	2.48 (2.09-2.93)	0.238
Day 01	2.60 (2.22-3.11)	2.70 (2.26-3.17)	2.58 (2.21-3.07)	0.121
Day 02	2.75 (2.29-3.32)	3.18 (2.45-3.67)	2.64 (2.23-3.17)	< 0.001
Day 03	2.80 (2.34-3.40)	3.00 (2.42-3.76)	2.72 (2.30-3.34)	0.003
P-value (interaction survival × day)	<0.001			
VR				
At start of ventilation	1.63 (1.29-2.09)	1.69 (1.30-2.14)	1.62 (1.23-2.05)	0.081
Day 01	1.80 (1.49-2.19)	1.92 (1.62-2.32)	1.76 (1.45-2.15)	< 0.001
Day 02	1.92 (1.59-2.35)	2.06 (1.73-2.54)	1.87 (1.55-2.28)	< 0.001
Day 03	2.00 (1.69-2.41)	2.18 (1.82-2.62)	1.96 (1.65-2.32)	< 0.001
P-value (interaction survival × day)	<0.001			
End-tidal to-arterial $PaCO_2$ ratio				
At start of ventilation	0.81 (0.69-0.93)	0.76 (0.65-0.88)	0.83 (0.70-0.94)	< 0.001
Day 01	0.80 (0.70-0.89)	0.77 (0.67-0.85)	0.82 (0.72-0.90)	< 0.001
Day 02	0.79 (0.68-0.88)	0.73 (0.65-0.82)	0.80 (0.70-0.89)	< 0.001
Day 03	0.78 (0.68-0.86)	0.73 (0.65-0.82)	0.79 (0.70-0.88)	< 0.001
P-value (interaction survival × day)	0.143			

Data are presented as median (interquartile range)
HB: Harris–Benedict; $PaCO_2$: Partial pressure of carbon dioxide; VR: Ventilatory ratio.

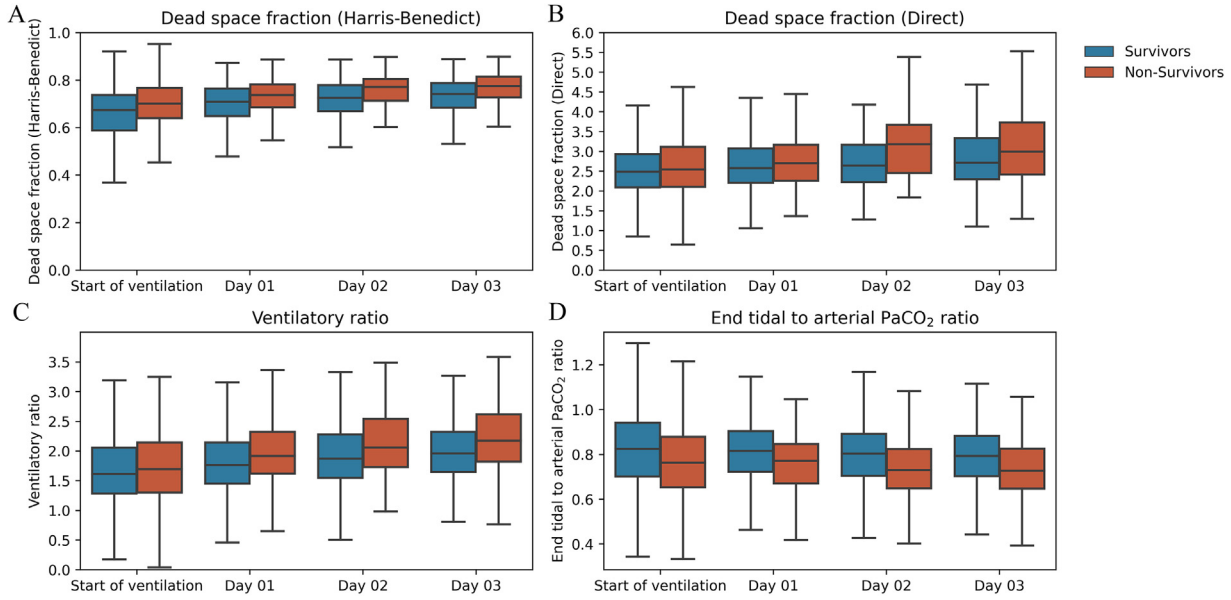

Figure 1. Lung-specific physiological variables over the first four days of ventilation, stratified by survival outcome. Dead space fraction calculated using (A) Harris-Benedict formula, (B) direct dead space fraction, (C) Ventilatory ratio and (D) end-tidal-to-arterial $PaCO_2$ ratio.
$PaCO_2$: Partial pressure of carbon dioxide.

showed similar trends to dead space fraction, with lower values in non-survivors and further decreases over time (all $P<0.001$).

Impact of ventilatory variables on 28-day mortality

Mortalities by tertiles of each variable are presented in Figure 2 and Table 4. Tertiles were calculated separately for each variable and each day to account for potential differences in scaling and measurements. Mortality increased with successive tertiles of dead space fraction based on the HB formula and of VR and decreased with successive tertiles of end-tidal-to-arterial $PaCO_2$ ratio. When considering lung-specific variables measured at the start of ventilation, and groups created according to the median of the variables at start of ventilation, 28-day mortality was higher in patients in the high group of dead space fraction estimated using HB estimation (10.1% *vs.* 13.5%, $P<0.005$), but did not differ by direct dead space fraction (13.5% *vs.* 10.0%, $P=0.550$). The log-rank test showed a slight difference in survival VR (10.8% *vs.* 12.9%, $P=0.040$). Assessment of end-tidal-to-arterial $PaCO_2$ ratio showed lower 28-day mortality (8.5% *vs.* 15.1%, $P<0.001$). The results are presented in Figure 3.

The unadjusted impact of each marker of impaired ventilation is shown in Supplementary Table S1. None of the lung-specific baseline variables were associated with 28-day mortality. In the univariate model, the following variables were associated with 28-day mortality: direct dead fraction, VR and end-tidal-to-arterial $PaCO_2$, age, use of ACE inhibitors and ARBs, initial pH, and fluid balance on day 1.

After adjustment for the base risk model, none of the markers of impaired ventilation measured at the start of ventilation or the following day was significantly associated with 28-day mortality (Table 5). The inclusion of these variables did not improve the AUC-ROC compared to the base model (Figure 4).

Discussion

This study demonstrated that findings on the association between ventilatory dead space and mortality in COVID-19-related ARDS obtained using manual data abstraction can be replicated using automatically extracted data. Consistent with the original paper, our results confirm that estimates of dead space fraction increased over time and were higher in non-survivors than survivors. We also confirmed the absence of a significant association between V_d/V_t and 28-day mortality after controlling for potential confounding factors.

We chose the original study as our target for replication because the lack of an independent association between estimated ventilatory dead space and mortality was unexpected. Since these findings are clinically counterintuitive, any flaws in the

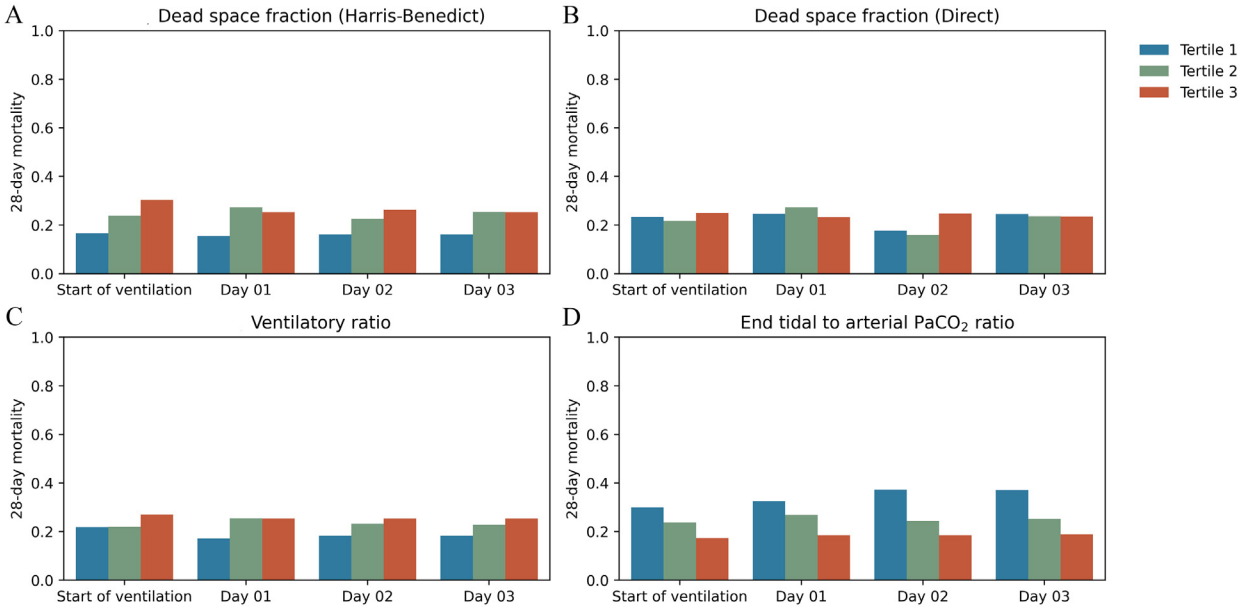

Figure 2. 28-Day mortality according to tertiles of lung-specific physiological variables over the first four days of ventilation. Dead space fraction calculated using (A) Harris-Benedict formula, (B) direct dead space fraction, (C) ventilatory ratio and (D) end-tidal-to-arterial $PaCO_2$ ratio.
$PaCO_2$: Partial pressure of carbon dioxide.

Table 4
Tertile cut-off values for ventilatory dead space according to 28-day mortality.

Time	Dead space fraction (HB) tertiles	Dead space fraction (direct) tertiles	VR tertiles	End-tidal-to-arterial $PaCO_2$ ratio tertiles
Start of ventilation	< 0.63, 0.63–0.72, > 0.72	< 2.22, 2.22–2.76, > 2.76	< 1.40, 1.40–1.88, > 1.88	< 0.73, 0.73–0.88, > 0.88
Day 1	< 0.68, 0.68–0.75, > 0.75	< 2.35, 2.35–2.90, > 2.90	< 1.60, 1.60–2.05, > 2.05	< 0.74, 0.74–0.85, > 0.85
Day 2	< 0.70, 0.70–0.77, > 0.77	< 2.45, 2.45–3.10, > 3.10	< 1.71, 1.71–2.15, > 2.15	< 0.72, 0.72–0.83, > 0.83
Day 3	< 0.71, 0.71–0.78, > 0.78	< 2.49, 2.49–3.18, > 3.18	< 1.80, 1.80–2.25, > 2.25	< 0.71, 0.71–0.83, > 0.83

HB: Harris–Benedict; $PaCO_2$: Partial pressure of carbon dioxide; VR: Ventilatory ratio.

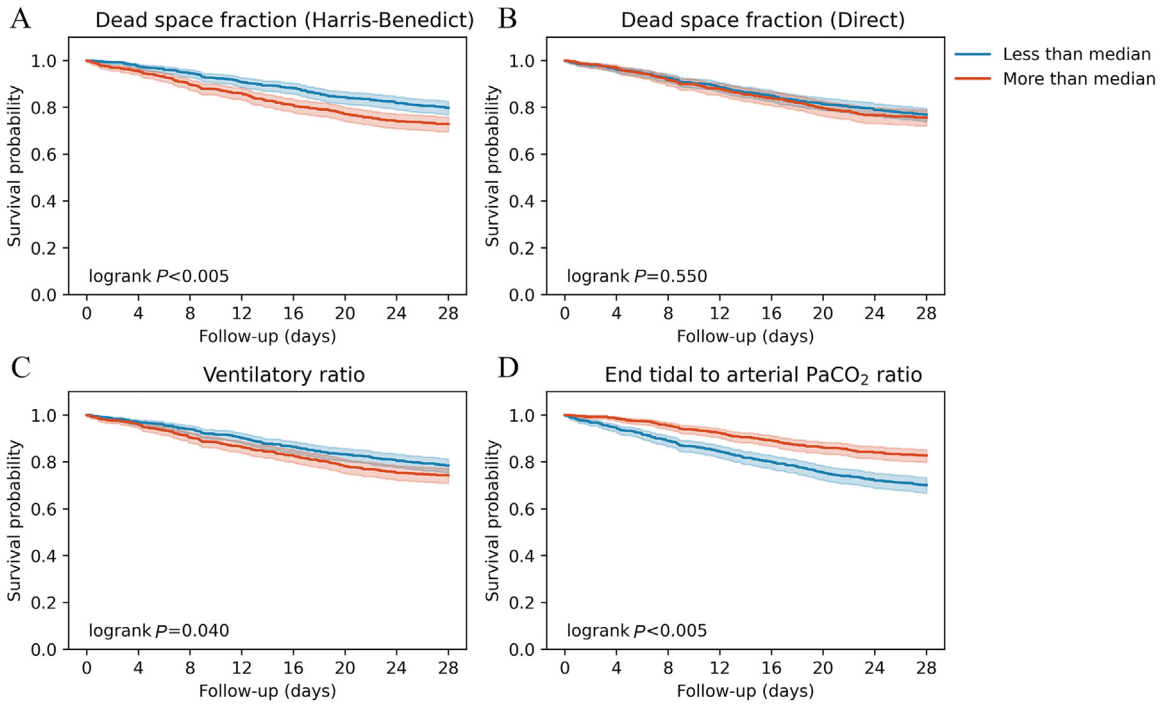

Figure 3. 28-day mortality according to lung-specific variables measures at the start of ventilation. Groups were created according to the median of the variables at start of ventilation; *P* values are from log-rank test. Dead space fraction calculated using (A) Harris-Benedict formula, (B) direct dead space fraction, (C) Ventilatory ratio and (D) end-tidal-to-arterial $PaCO_2$ ratio.
$PaCO_2$: Partial pressure of carbon dioxide.

Table 5
Predictive accuracy of lung-specific physiological variables.

Variables	OR (95% CI)*	*P*-value	AUC (95% CI)	Brier score	NRI (95% CI)	*P*-value	IDI (95% CI)	*P*-value
Base model	/	/	0.795 (0.724 to 0.863)	/	/	/	/	/
At start of ventilation								
+Dead space fraction by HB	0.95 (0.89 to 1.02)	0.157	0.802 (0.729 to 0.863)	0.166	0.010 (−0.065 to 0.081)	0.487	0.007 (−0.132 to 0.144)	0.470
+Direct dead space	1.01 (0.92 to 1.11)	0.787	0.800 (0.731 to 0.865)	0.166	0.011 (−0.065 to 0.086)	0.473	−0.002 (−0.128 to 0.124)	0.467
+VR	1.16 (0.94 to 1.44)	0.167	0.810 (0.732 to 0.874)	0.164	0.011 (−0.065 to 0.919)	0.463	0.001 (−0.131 to 0.134)	0.451
+End-tidal-to-arterial $PaCO_2$	1.01 (0.94 to 1.09)	0.738	0.810 (0.741 to 0.873)	0.164	0.005 (−0.070 to 0.081)	0.485	0.000 (−0.131 to 0.122)	0.500
Day 01								
+Dead space fraction by HB	1.01 (0.94 to 1.08)	0.731	0.739 (0.730 to 0.860)	0.168	0.010 (−0.070 to 0.092)	0.449	−0.001 (−0.130 to 0.134)	0.493
+Direct dead space	1.07 (0.99 to 1.17)	0.096	0.801 (0.730 to 0.870)	0.165	0.005 (−0.070 to 0.081)	0.467	0.008 (−0.123 to 0.125)	0.474
+VR	1.10 (0.95 to 1.29)	0.182	0.801 (0.725 to 0.862)	0.164	NA	NA	−0.000 (−0.118 to 0.135)	0.503
+End-tidal-to-arterial $PaCO_2$	0.94 (0.87 to 1.02)	0.126	0.813 (0.743 to 0.876)	0.161	0.010 (−0.065 to 0.097)	0.475	0.012 (−0.132 to 0.134)	0.444

AUC: Area under the curve; CI: Confidence interval; HB: Harris–Benedict; IDI: Integrated discrimination index; NA: Not applicable; NRI: Net reclassification index; OR: Odds ratio; $PaCO_2$: Partial pressure of carbon dioxide; VR: Ventilatory ratio.

* Represents the OR for the lung-specific physiological variables in the multivariable model.All models are mixed-effect models with centers as random effect and considering a binominal distribution.All continuous variables were entered after standardization to improve the convergence of the model, and OR represents the increase in one standard deviation of the variable.

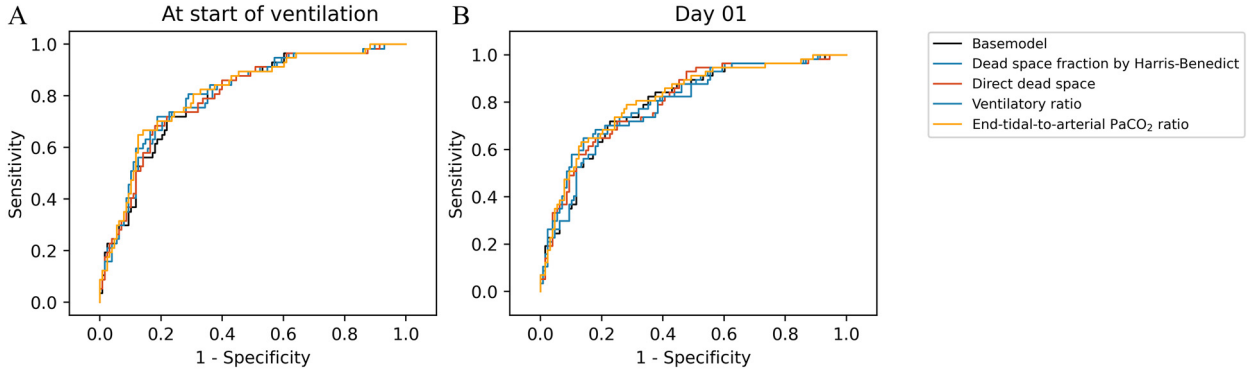

Figure 4. ROC curve of the base model and with the inclusion of lung-specific physiological variables at (A) start of ventilation and (B) first day of mechanical ventilation.
$PaCO_2$: Partial pressure of carbon dioxide; ROC: Receiver operating characteristics.

study design or errors during manual data abstraction would yield different results upon replication with automated data extraction. However, as our results align with the original study, this supports the feasibility and credibility of automated data extraction for intensive care research.

ICUs are among the most data-rich environments in healthcare chiefly because devices for monitoring and life support generate tens of thousands of discrete data points per patient per day.[19] It is therefore unsurprising that, in recent years, automatically extracted data has significantly contributed to advancing knowledge and clinical decision-making in intensive care medicine.[20,21] However, the reliability of automatically extracted data compared to data manually abstracted by trained clinicians has been studied only sporadically, and the prevailing view remains that manually abstracted data are superior in quality and accuracy.

Conversely, previous studies suggest that typing errors, mental lapses, subjectivity, distractions, and fatigue can lead to inaccuracies in manual data abstraction.[22–24] In addition, manual data abstraction is resources-intensive, typically limiting it to a small subset of available data elements. This constraint implies that key features for model development may be underutilized or wasted, especially when high-frequency data are considered. For instance, in the original study, ventilator settings and parameters were collected 1 h after initiating invasive ventilation and subsequently every 8 h, resulting in a model built from three daily data aggregations. In contrast, automated data collection allowed us to aggregate our predictive variables from over 500,000 data points without being constrained by practical time points.

It should be noted that our approach also comes with limitations. First, while the patient cohorts in the original study and our study were overlapping, they were not identical. In addition, automated data extraction involves significant technical, legal, and privacy challenges requiring extensive resources and associated costs, especially when data are sourced from multiple ICUs using different EHRs. This process requires complete data extract, transform, and load pipelines and strict procedures on which parameters to extract with what granularity, managing irregularly recorded data elements, missing data, and outliers. Multiple factors affecting data quality due to data transform pipeline issues have been described, including data characteristics and management issues, personnel training and experience, infrastructure availability, data complexity, cleaning practices, and code quality.[25] However, once such data infrastructure has been established, reusing routinely collected data at scale is easier, saving time, resources, and costs associated with repeated manual data abstraction. This is particularly relevant for accelerating clinical insights in intensive care medicine within the context of pandemic preparedness.

Conclusions

The concordance of our results with those of the original study adds credibility to the notion that automatically extracted data from EHRs can serve as a high-quality, reliable, and faster resource, circumventing the need for manual data collection and curation. This approach saves time and resources, ultimately improving care and outcomes for critically ill patients.

CRediT Authorship Contribution Statement

Lada LijoviLijović: Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis, Data curation. **Harm Jan de Grooth:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Patrick Thoral:** Writing – review & editing, Resources, Project administration. **Lieuwe Bos:** Writing – review & editing, Resources. **Zheng Feng:** Investigation, Formal analysis, Data curation. **Tomislav Radočaj:** Writing – review & editing, Supervision. **Paul Elbers:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Conceptualization.

Acknowledgments

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Statement

Not applicable.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data sets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jointm.2024.10.003.

References

[1] Yin AL, Guo WL, Sholle ET, Rajan M, Alshak MN, Choi JJ, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. Int J Med Inform 2022;157:104622. doi:10.1016/j.ijmedinf.2021.104622.

[2] Brazeal JG, Alekseyenko AV, Li H. Assessing quality and agreement of structured data in automatic versus manual abstraction of the electronic health record for a clinical epidemiology study. Res Methods Med Health Sci 2021;2:168–78. doi:10.1177/26320843211061287.

[3] Martin S, Wagner J, Lupulescu-Mann N, Ramsey K, Cohen A, Graven P, et al. Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. Appl Clin Inform 2017;8:794–809. doi:10.4338/ACI-2016-12-RA-0210.

[4] Morales-Quinteros L, Neto AS, Artigas A, Blanch L, Botta M, Kaufman DA, et al. Dead space estimates may not be independently associated with 28-day mortality in COVID-19 ARDS. Crit Care 2021;25:171. doi:10.1186/s13054-021-03570-0.

[5] Nuckton TJ, Alonso JA, Kallet RH, Daniel BM, Pittet JF, Eisner MD, et al. Pulmonary dead-space fraction as a risk factor for death in the acute respiratory distress syndrome. N Engl J Med 2002;346:1281–6. doi:10.1056/NEJMoa012835.

[6] Villar J, Szakmany T, Grasselli G, Camporota L. Redefining ARDS: a paradigm shift. Crit Care 2023;27:416. doi:10.1186/s13054-023-04699-w.
[7] Cepkova M, Kapur V, Ren X, Quinn T, Zhuo H, Foster E, et al. Pulmonary dead space fraction and pulmonary artery systolic pressure as early predictors of clinical outcome in acute lung injury. Chest 2007;132:836–42. doi:10.1378/chest.07-0409.
[8] Graf J, Pérez R, López R. Increased respiratory dead space could associate with coagulation activation and poor outcomes in COVID-19 ARDS. J Crit Care 2022;71:154095. doi:10.1016/j.jcrc.2022.154095.
[9] Ackermann M, Verleden SE, Kuehnel M, Haverich A, Welte T, Laenger F, et al. Pulmonary vascular endothelialitis, thrombosis, and angiogenesis in COVID-19. N Engl J Med 2020;383(2):120–8. doi:10.1056/NEJMoa2015432.
[10] Torres A, Motos A, Riera J, Fernández-Barat L, Ceccato A, Pérez-Arnal R, et al. The evolution of the ventilatory ratio is a prognostic factor in mechanically ventilated COVID-19 ARDS patients. Crit Care 2021;25:331. doi:10.1186/s13054-021-03727-x.
[11] Boers NS, Botta M, Tsonas AM, Algera AG, Pillay J, Dongelmans DA, et al. PRactice of VENTilation in patients with novel coronavirus disease (PRoVENT-COVID): rationale and protocol for a national multicenter observational study in The Netherlands. Ann Transl Med 2020;8:1251. doi:10.21037/atm-20-5107.
[12] Fleuren LM, Dam TA, Tonutti M, de Bruin DP, Lalisang RCA, Gommers D, et al. The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients. Crit Care 2021;25:304. doi:10.1186/s13054-021-03733-z.
[13] Weir JB. New methods for calculating metabolic rate with special reference to protein metabolism. J Physiol 1949;109:1–9. doi:10.1113/jphysiol.1949.sp004363.
[14] Harris JA, Benedict FG. A biometric study of human basal metabolism. Proc Natl Acad Sci U S A 1918;4:370–3. doi:10.1073/pnas.4.12.370.
[15] Beitler JR, Thompson BT, Matthay MA, Talmor D, Liu KD, Zhuo H, et al. Estimating dead-space fraction for secondary analyses of acute respiratory distress syndrome clinical trials. Crit Care Med 2015;43:1026–35. doi:10.1097/CCM.0000000000000921.
[16] Murray JF, Matthay MA, Luce JM, Flick MR. An expanded definition of the adult respiratory distress syndrome. Am Rev Respir Dis 1988;138:720–3. doi:10.1164/ajrccm/138.3.720.
[17] Sinha P, Fauvel NJ, Singh S, Soni N. Ventilatory ratio: a simple bedside measure of ventilation. Br J Anaesth 2009;102:692–7. doi:10.1093/bja/aep054.
[18] Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, et al. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med 2000;342(18):1301–8. doi:10.1056/NEJM200005043421801.
[19] Schenck EJ, Hoffman KL, Cusick M, Kabariti J, Sholle ET, Campion TR Jr. Critical carE Database for Advanced Research (CEDAR): an automated method to support intensive care units with electronic health record data. J Biomed Inform 2021;118:103789. doi:10.1016/j.jbi.2021.103789.
[20] Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA 2019;321:2003–17. doi:10.1001/jama.2019.5791.
[21] Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. Lancet Respir Med 2014;2:611–20. doi:10.1016/S2213-2600(14)70097-9.
[22] Vawdrey DK, Gardner RM, Evans RS, Orme JF Jr, Clemmer TP, Greenway L, et al. Assessing data quality in manual entry of ventilator settings. J Am Med Inform Assoc 2007;14:295–303. doi:10.1197/jamia.M2219.
[23] Zozus MN, Pieper C, Johnson CM, Johnson TR, Franklin A, Smith J, et al. Factors affecting accuracy of data abstracted from medical records. PLoS One 2015;10:e0138649. doi:10.1371/journal.pone.0138649.
[24] Feng JE, Anoushiravani AA, Tesoriero PJ, Ani L, Meftah M, Schwarzkopf R, et al. Transcription error rates in retrospective chart reviews. Orthopedics 2020;43:e404–8. doi:10.3928/01477447-20200619-10.
[25] Foidl H, Golendukhina V, Ramler R, Felderer M. Data pipeline quality: influencing factors, root causes of data-related issues, and processing problem areas for developers. J Syst Soft 2024;207:111855. doi:10.1016/j.jss.2023.111855.