medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Machine learning for the prediction of spontaneous preterm birth using early second and third trimester maternal blood gene expression: A Cautionary Tale

Kylie K Hornaday*[1¶], Ty Werbicki[1¶], Suzanne C Tough[2,3], Stephen L Wood[4], David W Anderson[5#], Donna M Slater[1,4]

[1]Department of Physiology and Pharmacology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[2]Department of Community of Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[3]Department of Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[4]Department of Obstetrics and Gynaecology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

[5]Department of Biochemistry and Molecular Biology, Cumming School of Medicine, University of Calgary, Alberta, Canada

[#]Current address: Department of Science, Langara College, Vancouver, British Columbia, Canada

*Corresponding author

E-mail: kylie.hornaday@ucalgary.ca (KKH)

¶These authors contributed equally to this work

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Abstract

Preterm birth (PTB) remains a significant global health challenge and a leading cause of neonatal mortality and morbidity. Despite advancements in neonatal care, the prediction of PTB remains elusive, in part due to complex etiologies and heterogeneous patient populations. This study aimed to validate and extend information on gene expression biomarkers previously described for predicting spontaneous PTB (sPTB) using maternal whole blood from the All Our Families pregnancy cohort study based in Calgary, Canada. The results of this study are two-fold: first, using additional replicates of maternal blood samples from the All Our Families cohort, we were unable to repeat the findings of a 2016 study which identified top maternal gene expression predictors for sPTB. Second, we conducted a secondary analysis of the original gene expression dataset from the 2016 study, including external validation using a pregnancy cohort based in Detroit, USA. While initial results of our machine learning model suggested promising performance (area under the receiver operating curve, AUC 0.90 in the training set), performance was significantly degraded on the test set (AUC 0.54), and further degraded in external validation (AUC 0.51), suggesting poor generalizability, likely due to overfitting exacerbated by a low feature-to-noise ratio. Prediction was not improved when using machine learning approaches over traditional statistical learning. These findings underscore the challenges in translating biomarker discovery into clinically useful predictive models for sPTB. This study highlights the critical need for rigorous methodological safeguards and external validation in biomarker research. It also emphasizes the impact of data noise and overfitting on model performance, particularly in high-dimensional omics datasets. Future research should prioritize

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

robust validation strategies and explore mechanistic insights to improve our understanding and prediction of PTB.

Introduction

Preterm birth, defined as delivery of a live infant prior to 37 weeks of gestation, occurs in 13.4 million births worldwide, and is a significant contributor to mortality and morbidity in neonates and children under five [1, 2]. While approximately one third of preterm births occur following known maternal or fetal indications, the remaining two thirds occur following spontaneous onset of labour and/or premature rupture of the fetal membranes (sPTB) without known indication, making prediction and subsequent clinical management of those at risk challenging [2]. Child mortality related to PTB complications has declined since 2000, in part due to advancements in treatments for neonatal complications of prematurity such as respiratory distress syndrome (RDS). However, an estimated 900,000 PTB-associated deaths of children under five still occurred in 2019 worldwide [3].

As one of the great obstetrical syndromes, the ability to predict sPTB may be key to improving outcomes. Indeed, considerable efforts have aimed to identify predictive biomarkers of sPTB, however none so far have emerged to have clinical utility, possibly due to heterogeneity within both patient populations and preterm birth phenotypes, as well as risk of bias within study design [4-7]. Methodological safeguarding and appropriate validation of models is important to determine the feasibility, repeatability, robustness, and generalizability of prediction [8-11]. Best practices for prediction modelling are well defined in the literature [12,

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

13], and primary research articles reporting external validation of prediction models have been increasingly published over the last five years [14-17]. However, studies externally validating prediction models in the reproductive field are limited [10, 18, 19]. Additionally, strategies to qualify and understand heterogeneity and regular updating and assessment of predictive models is important to ensure best practice and clinical utility for prediction [20]. Thus, we sought herein to determine the predictive relationship between gene expression biomarkers and spontaneous preterm birth, and to repeat and validate previous findings on prediction of sPTB.

Gene expression biomarkers have been identified in maternal whole blood for the prediction of sPTB, which presents a promising avenue for minimally invasive prediction as peripheral blood can reflect global and uterine physiological and immunological changes during pregnancy [21]. One example includes eight genes, *LOC100128908, MIR3691, LOC101927441, CST13P, ACAP2, ZNF324, SH3PXD2B, TBX21* that were identified as significantly predictive of sPTB (65% sensitivity and 88% specificity after adjusting for history of abortion and anaemia) in a stepwise logistic regression model [22]. These gene expression biomarkers were identified using an Affymetrix chip microarray analysis of maternal whole blood from the All Our Families pregnancy cohort based in Calgary, Canada [22]. The All Our Families pregnancy cohort presents a rare opportunity for testing experimental repeatability, as maternal blood samples were collected and stored in four separate PAXgene RNA tubes, two were used for the original study (22), and one for validating RNA quality and integrity [23], leaving a remaining fourth sample for experimental validation. The study herein sought to use the additional PAXgene tube to repeat and validate this predictive model to test feasibility for clinical use.

Further, though the original study, which used a logistic regression based model, presented promising predictive performance, we hypothesized that machine learning approaches

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

could improve predictive performance. Machine learning and other complex data analysis methods are particularly well suited for mining high dimensional datasets, such as transcriptomic datasets, as they do not generally require the data to adhere to any *a priori* assumptions [24]. Machine learning allows for the identification of non-obvious, interactive, complex, and/or non-linear patterns which can go undetected when using traditional statistical linear models [24, 25]. These patterns can be leveraged both toward outcome prediction (something highly valuable for complex medical conditions such as preterm birth) and characterizing underlying disease mechanisms [26, 27]. This is particularly enticing, as the underlying causes of sPTB remain poorly understood. The authors have identified an external pregnancy cohort based in Detroit, USA for external validation of both regression- and machine learning-based prediction of sPTB to determine the generalizability of prediction. Prediction algorithms that match too closely to the training data, in other words, suffer from overfitting, are not generalizable to other populations, which is one of the major limitations of machine learning and other methods for prediction. This problem is exacerbated by small or non-representative training sets, where patterns identified may not be meaningfully associated with the outcome, or "noise" and thus the prediction does not translate effectively beyond the original training observations. This stresses the importance of external validation in order to identify robust, generalizable models to meaningfully push forward the prediction of preterm birth.

The overarching aim is to explore the repeatability, generalization, and robustness of a prediction model for sPTB using maternal blood gene expression biomarkers. Specific aims are as follows:

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

1. To test the predictive utility of *LOC100128908, MIR3691, LOC101927441, CST13P, ACAP2, ZNF324, SH3PXD2B, TBX21* expression in maternal blood as biomarkers of spontaneous preterm birth.
2. To externally validate a prediction model for spontaneous preterm birth using maternal blood gene expression data with machine and statistical learning.

Methods

Biological samples and validation of top biomarkers

To test the reproducibility of top biomarkers identified in the literature, historical biological samples were collected from the All Our Families cohort [28-30]. In brief, participants were recruited between May 1st, 2008 and December 31st, 2011 at <25 weeks gestation and provided consent for blood sample collection, and complete questionnaires including information related to demographics, emotional and physical health. Participants provided informed written consent at the time of recruitment from healthcare offices, community and through Calgary Laboratory services and were provided copies of their consent forms for their records. This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary #REB15-0248 Predicting Preterm Birth Study. Biological samples were collected at two points in pregnancy, timepoint 1 (T1) at 17-23 weeks gestation and timepoint 2 (T2) 28-32 weeks gestation. Maternal whole blood was collected directly into four separate PAXgene blood RNA tubes which were then stored at -80°C prior to RNA isolation (PAXgene Blood RNA Kit, Qiagen). Samples were de-identified prior to data collection. Two tubes were previously used for the original prediction modelling and biomarker identification by Heng et al., [22], a third tube

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

was used to assess RNA integrity in long term storage [23], and a fourth was collected from storage from August 16th to November 3rd, 2018 for use in the current study. For the current study, n=47 participants who subsequently had an sPTB (<37 weeks) were included (n=44 T1, n=42 T2 samples), in addition to n=45 participants who had a healthy term (38-42 weeks) delivery (n=40 T1, n=44 T2). A total of n=13 samples were missing from storage or insufficient sample remaining (n=3 T1 sPTB, n=3 T2 sPTB, n=5 T1 term, n=2 T2 term), and n=2 T2 samples in the sPTB group were not included as delivery occurred prior to the second sample collection. Maternal blood samples were collected, and RNA was isolated according to manufacturer's instructions (RNAeasy minikit, Qiagen). The following genes were measured using a probe-based assay (Quantigene, Invitrogen, ThermoFisher Scientific), which uses identical probes to an Affymetrix microarray chip: *LOC100128908* (*LMLN2), LOC101927441, CST13P, ACAP2, ZNF324, SH3PXD2B, TBX21.* Due to limitations with measuring microRNA (miRNA) using the assay, *MIR3691* was not measured.

Novel prediction model

Population and expression dataset

To test whether machine learning could improve predictive performance, a secondary analysis of the maternal blood microarray data, as previously published [22], was conducted. Gene expression data was downloaded as raw Affymetrix Chip output files from the National Center for Biotechnology Information Gene Expression Omnibus (accession number: GSE59491) (All Our Families, AOF- Calgary cohort) [22]. The dataset used herein contains high-throughput expression data from n=165 subjects (n=51 sPTB, n=114 matched term delivery

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

controls) nested within the Calgary AOF cohort. Matched gene expression data from two timepoints, 17-23 weeks (T1) and 28-33 weeks (T2) was available for each participant. Two observations from the sPTB group were removed from the dataset as these deliveries occurred prior to the T2 collection and therefore have only one expression dataset. An external dataset was additionally identified from a pregnancy cohort based in Detroit, USA [31], which collected maternal blood samples at comparable timepoints for gene expression analysis, (accession number: GSE149440), and was used for external validation of the model. Participants that had at least two matched blood samples collected within the same two timeframes (T1 and T2) were selected from within the Detroit cohort dataset for analysis, for a total of n=98 subjects (n=34 sPTB and n=64 matched term delivery controls) included for external validation.

Differential expression analysis

The Calgary dataset was randomly split into 80:20 training and test sets, and the differential expression analysis was conducted on the training set, with 2 times 5-fold cross validation. Differential expression analysis was performed initially as described previously [22]. In brief, differential expression was explored using the following comparisons:

- sPTB group compared to term group at T1
- sPTB group compared to term group at T2
- T1 compared to T2 in sPTB group
- T1 compared to T2 in term group
- dT (T2-T1) in sPTB compared to term

Genes with a family-wise error rate less than 0.05 were considered differentially expressed and kept for downstream modelling.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Feature selection analysis

Three features from each differentially expressed gene were fed into the downstream modelling pipeline: the log2 of its intensity value at T1 (T1), the log2 of its intensity value at T2 (T2), and the difference between these two measurements (T2-T1, or dT). Input data for feature selection included all genetic features and their associated target labels (sPTB or term group) for the training set. Feature selection was conducted using a supervised stepwise feature selection approach using two times five-fold cross validation. In brief, this stepwise additive approach to feature selection iteratively includes each feature and retains only those features that significantly improve the training model. The feature selection algorithm assigns a gene score for each feature as a measure of relative importance, and subsequently discards non-explanatory or noisy features.

Model training and testing

Each training set was fitted using two learning algorithms: logistic regression (LR) and multilayer perceptron artificial neural network (MLP). Hyperparameters for MLP model training were selected using the *Hyperopt* package, and models were evaluated using two repeats of five-fold stratified cross validation. The resultant two models were assessed for predictive performance by fitting them on the internal (Calgary) test set, or the external (Detroit) dataset. Full details on the computational methods are described elsewhere [32].

Results

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Validation of biomarkers of preterm birth

Demographic characteristics of the population used for biomarker validation are described in Table 1. Participants with an sPTB did not significantly differ from the term group in age, smoking status, alcohol use during pregnancy, history of abortion, history of PTB, gravidity or parity. Of the biomarkers measured, only five of seven were detectable in the study population (Table 2, S2 Table). Samples were tested at four concentrations (1.875, 3.75, 6.25, 25ng/uL RNA standards). Biomarkers *CST13P* and *LMLN2* were below the limit of detection (<LOD) in over 50% of the population (68% and 56% respectively) at all concentrations and thus were excluded from further analysis. One sample (term T2) was <LOD across all biomarkers, which was likely a technical issue with sample processing and thus excluded. Levels of *SH3PXD2B* were <LOD in 22% of the population and those <LOD were assigned as one half of the basement level (3 MFI, mean fluorescence index units). The remaining four biomarkers were present above the limit of detection in all samples.

Table 1. Patient demographic and clinical characteristics

	sPTB (n=47)	**Term (n=45)**
Maternal age (years)	31.5[30.2-32.8]	31.8[30.6-33.0]
Maternal ethnicity		
White	37, 79[64-89]%	35, 78[63-89]%
Non-white	10, 21[11-36]%	10, 22[11-37]%
Smoking during pregnancy		
Yes	5, 11[4-23]%	10, 22[11-37]%
No	36, 77[62-88]%	34, 76[60-87]%
Unknown	6, 13[5-26]%	1, 2[0.06-12]%
Alcohol during pregnancy		
Yes	17, 36[23-52]%	24, 53[38-68]%
No	22, 47[32-62]%	18, 40[26-56]%
Unknown	8, 17[8-31]%	3, 7[1-18]%
History of abortion		

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Yes **No**	4, 9[2-20]% 43, 91[80-98]%	8, 18[8-32]% 37, 82[68-92]%
History of PTB **Yes** **No**	 8, 17[8-31]% 39, 83[69-92]%	 2, 4[0.5-15]% 43, 96[85-99]%
Gravidity	1.9[1.6-2.2]	1.9[1.7-2.2]
Parity	0.6[0.3-0.8]	0.5[0.4-0.7]
Gestational age at delivery (weeks)	33.9[33.2-34.6]	39.2[38.9-39.4]

Values are represented as mean[95% confidence interval] for continuous variables or n, %[95% confidence interval] for categorical variables.

Table 2. Biomarker levels in maternal blood

		ACAP2	***LOC101927441***	***ZNF324***	***SH3PXD2B***	***TBX21***
TERM	**T1 (n=40)**	11649[10656-12641]	546[479-613]	572[513-631]	27[18-36]	1128[100
	T2 (n=43)	10711[9727-11695]	465[397-533]	484[432-536]	24[17-31]	989[852-1
SPTL	**T1 (n=44)**	9868[8744-10629]	422[364-480]	444[394-494]	21[17-25]	895[757-1
	T2 (n=42)	11827[10049-13605]	471[388-554]	515[443-587]	39[26-52]	1042[864-

Values are represented as mean [95% confidence interval] of mean fluorescence index MFI.

Four of the five measured biomarkers, *ACAP2* (p=0.0068), *LOC101927441* (p=0.0082), *ZNF324* (p=0.0019), and *TBX21* (p=0.0182) exhibited significantly lower levels in the sPTL group compared to the term group at T1, and not at T2. When assessing biomarkers as a measurement of T2/T1 ratios, *ACAP2* (p=0.0074), *LOC101927441* (p=0.0273), *ZNF324* (p=0.0170), *TBX21* (p=0.0119) ratios were significantly higher in the sPTL group than the term group, suggesting a greater trajectory of increased expression through gestation in those with sPTL (Fig 1). Though we do observe some differences in biomarker levels between sPTL and

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

term samples, only five of the eight originally identified biomarkers [22] could be measured using the same population and methodology, and only four exhibited significant differences between term and preterm groups.

Figure 1. Biomarkers of preterm birth. Values are reported as mean fluorescence index (MFI) at either timepoint or a ratio of MFI values at T2 over T1. Analysed by one-way ANOVA followed by Dunnett correction for multiple comparisons. *p-value<0.05, **p-value<0.01. *ACAP2*: ArfGAP with coiled-coil, ankyrin repeat and PH domains 2, *ZNF324*: zinc finger protein 324. *SH3PCD2B*: SH3 and PX domains 2B, *TBX21*: T-box transcription factor 21.

Novel prediction model: Feature selection

In the interest of identifying other potential biomarkers that can robustly predict sPTB, we conducted feature selection analysis on the publicly available AOF microarray dataset (GSE59491) [22]. The top gene features selected from the complete microarray dataset, along with their assigned gene scores for each iteration of cross validation (two times five-fold cross validation for a total of ten iterations), are represented in Table 3. Notably, the top predictive genes did not show consistency in assigned gene scores across iterations. The topmost explanatory feature as selected by the feature selection algorithm, *FPR3* dT, was assigned a score of 1 (#1 most predictive) and 2 (#2 most predictive) but was assigned a score of zero (uninformative) the remaining eight iterations, indicating that top features are not robust to noise within the dataset.

Table 3. Top overall explanatory features and assigned feature selection scores by iteration

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Feature	***Iteration:***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
***FPR3* dT** Formyl peptide receptor 3		0	0	0	0	0	1	0	2	0	0
***TRBJ2-6* T1** T cell receptor beta joining 2-6		0	0	0	1	0	0	0	0	0	2
***PLD4* T2** phospholipase D family member 4		0	0	0	12	0	0	4	0	0	0
***ZFP14* T2** zinc finger protein		0	0	0	0	1	0	0	0	0	0

Rows represent top four overall explanatory genetic features selected. Columns represent individual analysis (2 times 5-times cross validation for a total of 10 iterations). Values represent predictive score assigned by the feature selection algorithm where 1 is the highest possible value, indicating the most predictive genetic feature and a score of 0 indicates that the feature was not considered predictive for that given iteration.

Model performance for prediction of sPTB

The MLP model showed promising performance in the training set (area under the receiver operating curve, AUC 0.9), and an improvement over traditional LR (AUC 0.85), though performance was notably degraded when applied to the internal test set (AUC 0.54), and further degraded when validated externally (AUC 0.51 MLP, AUC 0.53 LR), which indicates a high degree of overfitting (Table 4).

Table 4. Model Performance

Model	***Training Set***	***Test Set***	***Training AUC***	***Test AUC***
MLP	Calgary	Calgary	0.90	0.54
LR	Calgary	Calgary	0.85	0.54

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

MLP	Calgary	Detroit	0.74	0.51
LR	Calgary	Detroit	0.77	0.53

Model performance as reported by area under the receiver operating curve (AUC) for two prediction algorithms, multilayer perceptron (MLP) and logistic regression (LR) validated internally and externally.

Assessing overfitting

Both LR and MLP models showed significant degradation of performance in both internal and external test sets as compared to training performance, indicating a high degree of overfitting during training, particularly in the MLP model. To test the degree of overfitting, models were retrained using permuted data. In brief, target labels (sPTL or term) were scrambled to remove any potential true pattern within the data before proceeding with model training as before. Using scrambled data, high performance was still observed in the training set, with the highest performance by the MLP algorithm (AUC 0.80). Model performance was degraded when applied to either the internal or external test sets (Table 5).

Table 5. Model performance using permuted data

Model	*Training Set*	*Test Set*	*Training AUC*	*Test AUC*
MLP	Calgary	Calgary	0.80	0.49
LR	Calgary	Calgary	0.75	0.50
MLP	Calgary	Detroit	0.72	0.52
LR	Calgary	Detroit	0.63	0.52

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Model performance as reported by area under the receiver operating curve (AUC) for two prediction algorithms, multilayer perceptron (MLP) and logistic regression (LR) tested internally and validated externally.

Discussion

We were unable to repeat the findings of Heng et al., [22] to predict spontaneous preterm birth using maternal blood gene expression. Most alarmingly, two of the eight topmost predictive genes were not detectable in blood samples from the same patients, suggesting issues with repeatability in probe-based RNA array methods, despite validation of RNA integrity over long-term storage [23]. Indeed, array reliability may be particularly problematic in lowly expressed genes and certain genes may be more subject to poor probe specificity [33, 34], and we were unable to conduct assessment of gene-specific expression levels over long-term storage. Additionally, we were unable to produce a more generalizable model through secondary analysis of the microarray data using machine learning, and our results suggest a high degree of overfitting following external validation. This highlights the importance of repeat and validation studies in order to meaningfully progress the field of preterm birth prediction.

One of the primary limitations to prediction using maternal blood gene expression encountered was overfitting and noise within the dataset, which significantly skewed performance estimates. A noisy dataset likely also contributed to the inability to repeat and/or validate previous findings. Possible consequences of data noise are further exacerbated when using advanced methods such as machine learning, and, as evidenced in the study herein, high complexity/machine learning approaches often do not demonstrate improved predictive

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

performance over traditional statistical methods [35]. Feature selection approaches were unable to effectively reduce the noise within this dataset to obtain clinically useful patterns as markers for spontaneous preterm birth. Many prediction studies, at least in the field of reproduction, are preceded by observational experiments to explore biomarker patterns as possible predictors, such as differential gene expression analysis to identify genes associated with an outcome. Often, these observational experiments are conducted on the whole dataset, not the training set, and as such, prediction models trained on this data are biased by patterns that exist in the test set. This phenomenon, in which information from outside the training dataset is used to create the prediction model, is known as data leakage. Consequently, the training dataset contains information about the outcome that would not be otherwise available when using the model for prediction, artificially overinflating the predictive performance when the model is applied to the test set. Unintentional data leakage likely contributes to the lack of reproducibility in such prediction studies. High levels of noise inherent in gene expression data exacerbated the consequences of data leakage during differential expression analysis and feature selection, highlighting the importance of cautious interpretation and robust methodological safeguards.

To illustrate the extent to which data leakage may impact performance results, we also performed model training in the training set in which differential expression analysis was conducted *before* the splitting the training and test sets (S1 Table). While we observe high predictive performance in the training set (AUC 0.72 with LR, 0.79 with MLP), performance was not significantly degraded in the test set (AUC 0.65 with LR, 0.85 with MLP). Note that with the presence of data leakage, the machine learning MLP model had substantially higher performance (AUC 0.85) as compared to analysis without data leakage (AUC 0.54 for comparable dataset). This stresses the importance of methodological safeguarding and careful

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

study design, to avoid possible sources of bias and data leakage, particularly in omics or similar datasets that are prone to a high degree of noise. External validation is also a highly powerful tool for testing the generalizability of models that may have been subject to data leakage.

Our findings also underscore the broader implications for omics studies for discovery analysis, where high feature-to-observation ratios are common, which exacerbates the challenge for mitigating bias and ensuring the reliability of predictive models. For example, differential expression analysis without appropriate training and testing sets for validation introduces inherent bias and limits the generalizability of patterns identified. Testing on internal test sets alone is insufficient for measuring generalizability, especially in the instance of data leakage. Yet, assessments of overfitting and external validation are not standard practice in preterm birth prediction, and the authors stress their importance for meaningful future work in this field. As such, our study serves as a cautionary tale for researchers, emphasizing the need for transparency, rigorous methodological standards, as well as not only repeating results but validation in external cohorts in order to advance the field of spontaneous preterm birth prediction responsibly.

While maternal blood presents an enticing opportunity for minimally invasive prediction, peripheral blood is subject to a high noise signal from various physiological processes occurring within the body possibly unrelated to uterine function during pregnancy. This stresses the need for improved feature identification. Biological compartments including cervicovaginal fluid, amniotic fluid, and the vaginal microbiome may better reflect the physiology of pregnancy [36, 37] though sample availability of reproductive and gestational tissues for research purposes are limited. An emerging strategy involves the use of cell-free nucleic acid biomarkers, which can be utilized to identify biomarkers with uterine origin in maternal blood for improved prediction of

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

adverse pregnancy outcomes [38, 39]. Considerable research has been conducted to review the most robust predictors for sPTB, including but not limited to inflammatory biomarkers, maternal characteristics and genetic contributions [40-42], yet the most frequently used risk factors in current literature show variable predictive performance and poor robustness [43]. A recent meta-analysis identified the most robust predictors of PTB, including low gestational weight gain, interpregnancy interval following miscarriage <6months, and sleep-disordered breathing [42], and it is likely that combined biomarker approaches are necessary for prediction [44, 45]. Additionally, current literature often does not distinguish those predictors for general PTB from those for sPTB, despite likely distinct aetiologies. It is also worth noting that the pervasive use of convenience sampling in reproductive studies (e.g. secondary analysis of biosamples used for routine antenatal screening) are not necessarily performed proximal to the outcome of interest (PTB). For many subjects, the delay from testing to outcome may make identifying true associations difficult.

Looking ahead, our recommendations for future research include safeguarding against sources of data leakage, implementing cross-validation techniques as a measure of robustness, and prioritizing repeatability and reproducibility of findings. This likely includes incentivizing repeated studies in published literature and improving data management, storage, and sharing infrastructure [10, 11]. Additionally, as unsupervised feature selection techniques were not shown to be beneficial in improving prediction of spontaneous preterm birth, future research in identifying biomarkers for the mechanism preterm labour are important. The best models combine an understanding of the features (such as genes, proteins, or patient characteristics) that are most important for determining the outcome and robust methodologies. In the case of spontaneous preterm birth, this should involve a return to the bench to better elucidate those

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

pathways and biomarkers and their possible contribution to preterm birth outcomes. In better understanding the mechanisms of labour and preterm birth, we stand to better approach its prediction, and consequently, improving maternal and neonatal health outcomes for those impacted by preterm birth.

References

1. Ohuma EO, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. Lancet. 2023;402(10409):1261-71.

2. Purisch SE, Gyamfi-Bannerman C. Epidemiology of preterm birth. Seminars in Perinatology. 2017;41(7):387-91.

3. Perin J, Mulick A, Yeung D, Villavicencio F, Lopez G, Strong KL, et al. Global, regional, and national causes of under-5 mortality in 2000-19: an updated systematic analysis with implications for the Sustainable Development Goals. Lancet Child Adolesc Health. 2022;6(2):106-15.

4. Hornaday KK, Wood EM, Slater DM. Is there a maternal blood biomarker that can predict spontaneous preterm birth prior to labour onset? A systematic review. PLOS ONE. 2022;17(4):e0265853.

5. Marić I, Stevenson DK, Aghaeepour N, Gaudillière B, Wong RJ, Angst MS. Predicting Preterm Birth Using Proteomics. Clin Perinatol. 2024;51(2):391-409.

6. Ramachandran A, Clottey KD, Gordon A, Hyett JA. Prediction and prevention of preterm birth: Quality assessment and systematic review of clinical practice guidelines using the AGREE II framework. Int J Gynaecol Obstet. 2024.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

7. Yang Q, Fan X, Cao X, Hao W, Lu J, Wei J, et al. Reporting and risk of bias of prediction models based on machine learning methods in preterm birth: A systematic review. Acta Obstet Gynecol Scand. 2023;102(1):7-14.

8. Staffa SJ, Zurakowski D. Statistical Development and Validation of Clinical Prediction Models. Anesthesiology. 2021;135(3):396-405.

9. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7.

10. Sharifi-Heris Z, Laitala J, Airola A, Rahmani AM, Bender M. Machine Learning Approach for Preterm Birth Prediction Using Health Records: Systematic Review. JMIR Med Inform. 2022;10(4):e33875.

11. Mennickent D, Rodríguez A, Opazo MC, Riedel CA, Castro E, Eriz-Salinas A, et al. Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications. Front Endocrinol (Lausanne). 2023;14:1130139.

12. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, et al. Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals. Crit Care Med. 2020;48(5):623-33.

13. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. Bmj. 2024;384:e074819.

14. Lenain R, Dantan E, Giral M, Foucher Y, Asar Ö, Naesens M, et al. External Validation of the DynPG for Kidney Transplant Recipients. Transplantation. 2021;105(2):396-403.

15. Russell FM, Herbert A, Kennedy S, Nti B, Powell M, Davis J, et al. External validation of the ultrasound competency assessment tool. AEM Educ Train. 2023;7(3):e10887.

16. Yun JS, Han K, Choi SY, Cha SA, Ahn YB, Ko SH. External validation and clinical application of the predictive model for severe hypoglycemia. Front Endocrinol (Lausanne). 2022;13:1006470.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

17. Slieker RC, van der Heijden A, Siddiqui MK, Langendoen-Gort M, Nijpels G, Herings R, et al. Performance of prediction models for nephropathy in people with type 2 diabetes: systematic review and external validation study. Bmj. 2021;374:n2134.

18. Chaemsaithong P, Sahota DS, Poon LC. First trimester preeclampsia screening and prediction. Am J Obstet Gynecol. 2022;226(2s):S1071-S97.e2.

19. Neary C, Naheed S, McLernon DJ, Black M. Predicting risk of postpartum haemorrhage: a systematic review. Bjog. 2021;128(1):46-53.

20. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Medicine. 2023;21(1):70.

21. Feyaerts D, Marić I, Arck PC, Prins JR, Gomez-Lopez N, Gaudillière B, et al. Predicting Spontaneous Preterm Birth Using the Immunome. Clin Perinatol. 2024;51(2):441-59.

22. Heng YJ, Pennell CE, McDonald SW, Vinturache AE, Xu J, Lee MWF, et al. Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women. PLoS ONE. 2016;11(6).

23. Stephenson NL, Hornaday KK, Doktorchik CTA, Lyon AW, Tough SC, Slater DM. Quality assessment of RNA in long-term storage: The All Our Families biorepository. PLoS One. 2020;15(12):e0242404.

24. Alpaydin E. Introduction to machine learning. Ieee Xplore d, Mit Press p, ebrary I, editors: Cambridge, Massachusetts : MIT Press; Third edition.; 2014.

25. Theodoridis S. Machine learning : a Bayesian and optimization perspective: Amsterdam, Netherlands : Academic Press; First edition.; 2015.

26. Dhar V. Data science and prediction. Commun ACM. 2013;56(12):64–73.

27. Arain Z, Iliodromiti S, Slabaugh G, David AL, Chowdhury TT. Machine learning and disease prediction in obstetrics. Curr Res Physiol. 2023;6:100099.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

28. McDonald CR, Darling AM, Conroy AL, Tran V, Cabrera A, Liles WC, et al. Inflammatory and angiogenic factors at mid-pregnancy are associated with spontaneous preterm birth in a cohort of Tanzanian women. PLoS ONE. 2015;10(8).

29. Gracie SK, Lyon AW, Kehler HL, Pennell CE, Dolan SM, McNeil DA, et al. All Our Babies Cohort Study: recruitment of a cohort to predict women at risk of preterm birth through the examination of gene expression profiles and the environment. BioMed Central;http://www.biomedcentral.com/1471-2393/10/87;University of Calgary;Medicine; 2011.

30. Tough SC, McDonald SW, Collisson BA, Graham SA, Kehler H, Kingston D, et al. Cohort Profile: The All Our Babies pregnancy cohort (AOB). International Journal of Epidemiology. 2017;46(5):1389-90k.

31. Tarca AL, Pataki BÁ, Romero R, Sirota M, Guan Y, Kutum R, et al. Crowdsourcing assessment of maternal blood multi-omics for predicting gestational age and preterm birth. Cell Reports Medicine. 2021;2(6):100323-.

32. Werbicki T. Slater-Lab-SPTB2022. Available from: https://github.com/tywerbicki/Slater-Lab-SPTB.

33. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 2006;22(2):101-9.

34. Kothapalli R, Yoder SJ, Mane S, Loughran TP, Jr. Microarray results: how accurate are they? BMC Bioinformatics. 2002;3:22.

35. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.

36. Chakoory O, Barra V, Rochette E, Blanchon L, Sapin V, Merlin E, et al. DeepMPTB: a vaginal microbiome-based deep neural network as artificial intelligence strategy for efficient preterm birth prediction. Biomarker Research. 2024;12.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

37. Chang Y, Li W, Shen Y, Li S, Chen X. Association between interleukin-6 and preterm birth: a meta-analysis. Ann Med. 2023;55(2):2284384.

38. Cowan AD, Rasmussen M, Jain M, Tribe RM. Predicting Preterm Birth Using Cell-Free Ribonucleic Acid. Clin Perinatol. 2024;51(2):379-89.

39. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature. 2022;602(7898):689-94.

40. Tang ID, Mallia D, Yan Q, Pe'er I, Raja A, Salleb-Aouissi A, et al. A Scoping Review of Preterm Birth Risk Factors. Am J Perinatol. 2024;41(S 01):e2804-e17.

41. Li J, Ge J, Ran N, Zheng C, Fang Y, Fang D, et al. Finding the priority and cluster of inflammatory biomarkers for infectious preterm birth: a systematic review. J Inflamm (Lond). 2023;20(1):25.

42. Mitrogiannis I, Evangelou E, Efthymiou A, Kanavos T, Birbas E, Makrydimas G, et al. Risk factors for preterm birth: an umbrella review of meta-analyses of observational studies. BMC Med. 2023;21(1):494.

43. Ferreira A, Bernardes J, Gonçalves H. Risk Scoring Systems for Preterm Birth and Their Performance: A Systematic Review. J Clin Med. 2023;12(13).

44. Mirzaei A, Hiller BC, Stelzer IA, Thiele K, Tan Y, Becker M. Computational Approaches for Connecting Maternal Stress to Preterm Birth. Clin Perinatol. 2024;51(2):345-60.

45. Creswell L, Rolnik DL, Lindow SW, O'Gorman N. Preterm Birth: Screening and Prediction. Int J Womens Health. 2023;15:1981-97.

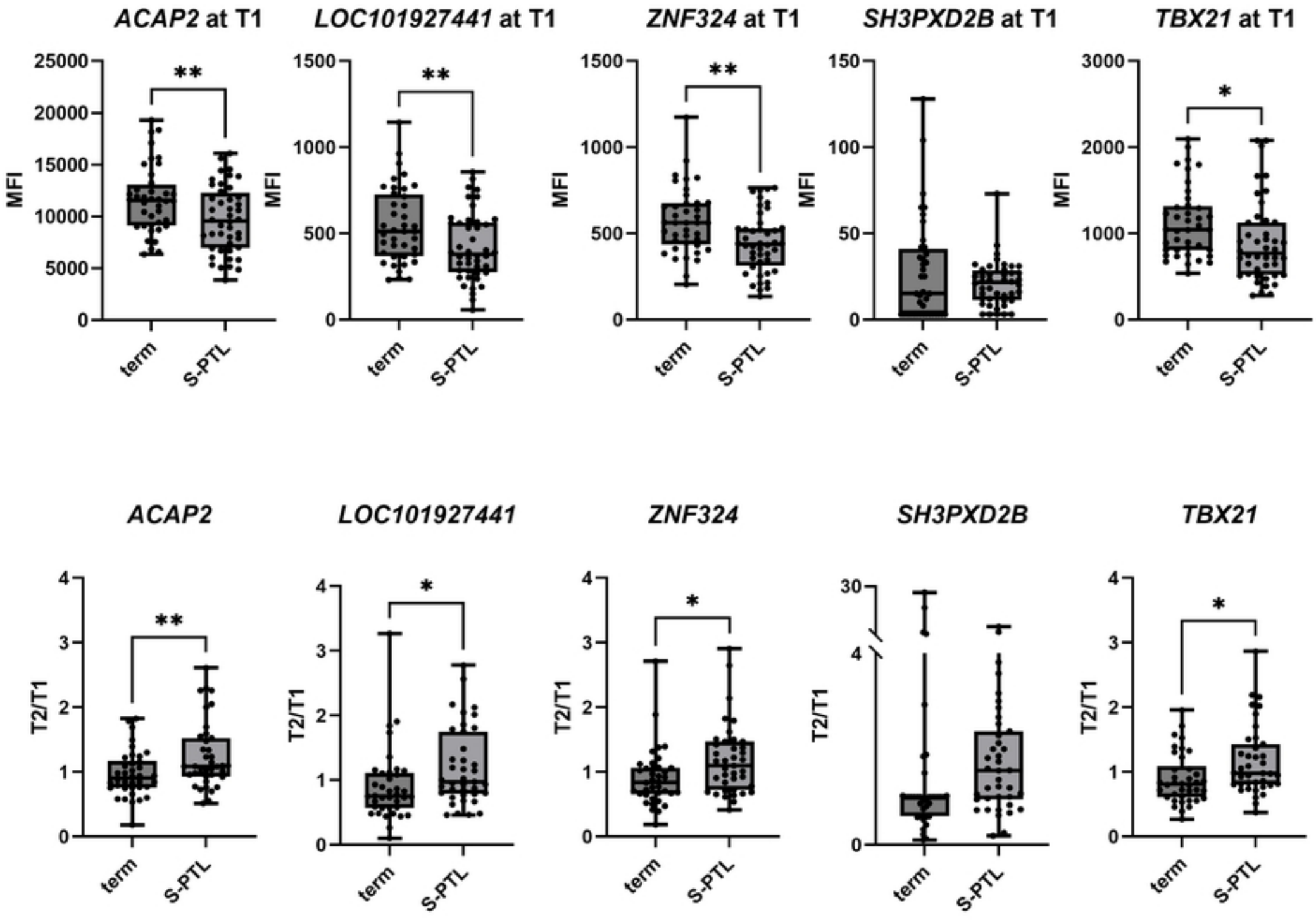
Supporting Information

medRxiv preprint doi: https://doi.org/10.1101/2024.09.11.24313497; this version posted September 12, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

S1 Table. Model performance with data leakage. LR: logistic regression, MLP: multilayer perceptron, AUC: area under the receiver operating curve. Trained on the Calgary dataset and tested on the Calgary test set.

S2 Table. Raw fluorescence index for predictive genes tested in maternal blood. Isolated RNA from whole maternal blood was analyzed for gene expression using a QuantiGene Plex custom assay (Qiagen).


ACAP2 at T1
LOC101927441 at T1
ZNF324 at T1
SH3PXD2B at T1
TBX21 at T1
MFI
term
S-PTL
ACAP2
LOC101927441
ZNF324
SH3PXD2B
TBX21
T2/T1

Figure