

## A Pre-trained Clinical Language Model for Predicting Neonatal Morbidities: Development, Validation, and Multicentre Evaluation of NeonatalBERT

Youssef Hariri<sup>1\*</sup>, Hassan Nasser<sup>1</sup>, Fatima Zahra<sup>2</sup>

<sup>1</sup> Department of Neonatology and Clinical NLP, Faculty of Medicine, Mohammed V University, Rabat, Morocco.

<sup>2</sup> Department of Predictive Modeling and Neonatal Outcomes, Faculty of Medicine, University of Casablanca, Casablanca, Morocco.

\*E-mail ✉ youssef.hariri@um5.ac.ma

### Abstract

Prompt recognition and continuous monitoring of health issues in newborns are vital to enable rapid therapeutic responses, reduce potential complications, improve clinical resource allocation, and provide essential support to families. While established methods that rely on structured tabular information and laboratory markers have proven useful, they remain limited in assessing neonatal morbidity risk. This study presents NeonatalBERT, a pre-trained large language model (LLM) specifically designed to assess the likelihood of neonatal morbidities from clinical documentation. This prognostic analysis examined retrospective primary and external cohorts sourced from two major quaternary academic medical institutions in the United States: Stanford Health Care and Beth Israel Deaconess Medical Center. NeonatalBERT underwent initial pre-training using clinical notes from the primary cohort, followed by separate fine-tuning procedures tailored to each cohort. We compared the model's performance with other established LLMs, including BioBERT and Bio-ClinicalBERT, as well as standard machine learning techniques and tabular-variable logistic regression models. NeonatalBERT was tested on 19 neonatal morbidities (respiratory distress syndrome, bronchopulmonary dysplasia, pulmonary haemorrhage, pulmonary hypertension, atelectasis, aspiration syndrome, intraventricular haemorrhage, periventricular leukomalacia, neonatal seizures, other CNS disorders, patent ductus arteriosus, cardiovascular instability, sepsis, candidiasis, anaemia, jaundice, necrotising enterocolitis, retinopathy of prematurity, and death) within the primary cohort, and on ten morbidities in the external cohort (respiratory distress syndrome, bronchopulmonary dysplasia, pulmonary haemorrhage, intraventricular haemorrhage, patent ductus arteriosus, sepsis, jaundice, necrotising enterocolitis, retinopathy of prematurity, and death). For every outcome, evaluation metrics included the area under the receiver operating characteristic curve, area under the precision-recall curve (AUPRC), and F1 scores. A total of 32 321 newborns formed the primary cohort, consisting of 27 411 in the primary training set (mean gestational age 38.64 weeks [SD: 2.30]; 13 056 [47.6%] female and 14 355 [52.4%] male) and 4910 in the primary testing set (mean gestational age 38.64 [2.13] weeks; 2336 [47.6%] female and 2574 [52.4%] male). The external cohort comprised 7061 newborns, including 5653 in the external training set (1567 [27.7%] premature and 4086 [72.3%] term births; 2614 [46.2%] female and 3039 [53.8%] male) and 1408 in the external testing set (383 [27.2%] premature and 1025 [72.8%] term births; 624 [44.3%] female and 784 [55.7%] male). In the primary cohort, NeonatalBERT attained a mean AUPRC of 0.291 (95% CI: 0.268–0.314) across the 19 outcomes, compared with 0.238 (0.217–0.259) for Bio-ClinicalBERT, 0.217 (0.197–0.236) for BioBERT, and 0.194 (0.177–0.211) for the conventional tabular model. In the external validation cohort, NeonatalBERT achieved a mean AUPRC of 0.360 (0.328–0.393), exceeding those of all other models (range 0.224–0.333). In summary, the present work demonstrates that a neonatal-focused large language model can harness routinely recorded clinical notes to generate risk estimates for a wide range of neonatal morbidities across two distinct cohorts. These results support adopting automated risk assessment to enable earlier therapeutic action, optimize clinical processes, and strengthen interactions with families. At the same time, they underscore the necessity for careful, responsible deployment, ongoing validation, and appropriate oversight of artificial intelligence tools in neonatal medicine.

**Keywords:** Neonatal medicine, Neonatal morbidity, Large language models, Clinical natural language processing, Clinical notes, Prognostic modeling

Access this article online

<https://smerpub.com/>

Received: 03 May 2024; Accepted: 04 June 2024

Copyright CC BY-NC-SA 4.0

**How to cite this article:** Hariri Y, Nasser H, Zahra F. A Pre-trained Clinical Language Model for Predicting Neonatal Morbidities: Development, Validation, and Multicentre Evaluation of NeonatalBERT. J Med Sci Interdiscip Res. 2024;4(1):141-54. <https://doi.org/10.51847/3p2YKqX4Ji>

### Introduction

The neonatal phase constitutes a critical stage of human development marked by profound physiological transitions as infants adapt to independent life beyond the uterus. Although healthcare has advanced significantly, more than 1 million neonatal deaths still occur worldwide

each year [1]. Timely prediction of neonatal disorders is central to facilitating life-saving measures. Several validated instruments exist for estimating risks of common prematurity-associated conditions [2], such as death, sepsis, neurodevelopmental challenges, hyperbilirubinemia, and bronchopulmonary dysplasia.

#### *Research in context*

##### *Evidence before this study*

A literature search was performed on PubMed for publications between 1 January 2015 and 1 June 2025, employing the keywords (“neonatal” OR “newborn”) AND (“predict” OR “prediction” OR “estimate”) AND (“machine learning” OR “regression” OR “deep learning” OR “transformer” OR “language model”), with no restrictions on language. Previous work has shown that conventional machine learning approaches, which predominantly use structured datasets, have limited predictive power because they fail to capture the rich contextual details in free-text clinical notes. Although transformer models and BERT variants have improved performance across general medical applications, no prior research has constructed or tested a foundational model pre-trained exclusively on neonatal clinical data. Additionally, few investigations have addressed large-scale morbidity prediction in newborns using unstructured notes from multiple hospital settings.

##### *Added value of this study*

The current work introduces NeonatalBERT, a specialized language model pre-trained on neonatal clinical documentation. Distinct from broader clinical language models, NeonatalBERT integrates subtle neonatal terminology and characteristic clinical patterns. Its probabilistic predictions support early identification of morbidity risks, including those not yet clinically manifest, thereby aiding prompt clinical action and decision-making. Assessments performed across both internal and external cohorts confirm the practical value of unstructured clinical notes for forecasting conditions such as bronchopulmonary dysplasia, necrotizing enterocolitis, sepsis, and mortality.

##### *Implications of all the available evidence*

NeonatalBERT supports the early identification of potential health complications in newborns by leveraging free-text clinical documentation, providing clinicians with timely information to guide therapeutic decisions. Although the model’s accuracy differs depending on the

specific condition — providing more immediate predictions for disorders apparent at delivery and forward-looking estimates for those that emerge subsequently — its reliable results in both internal and external datasets highlight its ability to serve as a valuable addition to traditional assessment methods, particularly when comprehensive structured information is lacking.

Recent developments in machine learning and artificial intelligence (AI), alongside the increasing availability of comprehensive electronic health records (EHRs), have considerably strengthened clinicians’ capacity to anticipate serious conditions affecting newborns and to implement suitable interventions [3]. In the past, clinicians have relied primarily on structured EHR components, such as laboratory findings, diagnostic procedures, and physiological parameters, to recognize neonatal problems [2]. Nevertheless, the detailed insights embedded in unstructured clinical notes have received limited attention in neonatal practice. These narrative records contain rich, specialized information, including direct clinician impressions and exchanges between healthcare team members, elements typically absent from formatted datasets.

Free-text clinical notes represent an important reservoir of clinical information; however, exploiting them for predictive modeling is difficult because natural language is inconsistent and requires complex interpretation. Significant progress in natural language processing, notably through contextual embeddings and large language models (LLMs) [4], has produced strong results in medical text comprehension [5]. Foundational models such as Bidirectional Encoder Representations from Transformers (BERT) [6] demonstrate superior language-processing capabilities that support a wide range of predictive text applications. Domain-adapted versions, including BioBERT [7] for biomedical corpora and Bio-ClinicalBERT [8] for clinical records, have been created. While Bio-ClinicalBERT has been applied successfully to adult disease forecasting [9], neonatal disorders differ substantially from adult pathologies. As a result, the vocabulary and descriptive styles used in newborn care exhibit unique linguistic properties distinct from those found in adult medical documentation.

This investigation introduces NeonatalBERT, a specialized BERT-based architecture optimized for extracting predictive signals from unstructured neonatal clinical notes. By adapting language models to the neonatal domain, NeonatalBERT demonstrates the

advantages of context-specific pre-training for advancing the early recognition of morbidities in newborns.

## Materials and Methods

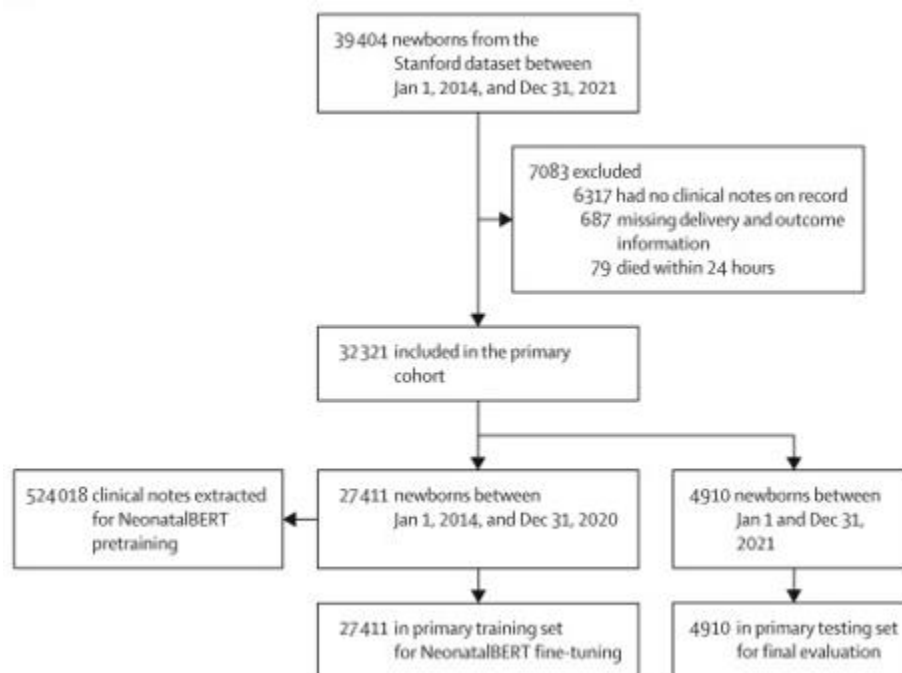
### *Study design and cohort*

This prognostic research employed retrospective electronic health record (EHR) data collected during standard newborn care at two major academic medical institutions in the United States. The Stanford University Institutional Review Board reviewed and approved the study protocol (approval number 39225) and determined that informed consent was unnecessary for this secondary analysis of existing data. Reporting conformed to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) standards [10]. Patients and the public had no role in planning, implementing, interpreting, or sharing study findings.

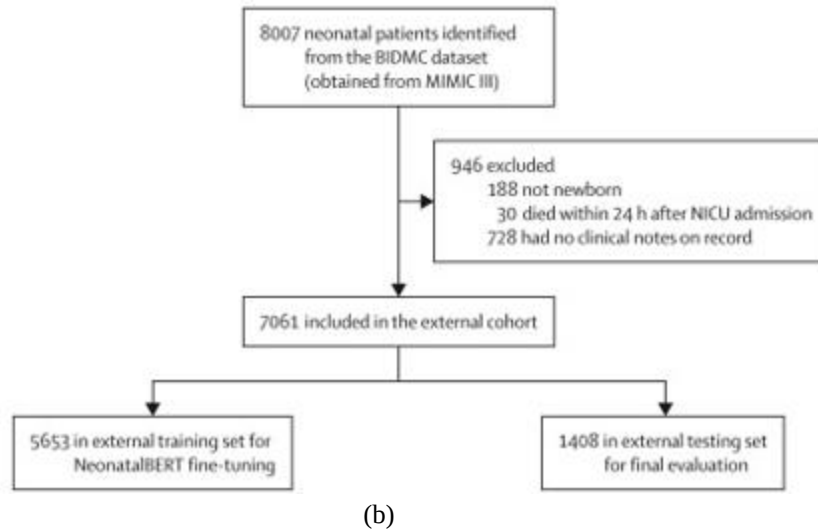
### *Data extraction and outcomes*

The primary cohort encompassed all infants delivered between 1 January 2014 and 31 December 2021 at hospitals and clinics affiliated with Stanford University, as documented in the Stanford Research Data Repository

(Figure 1a). Deliveries occurring from 1 January 2014 to 31 December 2020 constituted the primary training set, while those from 1 January to 31 December 2021 formed the primary testing set. This chronological separation was adopted to simulate real-world prospective use and to examine the influence of temporal changes in patient characteristics on model reliability. The analysis focused on 19 clinically relevant neonatal outcomes, grouped as follows: respiratory conditions (respiratory distress syndrome, bronchopulmonary dysplasia, pulmonary haemorrhage, pulmonary hypertension, atelectasis, aspiration syndrome), neurological complications (intraventricular haemorrhage, periventricular leukomalacia, neonatal seizures, and other CNS disorders), cardiac issues (patent ductus arteriosus and cardiovascular instability), infections (sepsis and candidiasis), haematologic abnormalities (anaemia and jaundice), and miscellaneous disorders (necrotising enterocolitis, retinopathy of prematurity, and death). We collected all clinical documentation for these patients, including admission notes, daily progress records, procedural descriptions, general entries, and specialist consultations. We extracted information on maternal and prenatal factors from the neonatal notes.



(a)



**Figure 1.** Assembly of the study cohorts; (a) Assembly of the primary study cohort; (b) Assembly of the external validation cohort.

The external cohort was assembled from newborn cases at Beth Israel Deaconess Medical Center using the Medical Information Mart for Intensive Care III database [11] maintained by the Massachusetts Institute of Technology Laboratory for Computational Physiology (**Figure 1b**). Ten key outcomes were examined: respiratory distress syndrome, bronchopulmonary dysplasia, pulmonary hemorrhage, intraventricular hemorrhage, patent ductus arteriosus, sepsis, jaundice, necrotizing enterocolitis, retinopathy of prematurity, and death. Because de-identification procedures altered date information, we randomly split the external cohort at the individual patient level into an external training set (80%) and an external testing set (20%). In the external cohort, ICD-9 codes [12] were used to define the outcome. Infants were excluded if they lacked clinical notes, had incomplete outcome records, or died within the initial 24 hours following birth.

#### *Model development and statistical analysis*

NeonatalBERT was constructed by performing domain-specific pre-training on top of the existing Bio-ClinicalBERT architecture [8]. The training corpus comprised neonatal clinical notes documented within the first three months of life, taken exclusively from the training subset of the primary cohort. This step was intended to align the model more closely with the distinctive linguistic features and clinical terminology used in neonatal care. The pre-training procedure replicated the original BERT self-supervised learning framework, incorporating both masked language

modeling and next-sentence prediction objectives. We identified the optimal model configuration through manual review and hyperparameter tuning on a dedicated portion of the training data, which served as an internal validation sample within the primary cohort.

Following pre-training, NeonatalBERT was fine-tuned on task-specific tasks to predict neonatal outcomes in each cohort. We added a logistic classification layer for each outcome and trained it on the corresponding cohort's training data. Fine-tuning involves refining a pre-trained model for a targeted application, here focused on generating risk estimates for particular neonatal morbidities. To maintain practical clinical value, fine-tuning relied solely on clinical notes recorded during the first 24 hours after birth from each cohort's training set. This approach resulted in two independent fine-tuned models: one optimized for the primary cohort and one for the external cohort. We evaluated these models on each cohort's test set (**Figure 2**).

Model performance was assessed through the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). While AUROC integrates recall with the true positive rate, AUPRC combines recall with positive predictive value. The latter metric is particularly suitable for evaluating rare and imbalanced conditions, which predominate among neonatal morbidities. To facilitate clearer interpretation, relative improvements in AUPRC were reported against a baseline no-skill classifier, given the measure's dependence on outcome prevalence. Given the pronounced class imbalance in most neonatal

outcomes (Table 1), we also used the F1 score [13] as an additional evaluation metric.

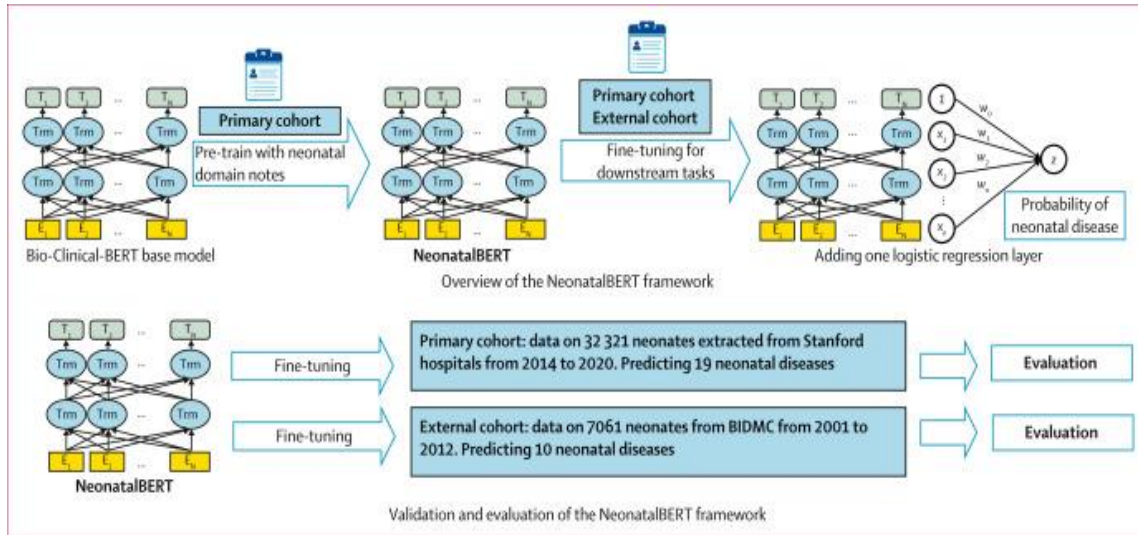


Figure 2. Development and validation process of NeonatalBERT

Table 1. Demographic and clinical features of the primary and external cohorts

| Parameter                                         | External cohort testing set (random 20% split; n = 1,408) | External cohort training set (random 80% split; n = 5,653) | External cohort (BIDMC) overall (n = 7,061) | Primary cohort testing set (2021 data; n = 4,910) | Primary cohort training set (2014–20 data; n = 27,411) | Primary cohort (Stanford) overall (n = 32,321) |
|---------------------------------------------------|-----------------------------------------------------------|------------------------------------------------------------|---------------------------------------------|---------------------------------------------------|--------------------------------------------------------|------------------------------------------------|
| Maternal age, years                               | ..                                                        | ..                                                         | ..                                          | 33.13 (5.37)                                      | 32.53 (5.62)                                           | 32.62 (5.59)                                   |
| Gestational age, weeks                            | ..                                                        | ..                                                         | ..                                          | 38.64 (2.13)                                      | 38.64 (2.30)                                           | 38.64 (2.28)                                   |
| Preterm birth (< 37 weeks)                        | 383 (27.2%)                                               | 1,567 (27.7%)                                              | 1,950 (27.6%)                               | 700 (14.3%)                                       | 4,075 (14.9%)                                          | 4,775 (14.8%)                                  |
| Extremely preterm (< 28 weeks)                    | ..                                                        | ..                                                         | ..                                          | 34 (0.7%)                                         | 254 (0.9%)                                             | 228 (0.7%)                                     |
| Very preterm (28 weeks to < 32 weeks)             | ..                                                        | ..                                                         | ..                                          | 55 (1.1%)                                         | 411 (1.5%)                                             | 466 (1.4%)                                     |
| Moderate-to-late preterm (32 weeks to < 37 weeks) | ..                                                        | ..                                                         | ..                                          | 611 (12.4%)                                       | 3,410 (12.4%)                                          | 4,021 (12.4%)                                  |
| Term birth (≥ 37 weeks)                           | 1,025 (72.8%)                                             | 4,086 (72.3%)                                              | 5,111 (72.4%)                               | 4,210 (85.7%)                                     | 23,336 (85.1%)                                         | 27,546 (85.2%)                                 |
| Birthweight, g                                    | ..                                                        | ..                                                         | ..                                          | 3,163.23 (661.04)                                 | 3,121.59 (544.63)                                      | 3,127.91 (564.05)                              |
| Sex                                               |                                                           |                                                            |                                             |                                                   |                                                        |                                                |
| Male                                              | 784 (55.7%)                                               | 3,039 (53.8%)                                              | 3,823 (54.1%)                               | 2,574 (52.4%)                                     | 14,355 (52.4%)                                         | 16,929 (52.4%)                                 |
| Female                                            | 624 (44.3%)                                               | 2,614 (46.2%)                                              | 3,238 (45.9%)                               | 2,336 (47.6%)                                     | 13,056 (47.6%)                                         | 15,392 (47.6%)                                 |
| Ethnicity                                         |                                                           |                                                            |                                             |                                                   |                                                        |                                                |
| White                                             | 833 (59.2%)                                               | 3,466 (61.3%)                                              | 4,299 (60.9%)                               | 1,445 (29.4%)                                     | 8,954 (32.7%)                                          | 10,399 (32.2%)                                 |
| Hispanic                                          | 67 (4.8%)                                                 | 285 (5.0%)                                                 | 352 (5.0%)                                  | 1,387 (28.2%)                                     | 8,337 (30.4%)                                          | 9,724 (30.1%)                                  |

|                                      |                |                  |                  |                  |                   |                   |
|--------------------------------------|----------------|------------------|------------------|------------------|-------------------|-------------------|
| <b>Asian</b>                         | 147<br>(10.4%) | 522 (9.2%)       | 669 (9.5%)       | 1,568<br>(31.9%) | 7,195<br>(26.2%)  | 8,763 (27.1%)     |
| <b>African</b>                       | 173<br>(12.3%) | 649<br>(11.5%)   | 822 (11.6%)      | 97 (2.0%)        | 596 (2.2%)        | 693 (2.1%)        |
| <b>Other</b>                         | 188<br>(13.4%) | 731<br>(12.9%)   | 919 (13.0%)      | 413 (8.4%)       | 2,329<br>(8.5%)   | 2,742 (8.5%)      |
| Respiratory morbidities              |                |                  |                  |                  |                   |                   |
| <b>Respiratory distress syndrome</b> | 233<br>(16.5%) | 1,031<br>(18.2%) | 1,264<br>(17.9%) | 315 (6.4%)       | 2,249<br>(8.2%)   | 2,564 (7.9%)      |
| <b>Bronchopulmonary dysplasia</b>    | 27 (1.9%)      | 133 (2.4%)       | 160 (2.3%)       | 36 (0.7%)        | 282 (1.0%)        | 318 (1.0%)        |
| <b>Pulmonary haemorrhage</b>         | 4 (0.3%)       | 16 (0.3%)        | 20 (0.3%)        | 4 (0.1%)         | 46 (0.2%)         | 50 (0.2%)         |
| <b>Pulmonary hypertension</b>        | ..             | ..               | ..               | 30 (0.6%)        | 152 (0.6%)        | 182 (0.6%)        |
| <b>Atelectasis</b>                   | ..             | ..               | ..               | 23 (0.5%)        | 305 (1.1%)        | 328 (1.0%)        |
| <b>Aspiration syndrome</b>           | ..             | ..               | ..               | 45 (0.9%)        | 356 (1.3%)        | 401 (1.2%)        |
| Neurological morbidities             |                |                  |                  |                  |                   |                   |
| <b>Intraventricular haemorrhage</b>  | 33 (2.3%)      | 151 (2.7%)       | 184 (2.6%)       | 36 (0.7%)        | 256 (0.9%)        | 292 (0.9%)        |
| <b>Periventricular leukomalacia</b>  | ..             | ..               | ..               | 4 (0.1%)         | 24 (0.1%)         | 28 (0.1%)         |
| <b>Seizures</b>                      | ..             | ..               | ..               | 19 (0.4%)        | 116 (0.4%)        | 135 (0.4%)        |
| <b>Other CNS disorders</b>           | ..             | ..               | ..               | 172 (3.5%)       | 844 (3.1%)        | 1,016 (3.1%)      |
| Cardiac morbidities                  |                |                  |                  |                  |                   |                   |
| <b>Patent ductus arteriosus</b>      | 64 (4.5%)      | 306 (5.4%)       | 370 (5.2%)       | 139 (2.8%)       | 948 (3.5%)        | 1,087 (3.4%)      |
| <b>Cardiovascular instability</b>    | ..             | ..               | ..               | 246 (5.0%)       | 1,775<br>(6.5%)   | 2,021 (6.3%)      |
| Infectious diseases                  |                |                  |                  |                  |                   |                   |
| <b>Neonatal sepsis</b>               | 40 (2.8%)      | 159 (2.8%)       | 199 (2.8%)       | 42 (0.9%)        | 560 (2.0%)        | 602 (1.9%)        |
| <b>Candidiasis</b>                   | ..             | ..               | ..               | 34 (0.7%)        | 186 (0.7%)        | 220 (0.7%)        |
| Blood disorders                      |                |                  |                  |                  |                   |                   |
| <b>Anaemia</b>                       | ..             | ..               | ..               | 89 (1.8%)        | 769 (2.8%)        | 858 (2.7%)        |
| <b>Jaundice</b>                      | 534<br>(37.9%) | 2,066<br>(36.5%) | 2,600<br>(36.8%) | 2,109<br>(43.0%) | 13,083<br>(47.7%) | 15,192<br>(47.0%) |
| Other conditions                     |                |                  |                  |                  |                   |                   |
| <b>Necrotising enterocolitis</b>     | 12 (0.9%)      | 46 (0.8%)        | 58 (0.8%)        | 11 (0.2%)        | 81 (0.3%)         | 92 (0.3%)         |
| <b>Retinopathy of prematurity</b>    | 29 (2.1%)      | 118 (2.1%)       | 147 (2.1%)       | 62 (1.3%)        | 576 (2.1%)        | 638 (2.0%)        |
| <b>Neonatal death</b>                | 4 (0.3%)       | 31 (0.5%)        | 35 (0.5%)        | 23 (0.5%)        | 125 (0.5%)        | 148 (0.5%)        |

Data are shown as mean (SD) or n (%). BIDMC = Beth Israel Deaconess Medical Center.

For comparison, we generated baseline models for each cohort by applying logistic regression to structured tabular EHR variables collected on the first day of life. We also evaluated NeonatalBERT against earlier BERT-derived language models, namely BioBERT and Bio-ClinicalBERT. All comparison models were developed and tested using the same patient cohorts and clinical notes used during NeonatalBERT fine-tuning. We also benchmarked the model against conversational large language models (such as Meta Llama-3.1-8B) to demonstrate the proposed model's superior applicability for clinical risk stratification and decision-making.

#### Role of the funding source

The funding organizations played no part in the study design, data acquisition, statistical analysis, interpretation of results, or the drafting of the report.

#### Results and Discussion

A total of 32 321 newborns formed the primary cohort. Of these, 27 411 were allocated to the primary training set (mean gestational age 38.64 weeks [SD 2.30]; 13 056 [47.6%] female and 14 355 [52.4%] male newborns) and

4910 to the primary testing set (mean gestational age 38.64 [2.13] weeks; 2336 [47.6%] female and 2574 [52.4%] male newborns); (**Table 1**). The study sample included both full-term and preterm infants, which permitted the model to identify patterns across diverse clinical presentations. Prevalence of neonatal morbidities differed substantially, with notable variation according to gestational age at birth. The external cohort consisted of 7061 newborns, including 5653 in the external training set (1567 [27.7%] premature and 4086 [72.3%] term births; 2614 [46.2%] female and 3039 [53.8%] male newborns) and 1408 in the external testing set (383 [27.2%] premature and 1025 [72.8%] term births; 624 [44.3%] female and 784 [55.7%] male newborns). Newborns in the external cohort experienced higher rates of morbidities than those in the primary cohort, reflecting their admission to neonatal intensive care and the larger proportion of premature deliveries.

NeonatalBERT achieved a mean AUROC of 0.910 (95% CI: 0.900–0.921) across these outcomes, surpassing 0.884 (0.870–0.901) for Bio-ClinicalBERT, 0.877 (0.867–0.890) for BioBERT, and 0.775 (0.749–0.802) for the standard regression model relying on tabular features. The model also recorded a mean AUPRC of 0.291 (95% CI: 0.268–0.314), outperforming Bio-ClinicalBERT at 0.238 (95% CI: 0.217–0.259), BioBERT at 0.217 (95% CI: 0.197–0.236), and the conventional tabular regression approach at 0.194 (0.177–0.211). Relative to a random classifier, AUPRC improved by up to 129.6-fold for mortality prediction, 13.5-fold for periventricular leukomalacia, 14.1-fold for patent ductus arteriosus, 25.7-fold for intraventricular hemorrhage, and 88.2-fold for pulmonary hemorrhage. Performance gains were more striking with AUPRC than with AUROC, underscoring the model's effectiveness in addressing rare, imbalanced outcomes.

**Table 2.** Performance metrics of different models evaluated on the testing sets of the primary and external cohorts

| Outcome                              | AUPRC (× baseline)*<br>– Bio-BERT | AUPRC (× baseline)*<br>– Bio-ClinicalBERT | AUPRC (× baseline)*<br>– Logistic regression† | AUPRC (× baseline)*<br>– NeonatalBERT | AUPRC (absolute)<br>– Bio-BERT | AUPRC (absolute)<br>– Bio-ClinicalBERT | AUPRC (absolute)<br>– Logistic regression† | AUPRC (absolute)<br>– NeonatalBERT |
|--------------------------------------|-----------------------------------|-------------------------------------------|-----------------------------------------------|---------------------------------------|--------------------------------|----------------------------------------|--------------------------------------------|------------------------------------|
| Respiratory outcomes                 |                                   |                                           |                                               |                                       |                                |                                        |                                            |                                    |
| <b>Respiratory distress syndrome</b> | 5.8                               | 9.2                                       | 6.8                                           | 11.0                                  | 0.371<br>(0.322–0.423)         | 0.587 (0.532–0.643)                    | 0.436<br>(0.384–0.493)                     | 0.707 (0.650–0.763)                |
| <b>Bronchopulmonary dysplasia</b>    | 29.6                              | 38.8                                      | 46.8                                          | 47.6                                  | 0.296<br>(0.238–0.442)         | 0.388 (0.283–0.492)                    | 0.468<br>(0.318–0.614)                     | 0.476 (0.338–0.618)                |
| <b>Pulmonary haemorrhage</b>         | 25.0                              | 26.0                                      | 78.0                                          | 88.2                                  | 0.025<br>(0.009–0.072)         | 0.026 (0.011–0.058)                    | 0.078<br>(0.005–0.283)                     | 0.088 (0.029–0.344)                |
| <b>Pulmonary hypertension</b>        | 18.0                              | 18.0                                      | 2.5                                           | 25.2                                  | 0.108<br>(0.067–0.190)         | 0.108 (0.067–0.190)                    | 0.015<br>(0.005–0.072)                     | 0.151 (0.086–0.254)                |
| <b>Atelectasis</b>                   | 17.8                              | 12.4                                      | 3.4                                           | 12.0                                  | 0.089<br>(0.022–0.247)         | 0.062 (0.021–0.200)                    | 0.017<br>(0.008–0.042)                     | 0.06 (0.028–0.141)                 |
| <b>Aspiration syndrome</b>           | 7.0                               | 6.3                                       | 2.3                                           | 10.8                                  | 0.063<br>(0.025–0.101)         | 0.057 (0.032–0.089)                    | 0.021<br>(0.015–0.034)                     | 0.097 (0.061–0.183)                |
| Neurological outcomes                |                                   |                                           |                                               |                                       |                                |                                        |                                            |                                    |
| <b>Intraventricular haemorrhage</b>  | 18.7                              | 18.1                                      | 18.4                                          | 25.8                                  | 0.131<br>(0.085–0.178)         | 0.127 (0.089–0.195)                    | 0.129<br>(0.086–0.199)                     | 0.180 (0.129–0.286)                |

|                                     |      |      |      |       |                        |                      |                        |                     |
|-------------------------------------|------|------|------|-------|------------------------|----------------------|------------------------|---------------------|
| <b>Periventricular leukomalacia</b> | 10.0 | 9.0  | 4.0  | 13.5  | 0.010<br>(0.002–0.030) | 0.009 (0.001–0.029)  | 0.004<br>(0.000–0.018) | 0.014 (0.003–0.036) |
| <b>Seizures</b>                     | 34.8 | 40.3 | 8.3  | 21.5  | 0.139<br>(0.052–0.287) | 0.161 (0.066–0.318)  | 0.033<br>(0.010–0.085) | 0.083 (0.042–0.177) |
| <b>Other CNS disorders</b>          | 6.9  | 5.4  | 0.1  | 6.0   | 0.242<br>(0.047–0.461) | 0.188 (0.051–0.361)  | 0.003<br>(0.001–0.007) | 0.208 (0.051–0.488) |
| Cardiovascular outcomes             |      |      |      |       |                        |                      |                        |                     |
| <b>Patent ductus arteriosus</b>     | 14.3 | 10.8 | 5.1  | 14.1  | 0.400<br>(0.326–0.466) | 0.301 (0.239–0.364)  | 0.144<br>(0.090–0.210) | 0.395 (0.318–0.500) |
| <b>Cardiovascular instability</b>   | 4.7  | 5.1  | 5.5  | 8.0   | 0.234<br>(0.194–0.263) | 0.256 (0.214–0.286)  | 0.274<br>(0.239–0.322) | 0.399 (0.339–0.466) |
| Infectious outcomes                 |      |      |      |       |                        |                      |                        |                     |
| <b>Neonatal sepsis</b>              | 14.3 | 26.9 | 8.0  | 17.7  | 0.129<br>(0.108–0.205) | 0.242 (0.156–0.358)  | 0.072<br>(0.033–0.140) | 0.159 (0.093–0.291) |
| <b>Candidiasis</b>                  | 1.3  | 1.6  | 1.7  | 2.0   | 0.009<br>(0.007–0.014) | 0.011 (0.008–0.015)  | 0.012<br>(0.007–0.035) | 0.014 (0.009–0.023) |
| Haematological outcomes             |      |      |      |       |                        |                      |                        |                     |
| <b>Anaemia</b>                      | 31.4 | 30.6 | 30.7 | 34.8  | 0.566<br>(0.510–0.677) | 0.551 (0.461–0.633)  | 0.552<br>(0.442–0.655) | 0.627 (0.549–0.727) |
| <b>Jaundice</b>                     | 1.3  | 1.3  | 1.2  | 1.3   | 0.545<br>(0.529–0.560) | 0.55 (0.538–0.574)   | 0.531<br>(0.507–0.555) | 0.576 (0.560–0.602) |
| Other neonatal outcomes             |      |      |      |       |                        |                      |                        |                     |
| <b>Necrotising enterocolitis</b>    | 46.5 | 87.5 | 32.0 | 46.4  | 0.093<br>(0.055–0.239) | 0.175 (0.061–0.435)  | 0.064<br>(0.030–0.155) | 0.093 (0.027–0.204) |
| <b>Retinopathy of prematurity</b>   | 17.5 | 17.5 | 22.4 | 27.7  | 0.35<br>(0.218–0.443)  | 0.349 (0.233, 0.443) | 0.447<br>(0.373–0.553) | 0.555 (0.437–0.679) |
| <b>Neonatal death</b>               | 63.0 | 74.4 | 78.4 | 129.6 | 0.315<br>(0.198–0.441) | 0.372 (0.226–0.526)  | 0.392<br>(0.249–0.548) | 0.648 (0.498–0.780) |
| Mean across all outcomes            | 19.4 | 23.1 | 18.7 | 28.6  | 0.217<br>(0.197–0.236) | 0.238 (0.217–0.259)  | 0.194<br>(0.177–0.211) | 0.291 (0.268–0.314) |

Data are presented as estimates with 95% CIs. AUPRC = area under the precision-recall curve.

\* Represents relative increase or decrease in AUPRC compared with a random classifier (baseline).

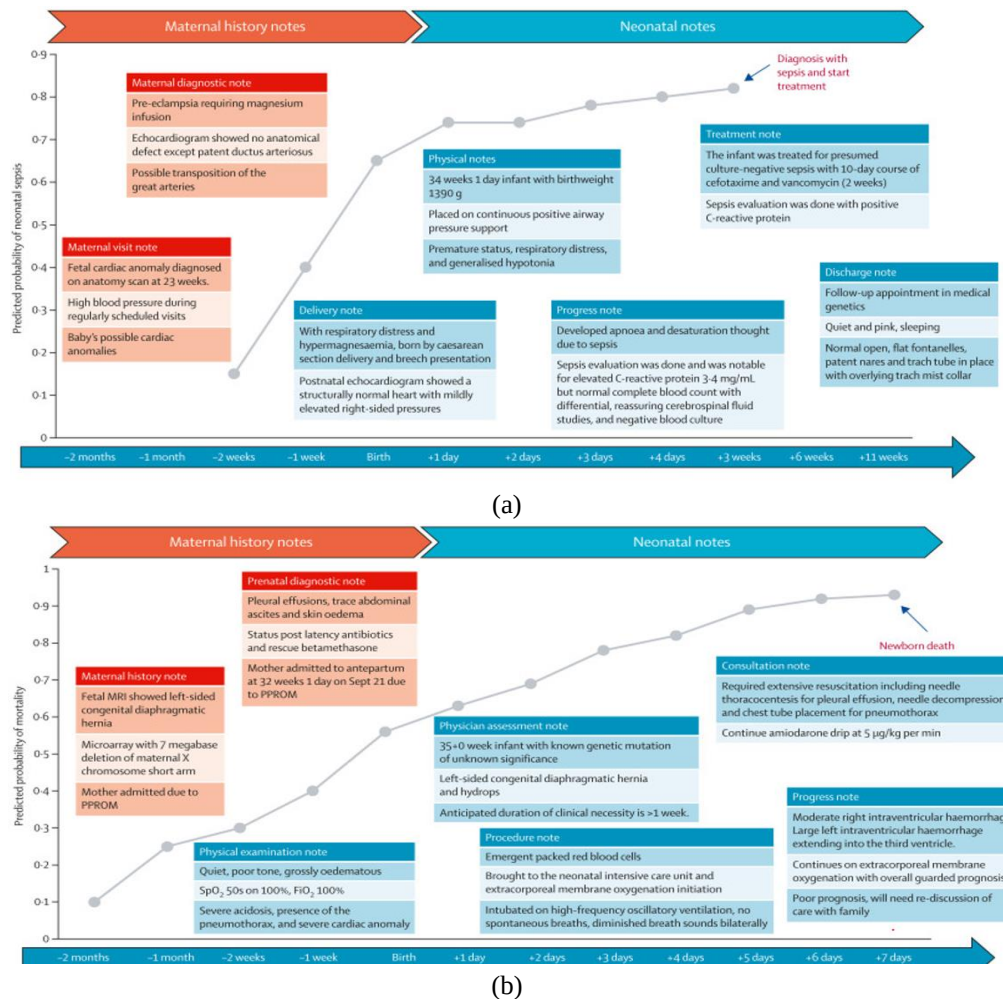
† Logistic regression using tabular variables only.

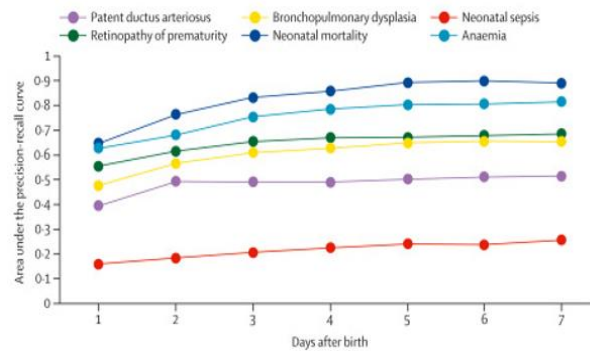
In external validation, NeonatalBERT demonstrated. By contrast, a logistic regression model based exclusively on tabular data achieved a mean AUROC of 0.816 (0.772–0.860) and a mean AUPRC of 0.224 (0.202–0.247).

BioBERT and Bio-ClinicalBERT produced AUROCs of 0.925 (0.917–0.933) and 0.925 (0.917–0.932), respectively, with corresponding AUPRCs of 0.333 (0.303–0.363) and 0.327 (0.301–0.353). In addition, NeonatalBERT achieved a mean F1 score of 0.416, exceeding logistic regression (0.267), BioBERT (0.379), and Bio-ClinicalBERT (0.386). These outcomes highlight NeonatalBERT's ability to generalize successfully to varied hospital settings, patient groups, and EHR platforms.

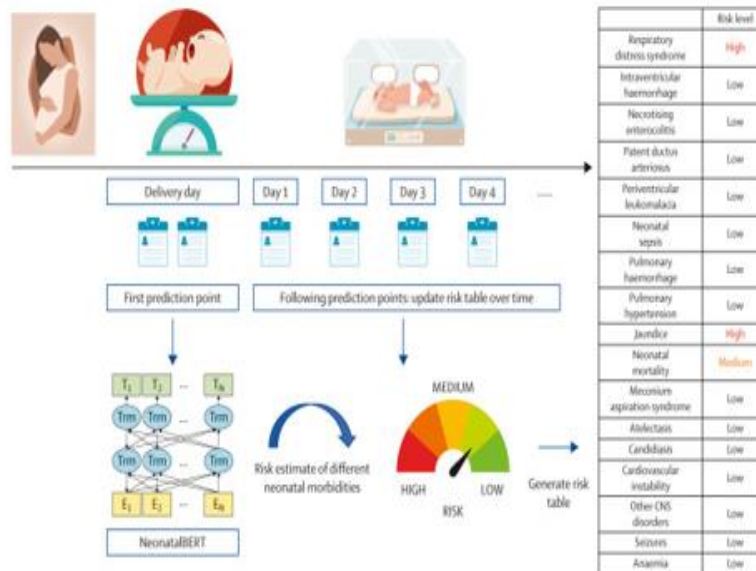
Infants diagnosed with multiple neonatal morbidities—including respiratory distress syndrome, patent ductus arteriosus, intraventricular hemorrhage, necrotizing enterocolitis, and sepsis—exhibited tight clustering. In contrast, cases of jaundice did not form a distinct group, which accounts for the observed differences in prediction reliability. **Figure 3** depicts illustrative clinical scenarios in which the model can assist decision-making for newborns immediately after birth. For an infant born at

35 weeks' gestation, the figure displays sepsis risk predictions derived from keywords documented in the patient's notes before and after delivery. The model effectively synthesized diverse inputs, including maternal background, nuanced clinical indicators, and laboratory results, and successfully detected an increased likelihood of sepsis well before confirmatory diagnosis in the third week of life. The same figure also presents mortality risk predictions for a separate newborn. Although recorded diagnoses of this nature are typically considered reliable, recent studies [14] reveal actual survival rates ranging from 60–80% following such predictions, indicating considerable prognostic heterogeneity. This temporal analysis confirmed that accuracy increased as additional notes and clinical data accumulated over time, as evidenced by average AUPRC values rising from 0.291 on day 1 to 0.380 on day 3 and 0.417 on day 7.





(c)



(d)

**Figure 3.** NeonatalBERT facilitating clinical decision support in neonatal care

This investigation introduced NeonatalBERT and evaluated its effectiveness across different hospital environments and patient populations. Unlike conventional statistical models that rely solely on structured tabular data to forecast neonatal morbidity, NeonatalBERT uses a transformer neural network architecture to extract meaningful patterns from detailed, free-text clinical documentation. This methodology strengthens contextual interpretation and supports broader transferability between institutions [15]. Overall, NeonatalBERT opens the door to multiple innovative applications that aim to advance neonatal healthcare delivery and improve hospital operational efficiency. Several key elements account for NeonatalBERT's superior results compared with baseline approaches. Thanks to its pre-training on extensive corpora, the model acquires rich contextual representations, as

language models primarily build their foundational knowledge during pre-training rather than during subsequent fine-tuning [16]. Furthermore, clinical narratives are routinely documented and widely accessible in standard practice, unlike many structured data elements, such as biomarker measurements, which are frequently unavailable for certain patients. These free-text records offer a wealth of clinical details, enabling NeonatalBERT to draw upon a substantially larger volume of observations throughout pre-training and task-specific adaptation for neonatal cases than is possible with tabular datasets alone. This comprehensive exploitation of textual information is a major advantage, driving the model's higher accuracy than standard machine learning techniques and regression models limited to structured data. Critical aspects of maternal and infant health—ranging from physical assessment

details and imaging interpretations to inter-professional communications and results from specialized genetic testing—are predominantly documented in the unstructured sections of electronic health records [17]. NeonatalBERT may also uncover signals related to uncommon disorders that are difficult to encode in conventional tabular formats. Relative to Bio-ClinicalBERT, NeonatalBERT received additional domain-specific training focused on neonatal conditions and terminology. Starting from Bio-ClinicalBERT weights provides an advantageous foundation of relevant medical knowledge, reducing the resources required compared with training general BERT models from scratch. In this manner, NeonatalBERT combines the core linguistic strengths of the BERT framework with specialized refinement for neonatal medicine, a targeted expertise not present in comparator models such as BioBERT or Bio-ClinicalBERT, which explains its improved predictive capability for neonatal outcomes. Although Llama-family large language models perform well in dialogue and text generation [18], they showed inferior performance in neonatal condition prediction in our evaluations, consistent with earlier reports [19, 20]. Moreover, their deployment in clinical settings is hindered by considerable computational costs, even at inference, limiting feasibility in resource-constrained environments [21]. BERT-derived models, on the other hand, provide a well-balanced trade-off between accuracy, efficiency, and practicality [19, 22], rendering them a more suitable choice for risk assessment of neonatal disorders.

NeonatalBERT holds promise for detecting prodromal risk indicators of neonatal complications by analyzing clinical documentation available before definitive diagnosis, thereby enabling earlier stratification of risk levels. The system can provide neonatologists with forward-looking analytics to help identify emerging problems and initiate timely measures that can positively influence patient trajectories. For example, the model successfully predicted sepsis in term infants without overt signs at delivery, thereby facilitating an accelerated diagnostic workup and therapeutic intervention [23]. Our findings indicated that NeonatalBERT identified elevated sepsis likelihood days in advance, whereas formal confirmation and management occurred during the third week of life. Accordingly, the model can consolidate disparate information sources to support earlier clinical suspicion than might otherwise occur. In another example involving neonatal death, the model

generated a high mortality risk score in the initial days, with the outcome materializing on day 7. NeonatalBERT could therefore help medical teams anticipate potential complications and engage families earlier through informed discussions about prognosis.

In addition, the model's predictive power may help clinicians recognize infants at risk for necrotizing enterocolitis in a timely manner, a condition notoriously difficult to predict [24]. Warnings would permit proactive preventive actions, including modifications to nutritional management and the introduction of prophylactic probiotics [25]. Predictions concerning pulmonary hypertension could similarly trigger prompt respiratory support to ensure adequate oxygenation and avoid tissue hypoxia. By forecasting intraventricular hemorrhage risk, NeonatalBERT enables early deployment of measures to stabilize brain perfusion among preterm neonates, potentially decreasing hemorrhage incidence and reducing associated neurodevelopmental consequences [26]. For respiratory distress syndrome, the model can highlight high-risk newborns, allowing clinical staff to prepare necessary surfactant replacement and ventilatory assistance without delay. The capacity to predict periventricular leukomalacia from birth-day records further facilitates recommendations for immediate imaging studies, such as ultrasound or magnetic resonance scans, in susceptible infants before clinical signs develop. Although existing tools such as BiliTool provide value for hyperbilirubinemia evaluation, they remain narrow in focus and dependent on available structured inputs [27]. NeonatalBERT, by thoroughly processing unstructured clinical narratives, offers versatile prediction across numerous neonatal disorders using only free-text data. Its relevance also extends to conditions that typically appear later in the neonatal course, notably bronchopulmonary dysplasia and retinopathy of prematurity. Early detection of bronchopulmonary dysplasia risk—often diagnosed beyond 28 days—supports individualized management, including refined mechanical ventilation approaches or corticosteroid use. For retinopathy of prematurity, which generally emerges between 4 and 6 weeks after birth, NeonatalBERT predictions could help intensify retinal screening and prompt timely therapeutic measures to avert visual impairment.

NeonatalBERT delivers objective assessments and an aggregated risk score by evaluating the full set of available clinical documentation. Neonatal conditions vary widely in onset and clinical detection. Certain

disorders, including meconium aspiration syndrome and respiratory distress syndrome, are generally recognized at or immediately following delivery. In contrast, others—such as sepsis or necrotizing enterocolitis—may arise several days or weeks later. Consequently, the model's risk scores function as descriptive tools for conditions with immediate presentation and as predictive indicators for those that evolve gradually. Although neonatologists typically develop an initial clinical impression and gauge risk based on their judgment, such assessments are shaped by personal experience and proficiency, which can differ substantially [28]. In this setting, the model's risk estimates provide an independent source of confirmation, especially beneficial in the fast-paced environment of neonatal intensive care. Prior research [29] has emphasized the value of data-informed decision-support systems in encouraging neonatologists to re-evaluate, reconsider, and verify their impressions of patients, thereby reducing the likelihood of overlooking morbidities or important risk factors. NeonatalBERT can be seamlessly incorporated into current health information systems and serves as a versatile platform for reimagining the use of clinical notes in neonatal medicine, bypassing the need for labor-intensive data extraction and standardization [22].

We suggest implementing a daily workflow in which the model generates updated risk profiles for each neonate at the end of each clinical shift. These profiles can serve as informative reference points during morning rounds, helping clinicians adjust and optimize daily management strategies. Integrating these assessments via a dedicated dashboard within the electronic health record system would give clinicians efficient access to practical insights that augment existing clinical information and promote more proactive, evidence-based choices in routine practice [30]. Deploying NeonatalBERT in neonatal units requires thorough attention to ethical considerations, patient data confidentiality, and possible biases embedded in clinical records [31]. Documentation practices may vary by individual provider habits, institutional conventions, and patient demographics (such as gestational age, sex, and racial or ethnic background) [32], potentially affecting model outputs. To address these issues, continued training on representative, varied datasets is essential to minimize bias and strengthen applicability across groups. Regular audits and equity assessments will be vital to ensure fair performance across subpopulations. To prevent reinforcing existing patterns, NeonatalBERT should

undergo periodic retraining using current electronic health record data to accommodate changes in documentation practices. Subsequent investigations should prioritize prospective trials to test NeonatalBERT under real-world clinical conditions and assess its impact on decision-making processes and patient outcomes. The model's scalability will also be critical for widespread implementation in varied healthcare contexts. In well-resourced facilities, supplementary pre-training on institution-specific data can accommodate local variations and refine domain adaptation [33]. In resource-limited settings, NeonatalBERT can function effectively in a fine-tuning-only configuration, processing clinical notes with minimal additional computational requirements.

This study has several limitations. First, the data originated from an electronic repository of routinely gathered clinical information. As a result, the dataset may contain missing or inaccurately entered variables, and certain outcomes were unavailable because they were not documented. In addition, more than 6000 newborns were omitted from the analysis because of missing notes or incomplete records. Such exclusions may have introduced selection bias if the omitted cases differed systematically from the retained cases. Moreover, reliance on ICD codes extracted from electronic health records to determine diagnosis dates may not precisely reflect the true onset of a condition, potentially leading to slight information leakage in some cases. Future efforts will also explore incorporating additional data sources, such as continuous physiological monitoring, medication logs, and structured elements, to refine outcome definitions and improve label accuracy for multifaceted neonatal disorders. NeonatalBERT was initially pre-trained using data from one healthcare organization (Stanford Health Care), and variations in clinical recording styles and procedures across systems may be substantial. Although the model maintained good performance on an independent external cohort from the Beth Israel Deaconess Medical Center after fine-tuning, this does not guarantee equivalent results when applied to notes from other healthcare organizations. Despite solid performance across two separate cohorts, the modest overall sample size and limited representation of preterm infants limit broader applicability. Additional validation across larger, more heterogeneous neonatal groups is needed to confirm the model's robustness.

## Conclusion

In summary, the present work demonstrates that a neonatal-focused large language model can harness routinely recorded clinical notes to generate risk estimates for a wide range of neonatal morbidities across two distinct cohorts. Although predictive strength varied by outcome—serving a descriptive role for conditions evident at birth and an anticipatory role for those appearing later—NeonatalBERT’s stable performance across institutions supports its role as a supportive instrument in clinical practice, especially when structured data are sparse or absent. These results support adopting automated risk assessment to enable earlier therapeutic action, optimize clinical processes, and strengthen interactions with families. At the same time, they underscore the necessity for careful, responsible deployment, ongoing validation, and appropriate oversight of artificial intelligence tools in neonatal medicine.

**Acknowledgments:** None

**Conflict of Interest:** None

**Financial Support:** None

**Ethics Statement:** None

## References

1. Stoll BJ, Hansen NI, Bell EF, Walsh MC, Carlo WA, Shankaran S, et al. Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993-2012. *JAMA*. 2015;314(10):1039-51.
2. De Francesco D, Reiss JD, Roger J, Foo PY, Liu S, Kim H, et al. Data-driven longitudinal characterization of neonatal health and morbidity. *Sci Transl Med*. 2023;15(710):eadc9854.
3. Mangold C, Zoretic S, Thallapureddy K, Moreira A, Guzman L, Patel RM, et al. Machine learning models for predicting neonatal mortality: a systematic review. *Neonatology*. 2021;118(4):394-405.
4. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. 2024;56(2):1-40.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-40.
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2019 Jun 2-7; Minneapolis, MN, USA. Minneapolis: Association for Computational Linguistics; 2019. p. 4171-86.
7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-40.
8. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*; 2019 Jun 7; Minneapolis, MN, USA. Minneapolis: Association for Computational Linguistics; 2019. p. 72-8.
9. Mao C, Xu J, Rasmussen L, Li Y, Pathak J, Wang F. AD-BERT: using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer’s disease. *J Biomed Inform*. 2023;144:104442.
10. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
11. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.
12. De Francesco D, Blumenfeld YJ, Marić I, Aghaeepour N, Rios M, Shaw GM, et al. A data-driven health index for neonatal morbidities. *iScience*. 2022;25(6):104143.
13. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
14. Politis MD, Bermejo-Sánchez E, Canfield MA, Contreras M, Lawrence G, Lee WR, et al. Prevalence and mortality in children with congenital diaphragmatic hernia: a multicountry study. *Ann Epidemiol*. 2021;56:61-9.

15. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med.* 2023;6:135.
16. Gekhman Z, Yona G, Aharoni R, Eyal M, Goldberg Y, Szpektor I, et al. Does fine-tuning LLMs on new knowledge encourage hallucinations? In: Findings of the Association for Computational Linguistics: EMNLP 2024; 2024 Nov 12-16; Miami, FL, USA. Miami: Association for Computational Linguistics; 2024. p. 1012-29.
17. Dong H, Suárez-Paniagua V, Zhang H, Wang Q, Wu H, Yang C. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Med Inform Decis Mak.* 2023;23:86.
18. Chen S, Guevara M, Moningi S, Hoerner F, Bisk Y, Turan B, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health.* 2024;6(6):e379-81.
19. Hager P, Jungmann F, Holland R, Bhagat K, Maurer A, Ezhil A, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med.* 2024;30:2613-22.
20. Brown KE, Yan C, Li Z, Bafna R, Poon B, Han B, et al. Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *J Am Med Inform Assoc.* 2025;32(4):811-22.
21. Verlingue L, Boyer C, Olgiati L, Bonastre J, Blanc T, Borthwick K, et al. Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. *Lancet Reg Health Eur.* 2024;46:101064.
22. Jiang LY, Liu XC, Nejatian NP, Nasir M, Chen Z, Kim S, et al. Health system-scale language models are all-purpose prediction engines. *Nature.* 2023;619(7969):357-62.
23. Chen KT, Ringer S, Cohen AP, Goldman DA, Aghai ZH, Graham DA, et al. The role of intrapartum fever in identifying asymptomatic term neonates with early-onset neonatal sepsis. *J Perinatol.* 2002;22(8):653-7.
24. Thompson AM, Bizzarro MJ. Necrotizing enterocolitis in newborns: pathogenesis, prevention and management. *Drugs.* 2008;68(9):1227-38.
25. Razak A, Patel RM, Gautham KS. Use of probiotics to prevent necrotizing enterocolitis: evidence to clinical practice. *JAMA Pediatr.* 2021;175(8):773-4.
26. Ballabh P, de Vries LS. White matter injury in infants with intraventricular haemorrhage: mechanisms and therapies. *Nat Rev Neurol.* 2021;17(4):199-214.
27. Petersen JD, Lozovatsky M, Markovic D, Kim J, Bertenthal D, Harris C, et al. Clinical decision support for hyperbilirubinemia risk assessment in the electronic health record. *Acad Pediatr.* 2020;20(6):857-62.
28. Cavolo A, Dierckx de Casterlé B, Naulaers G, Demarsin K, Dumont A, De Jonckheere J, et al. Neonatologists' resuscitation decisions at birth for extremely premature infants. A Belgian qualitative study. *Front Pediatr.* 2022;10:852073.
29. Vijlbrief D, Dudink J, van Solinge W, Benders M, van Bel F, de Vries LS. From computer to bedside, involving neonatologists in artificial intelligence models for neonatal medicine. *Pediatr Res.* 2023;93(2):437-9.
30. Alves M, Seringa J, Silvestre T, Almeida A, Dias I, Pereira J, et al. Use of artificial intelligence tools in supporting decision-making in hospital management. *BMC Health Serv Res.* 2024;24(1):1282.
31. Ning Y, Teixayavong S, Shang Y, Xu Z, Yang J, Liu D, et al. Generative artificial intelligence and ethical considerations in health care: a scoping review and ethics checklist. *Lancet Digit Health.* 2024;6(12):e848-56.
32. Lee CR, Aysola J, Chen X, Kogan A, He W, Ng R, et al. Race and ethnicity and clinician linguistic expressions of doubt in hospital admission notes. *JAMA Netw Open.* 2024;7(11):e2438550.
33. Carrell DS, Schoen RE, Leffler DA, Meyers C, Harrison M, Hsu W, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc.* 2017;24(5):986-91.