Journal of Biomedical Informatics 128 (2022) 104031

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Integrating longitudinal clinical and microbiome data to predict growth faltering in preterm infants

Jose Lugo-Martinez [a,1], Siwei Xu [b,1], Justine Levesque [c], Daniel Gallagher [c], Leslie A. Parker [d], Josef Neu [e], Christopher J. Stewart [f], Janet E. Berrington [f], Nicholas D. Embleton [f], Gregory Young [g], Katherine E. Gregory [h], Misty Good [i], Arti Tandon [c], David Genetti [c], Tracy Warren [c,*], Ziv Bar-Joseph [j,k,*]

[a] Department of Computer Science, University of Puerto Rico, San Juan, PR, USA
[b] School of Information and Computer Sciences, University of California, Irvine, CA, USA
[c] Astarte Medical, Yardley, PA, USA
[d] College of Nursing, University of Florida, Gainesville, FL, USA
[e] Department of Pediatrics, College of Medicine, University of Florida, Gainesville, FL, USA
[f] Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK
[g] Hub for Biotechnology in the Built Environment, Northumbria University, Newcastle upon Tyne, UK
[h] Department of Newborn Medicine, Brigham and Women's Hospital, Boston, Massachusetts, U.S.A. Harvard Medical School, Boston, MA, USA
[i] Division of Newborn Medicine, Washington University School of Medicine, St. Louis, MO, USA
[j] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA
[k] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Keywords:
Neonatal care
Precision nutrition
Integration of clinical and microbiome data
Early identification of growth faltering risk for preterm infants

ABSTRACT

Preterm birth affects more than 10% of all births worldwide. Such infants are much more prone to Growth Faltering (GF), an issue that has been unsolved despite the implementation of numerous interventions aimed at optimizing preterm infant nutrition. To improve the ability for early prediction of GF risk for preterm infants we collected a comprehensive, large, and unique clinical and microbiome dataset from 3 different sites in the US and the UK. We use and extend machine learning methods for GF prediction from clinical data. We next extend graphical models to integrate time series clinical and microbiome data. A model that integrates clinical and microbiome data improves on the ability to predict GF when compared to models using clinical data only. Information on a small subset of the taxa is enough to help improve model accuracy and to predict interventions that can improve outcome. We show that a hierarchical classifier that only uses a subset of the taxa for a subset of the infants is both the most accurate and cost-effective method for GF prediction. Further analysis of the best classifiers enables the prediction of interventions that can improve outcome.

1. Introduction

Preterm birth (<37 completed weeks of gestation) affects more than 10% of all births worldwide [1]. Furthermore, the number of preterm births continues to grow at an alarming rate increasing over 30% in the last 30 years [2]. Preterm birth is recognized as a critical public health concern due to its implications for morbidity and mortality as well as its socio-economic liability [3–5], including persistent health disparities across sub-populations [6,7].

Preterm infants born prior to 34 weeks gestational age are disproportionately at risk of morbidity and mortality due to their immaturity at birth. One of the most persistent health problems these infants experience is growth faltering (GF), typically defined as birth-to-discharge weight z-score decline greater than or equal to 1.2[8–17], in preterm infants. GF is associated with key neonatal morbidities, poor neurodevelopmental outcomes. and cardiometabolic and neurodevelopmental impairment throughout childhood [18–21].

Despite the prevalence of this problem, little is known regarding the

* Corresponding authors at: Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA (Z. Bar-Joseph).
E-mail addresses: tracy@astartemedical.com (T. Warren), zivbj@andrew.cmu.edu (Z. Bar-Joseph).
[1] Contributed equally to this work.

https://doi.org/10.1016/j.jbi.2022.104031
Received 12 September 2021; Received in revised form 20 December 2021; Accepted 14 February 2022
Available online 18 February 2022

mechanisms responsible for postnatal growth faltering. Known contributing factors, such as neonatal illness and inadequate nutrient intake, are not enough for predicting growth outcomes. Recent studies in monkeys and piglets have implicated gut microbiome in ponderal growth and maturation of the brain [22,23]. In the context of preterm infants, the gut microbiome of these infants is influenced heavily by organ immaturity resulting from preterm birth, the neonatal intensive care unit (NICU) environment [24,25], feeding [26], and clinical care (e.g., mode of birth [27], antibiotic exposure [28–31]). Early identification of infants at risk for GF using clinical and microbiome data remains a major unresolved challenge [32].

To address this challenge, we trained prediction models using three time windows to enable early clinical interventions using data from a large international multi-site cohort of preterm infants. Our results show that supervised machine learning methods can be used to effectively predict GF from clinical data.

To make better use of the temporal data, overcome differences in sampling and missing values and to integrate the clinical and microbiome data we next trained Hidden Markov Models (HMMs). Our results indicate that HMMs combining microbiome composition with clinical data result in improved performance when compared to models with clinical data alone. We further extended these models to learn Input/Output HMM (IO-HMM which can be used to predict interventions that can impact growth for specific infants leading to more personalized nutrition and treatments.

2. Materials and methods

2.1. Clinical data

This observational study is composed of four clinical sites: three in the United States (US) and one in the United Kingdom (UK). All preterm infants enrolled in these studies were born < 34 weeks gestational age between 2009 and 2019. Infants enrolled are followed from birth until discharged from the Neonatal Intensive Care Unit (NICU). All clinical data study procedures followed protocols that were approved by site-specific US IRBs (Mass General Brigham Protocol #2016-P-001020, Washington University Protocol #201706182, University of Florida Protocol #201501174) or UK RECs (SERVIS study: permission from North East and N Tyneside 2 #10/H0908/39 and Biobank: permission from North East and N Tyneside 1 #15/NE/0334). Written informed consent was obtained from the parents or guardians of the infants who served as subjects of the investigation. See Supporting Methods for details on how maternal and infant demographic and clinical data were collected from the different sites.

2.2. Microbiome data

Specimen collection Stool samples were obtained for a subset of preterm infants from three sites: two in the US and one in the UK. Samples were collected weekly (on average) from birth until discharged from NICU.

DNA extractions and quantification Bacterial DNA was extracted using the DNeasy PowerSoil Kit following the manufacturer's instructions. Once DNA extractions are complete, the DNA from each sample is stored at -20 °C. After extractions, prior to library prep, each sample of DNA was quantified fluorescently using a Qubit Fluorometer 2.0 and the dsDNA High Sensitivity kit.

Library preparation Following DNA quantification, individual libraries were constructed from each sample. See Supporting Methods for details.

Whole genome sequencing Sequencing was done on the HiSeq X (Illumina) with a target read depth of 10 M reads per sample. A sample sheet used for demultiplexing was created using Illumina Experiment Manager. Once the run was completed, FastQ files were generated for each sample.

Post-processing While the target read depth was 10 M reads per sample, it was determined that a minimum of 5 M reads per sample would be enough for the data to be collected. All samples below 5 M reads were individually normalized to a concentration of 2 nM and then pooled for re-sequencing. Both sets of sequence data were combined for bioinformatic analysis.

2.3. Problem formulation

We consider a binary classification task. Given a set of discharged preterm infants annotated with corresponding z-scores computed from birth weight and discharged weight (Fig. S1), we defined the following classification problem: *Prediction of growth faltering.* This classification problem aims at predicting each preterm infant as *growth faltering* (GF) or *growth normal* (GN), where GF (positive class) is defined as:

$$\text{GF} := \text{discharged z-score} - \text{birth z-score} \leq -1.2$$

and GN is defined as:

$$\text{GN} := \text{discharged z-score} - \text{birth z-score} > -1.2$$

2.4. Training set

Clinical data: We collected a longitudinal data set of preterm infants ($n = 357$) from three NICUs. The data set includes GF ($n = 111$) and GN ($n = 246$) infants. The data was collected from birth until discharge from NICU (354) or transfer to another unit (3). Several attributes were collected for each infant including clinical, and demographic information for both mother and infant, medications, feeding, and probiotics.

Microbiome data: We collected longitudinal stool samples from a subset of these preterm infants ($n = 259$). This set includes preterm infants with GF ($n = 97$) and GN ($n = 162$). Following standard metagenomic analysis (Supporting Methods), 444 microbial taxa were quantified at the species-level for a total of 2,923 samples.

Tables 1 and S1 summarize the attributes used in this study.

2.5. Independent validation set

We generated an independent test set to assess the performance of the clinical classifiers. This independent cohort consists of 135 preterm infants from a new site with a different distribution for GF ($n = 73$) and GN ($n = 62$). This dataset included a subset of clinical attributes: birth z-score, gestational age at birth, maternal age at time of delivery, biweekly body weight, and weekly feeding information (Tables 1 and S1).

2.6. Imputation of missing data

Clinical data: For each preterm infant profile, we imputed missing values for an attribute of interest by using k-nearest neighbors (k-NN) imputation approach [33] with $k = 5$. Specifically, we modified the *KNNImputer* function from *scikit-learn* package (version 0.23.1) such that:

- If the feature is discrete, the imputed value is determined based on the majority among k nearest neighbors, where ties are broken based on neighbor distance.
- If the feature is continuous, the imputed value is defined as the arithmetic mean of all k neighbors.

Microbiome data: In order to circumvent potentially missing or noisy microbiome samples as well as non-uniform sampling rates across infants, we followed prior work and used B-splines for fitting continuous curves to microbial composition time series data [34,35]. Additionally, we removed any infant with fewer than five measured timepoints.

2.7. Construction of feature matrices across different longitudinal periods

All available features across sites were encoded weekly, except for *Body weight* which was only available biweekly for the majority of infants. Next, we learned classifiers for three different periods (or windows). (1) A classifier only using attributes collected at *Birth*, (2) Up to *two weeks* and (3) Up to *one month*.

2.8. Model construction

2.8.1. Clinical-based models

For a given feature set for a specific period, we trained a random forest classifier, with and without imputed data, using the *fitcensemble* function in Matlab (version R2020a) as well as a logistic regression classifier on the imputed data using the *LogisticRegression* function from *scikit-learn* in Python. We performed a parameter sweep to select the best parameters for RF and LR (Supporting Methods). Finally, prediction scores were computed using the *predict* function (Matlab) with the default setting and *predict_proba* function (Python) which reports probability estimated for LR.

2.8.2. Microbiome-based models

Dirichlet Multinomial Mixtures (DMMs): DMM clustering [37] is a probabilistic method for community detection in microbial samples. Importantly, DMM is an infinite mixture model, thus, it can infer the optimal number of clusters (i.e., community types) for a given data set.

In this study, we defined a baseline classifier by first clustering the temporal gut microbiome samples using DMM into six clusters referred to as gut community types (GCTs). We then picked the GCT assignment of the microbiome sample closest to the 7-day interval in each week from week 28 to week 37 such that each infant could be represented by a fixed-length vector comprising these 10 weeks. After one-hot encoding of these categorical vectors, we next imputed any missing GCT assignments using a *k*-nearest neighbor imputer. Finally, we trained a LR classifier to discriminate between GF and GN.

Hidden Markov models (HMMs): We use HMMs to learn a flexible model for classifying longitudinal microbiome data. HMMs are defined using a set of states (GCTs in this work), the transitions between these states and the emissions of each state.

We learn two HMMs, one for GF infants and another for GN.

2.8.3. Initializing HMMs

We set the number of states in the HMMs to six which is the number determined by DMM [37] for this data. Emissions in this model consist of microbiome profiles. The emission values are modeled using Gaussian distributions. We assume that a microbiome profile g_i at state j is normally distributed with mean μ_j and variance ${\sigma_j}^2$,

$$g_{ij} \sim N(\mu_j, {\sigma_j}^2)$$

The microbiome data consists of the relative abundance of 444 bacterial taxa. We have also tested the HMM with a subset of taxa based on the top X as determined by variance, where X ranges from 1 to 25. We did not consider a larger range for X as performance plateaued around X $\geq$ 20.

2.8.4. Learning and inference for integrative HMM

A key challenge in clinical data analysis is missing values. As noted above, several methods for imputation of such data have been proposed and used [33]. HMMs provide an alternative that enables learning a model even when values are missing by integrating over all possible assignments taking into account their probability. This overcomes the challenge of selecting user defined parameters for the imputation (for example, the number k in the k-nearest neighbors approach) and so may lead to much better results.

To allow the use of data with missing values, we modified the standard Expectation Maximization (EM) algorithm used for learning and inference in HMMs. From the original HMM forward algorithm,

$$\alpha_t(i) = P(o_1, o_2, ..., o_t, s_t = i)$$

$$= P(o_t|s_t = i)\sum_{j\in S} P(s_t = i|s_{t-i} = j)\alpha_{t-1}(j)$$

In case that the observation is missing, o_t can be any value. Therefore, using the sum rule of probability,

$$\alpha_t(i) = \sum_{j\in S} P(s_t = i|s_{t-i} = j)\alpha_{t-1}(j)\sum_{k\in R} P(o_t = k|s_t = i)$$

$$= \sum_{j\in S} P(s_t = i|s_{t-i} = j)\alpha_{t-1}(j)\cdot 1$$

$$= \sum_{j\in S} P(s_t = i|s_{t-i} = j)\alpha_{t-1}(j)$$

Similarly, for the backward algorithm, we apply the same modification on the emission probability.

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, ..., o_T|s_t = i)$$

$$= P(o_t|s_t = i)\sum_{j\in S} P(s_{t+1} = i|s_t = j)\beta_{t+1}(j)$$

$$= \sum_{j\in S} P(s_{t+1} = i|s_t = j)\beta_{t+1}(j)$$

Thus, given the set of states S, initial start probability Π, transition probability matrix A, mean μ, and variance σ, the modified E-step now considers the emission probability, $b_j(o_t)$, to be 1 if the sample is missing. All other recurrences remain the same.

$$\alpha_t(i) = \{b_i(o_t)\Pi_i\sum_{j\in S} b_i(o_t)A_{ji}\alpha_{t-1}(j)\sum_{j\in S} A_{ji}\alpha_{t-1}(j)$$

$$\beta_t(i) = \{1\sum_{j\in S} b_i(o_{t+1})A_{ij}\beta_{t+1}(j)\sum_{j\in S} A_{ij}\beta_{t+1}(j)$$

$$S_t(i,j) = \{\frac{\alpha_t(i)A_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\sum_k \alpha_t(k)\beta_t(k)} \frac{\alpha_t(i)A_{ij}\beta_{t+1}(j)}{\sum_k \alpha_t(k)\beta_t(k)}$$

In the M − step, only available microbiome samples are used to re-estimate the means and variances.

$$\mu_i = \frac{\sum_t o_t}{t}$$

$$\sigma_i = \frac{\sum_t (o_t - \mu_t)(o_t - \mu_t)^T}{t}$$

Input/Output Hidden Markov Models: Input/Output HMM (IO-HMM), extend HMMs to enable the determination of the impact of various clinical decisions and interventions (e.g., feeding, medications). In standard HMMs we have a state (unobserved) layer and the emission (observed) layer. In IO-HMMs we have an additional input (observed) layer. See Fig. 1 for an example. The *input layer* encodes possible interventions that can impact transition between states which ultimately can change the outcome of the process. When using this layer, the standard transition probability table used in HMMs is modified to become a *transition probability function*. We use logistic regression to learn that function. Specifically, for each state in the model i we learn a function:

$$f_{j,k}(I) = LR_j(k|I)$$

where I is the input for a specific observation, j and k are states in the model and LR_j is the logistic regression function learned for state j. To learn the logistic regression model for a specific state we modify the M − step by using the Maximum Likelihood Estimate of $S_t(i,j)$ for each time point. with the cross-entropy loss

$$L_i = -\sum_t argmax_j(S_t(i,j))log(LR_i(j|I_t))$$

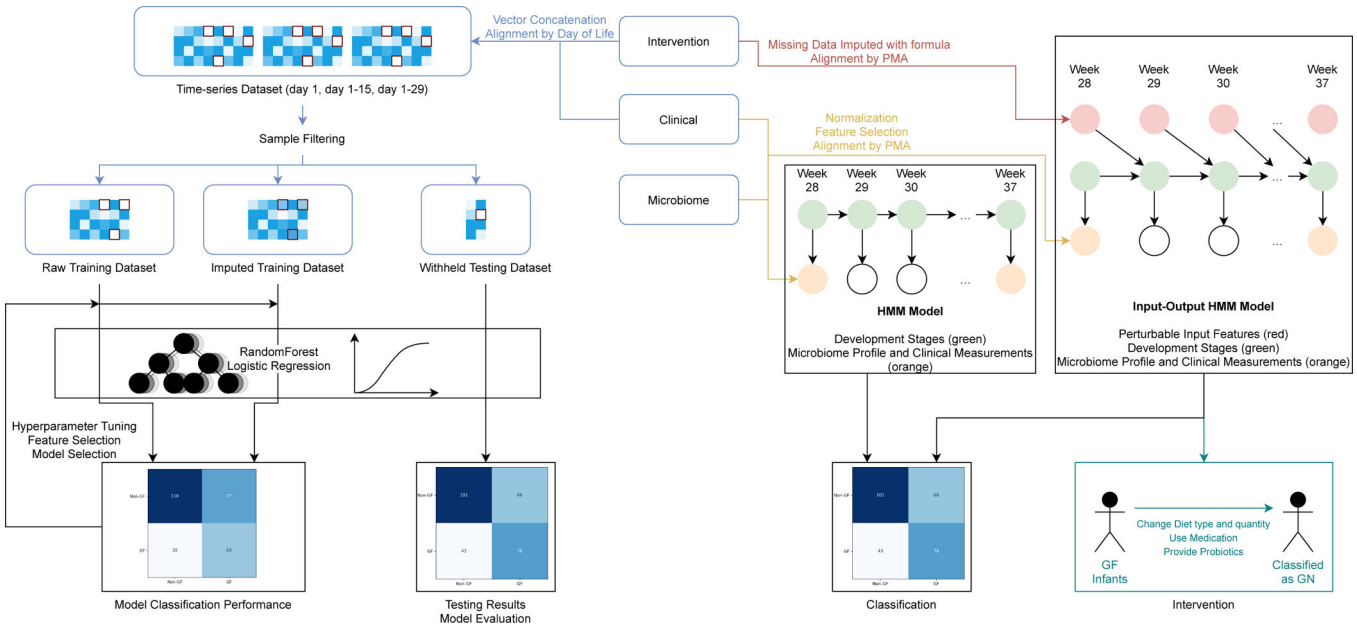

Fig. 1. Overview of the data and methods used. Left: Clinical data based models. The data for these models is from clinical features across each of the time-periods considered in this study: *Birth*, *Two weeks*, and *One month*. We split data to training and test data (with and without data imputation) and use two machine learning models (Random Forest and Logistic Regression) for learning a predictor. Several performance metrics are computed for each of the classifiers and post-model selection of key features is performed. Right: Integrating clinical and microbiome longitudinal data. We use two types of graphical models for integrating these data types. A standard Hidden Markov Model (HMM) (left) and an Input/Output HMM (IO-HMM). For each HMM-based approach, we learn two different models: one for preterm infants with growth failure (GF) and another for those infants with normal growth (GN). Moreover, we evaluate each approach using different performance metrics. Finally, we use the IOHMM model to predict intervention strategies outcomes.

where $LR_i(j|I_t)$ is the probability of transitioning from state i to state j based on the clinical inputs at time t (Supporting Methods). Here we used the input layer to encode possible interventions including different types of feeding, feeding quantity, medication, and probiotics.

2.9. Perturbation analysis on clinical data

We conducted a perturbation analysis of feeding intervention strategies using the clinical classification model. We first trained a LR model for the *One month* time-window. We then systematically changed the value of specific feeding type and quantity features and evaluated the new prediction score from the classifier on each perturbed instance. See Supporting Methods for more details on perturbation analysis and feature selection.

2.10. Evaluation methodology

The performance of each clinical-based method was evaluated through a train/test split, where 20% of preterm infants are randomly selected for the test set and the remaining 80% are used for training. In the case of the HMM-based models, the classification of growth faltering versus growth normal uses two models, one trained with growth failure samples only and one with growth normal samples only. Then, for each sample in the testing dataset, we calculate the likelihood of such a sample being in the growth failure model and that in the growth normal model. The final likelihood is the quotient of the two likelihood values.

3. Results

We developed a workflow for predicting growth faltering in preterm infants from clinical and microbiome data (Fig. 1). We start by computing weekly features from the observed clinical and microbiome data. Next, we replace missing clinical and microbiome data using a *k*-nearest neighbors (k-NN) or probabilistic imputation approach [33,34,35]. We then use clinical and microbiome information to learn various classification models and explore their impact on identifying preterm infants at risk of growth failure.

We applied our methods to study a longitudinal data set from preterm infants across three different sites (n = 357): two US-based sites (n_1 = 128 and n_2 = 50) and a UK site (n_3 = 179). The demographic characteristics and health conditions per site are summarized in Table 1 and Table S1. In addition to clinical data, for a subset of these infants we also obtained temporal stool samples (n = 259; n_1 = 39).

3.1. Classifying using clinical data

We first used supervised machine learning models to predict growth faltering (GF) versus normal growth (GN) at discharge which we defined as [9,10,12,13,17]:

GF:= birth-to-discharge weight z-score ≤ -1.2

GN:= birth-to-discharge weight z-score > -1.2

Fig. S1 shows the distribution of the birth-to-discharge weight z-score changes for the cohort of preterm infants (GF = 111, GN = 246). Differences in data collection between and within sites led to several preterms missing parts of the clinical features used by the classifier. To enable learning using these, we also developed and tested imputation methods to overcome differences in data collection across sites. Fig. S2 shows the accuracy of imputation for different clinical features (i.e., body weight and feeding quantity). We observe that our imputation method is reliable at imputing body weight values: body weight at birth (R^2 = 0.93) and body weight at 29 days of life (R^2 = 0.91). The imputation method is still useful when imputing feeding information though not as accurate as for weight (week 1: R^2 = 0.49 and week 2: R^2 = 0.66).

We next trained our prediction models for three-time windows (Methods): 1) *Birth*, 2) *Two weeks*, and 3) *One month*. For each, we tested different models using random forest (RF) and logistic regression (LR) by randomly splitting the dataset into 80% training and 20% test sets. Model performance was also systematically assessed through a 5-fold

Table 1
Summary of demographic characteristics and health conditions per site. For each site, we list the total number of preterm infants in each class: growth faltering (GF) or normal growth (GN) as well as demographic characteristics and health conditions for each site, if available.

Characteristics	Site (# of GF; # of GN)			
	$n_1 = 128$ (16; 112)	$n_2 = 50$ (12; 38)	$n_3 = 179$ (83; 96)	$n_4 = 135$ (73; 62)
Gestational age (weeks)				
Mean	30.0	29.0	26.9	27.2
Standard deviation	2.7	2.7	2.1	2.0
Range	23.6 to 33.7	24.1 to 34.9	23.0 to 32.0	23.0 to 32.0
Birth body weight (grams)				
Mean	1376	1291	920	914
Standard deviation	574	461	298	182
Range	410 to 3890	500 to 2320	500 to 2000	500 to 1247
Birth z-score				
Mean	−0.17	0.15	−0.05	0.02
Standard deviation	0.87	0.96	0.87	0.95
Range	−2.7 to 4.5	−1.6 to 2.0	−2.3 to 2.9	−2.4 to 3.0
Gender				
Male	60	21	103	72
Female	68	29	76	63
Race				
White	64	31	-	71
African American	24	16	-	59
American Indian/Alaska	1	0	-	0
Native	5	0	-	1
Asian	9	2	-	4
Other				
Ethnicity				
Hispanic	16	0	–	15
Non-Hispanic	94	50	-	120
Mode of delivery				
Vaginal	27	16	85	37
C-section	101	34	94	98
Multiple gestation?				
Yes	50	30	57	33
No	77	20	122	102

$n_2 = 45$ and $n_3 = 175$). Besides the differences in clinical practice between sites, they also vary in the number of infants profiled, the type and frequency of clinical.
information they collected, and the overall number of microbiome samples. Thus, these sites provide a good set to test the generality of our methods.

cross validation on the training set (n = 286: GF = 91, GN = 195) and using an independent test set (n = 71: GF = 20, GN = 51). For each, we computed multiple performance metrics: sensitivity, accuracy, and area under the ROC curve (AUC). Results are summarized in Table 2, Fig. 2, and Fig. S3. We find that LR outperforms RF across all three windows. Specifically, LR classifiers improved AUC performance between 2% and 9% when compared to RF with and without data imputation, respectively. Interestingly, Fig. 2 (left panel) shows that imputation of missing values did not improve model performance for the LR classifier on the test set (*Birth*: 0.75 vs. 0.75; *Two weeks*: 0.72 vs. 0.72; *One month*: 0.76 vs. 0.76). In contrast, results for RF classifiers were impacted by the addition of imputed data (*Birth*: 0.66 vs. 0.75; *Two weeks*: 0.69 vs. 0.72; *One month*: 0.72 vs. 0.76). See Supporting Results for further analysis of the effects of imputation on accuracy.

3.2. Validation on independent test set

To test the generality of our models, we validated their performance on an independent cohort of 135 preterm infants from a third site with a different GF profile from the training set, including fewer infants with GN (n = 62) than with GF (n = 73). As shown in Table 2 and Fig. 2 (right panel), the LR classifier outperforms random forest classifiers and more importantly, performs favorably on the independent validation set for these time-periods (*Birth*: 0.76; *Two weeks*: 0.59; *One month*: 0.75).

3.3. Minimum feature set with feature selection

To enable a faster and more efficient detection of GF we employed a series of feature selection techniques [36] to collect a ranked list of important features. As shown in Fig. S5, the AUC performance values for the selected minimal feature set decreased when compared to the full feature set model across all three periods. Performance differences were small (up to 5%) for the early time points *Birth* and *Two weeks* but increased for *One month* (LR: 0.73 vs. 0.67). This behaviour is likely because the most predictive features are post-menstrual age (PMA) of infant at birth, namely *Birth PMA* (Pearson correlation on class labels: −0.15; P-value: 7.7E-03) and *Birth z-score* (Pearson correlation on class labels: 0.16; P-value: 9.2E-03). Therefore, as time progresses, the importance of the information from these two features decreases. See Fig. S6 and Supporting Results for additional discussion.

3.4. Hidden markov models for integrating clinical and microbiome data

We also collected weekly stool samples for a subset of the preterm infants for which we had clinical data (n = 259: GF = 97, GN = 162). The number of profiled samples per infant ranged from 1 to 64 time-points. To overcome differences in data collection, we focused on samples obtained between weeks 28 and 37 based on PMA (Methods). We next used HMMs to integrate microbiome and clinical data to predict GF in preterm infants (Methods). We learn two different models, one for preterm infants with GF and another for GN. In addition to learning HMM models we have also learned models using an extension of HMMs termed IO-HMMs (Methods) which allows the separation of interventions and observations. We compared our integrative HMM models to a logistic regression model based solely on clinical data but restricted to these 259 infants, and a naive (baseline) approach. The baseline uses Dirichlet Multinomial Mixtures (DMM)[37] to cluster the microbiome samples across all preterm infants into six possible Gut Community Types (GCTs, Methods). Next, for each infant, a fixed-length vector comprising 10 weeks (i.e., between 28 and 37 weeks based on PMA) is created using the GCT weekly assignment. The baseline combines this with clinical data using a logistic regression model.

We observe (Fig. 3) that the HMM approach combining both clinical

Table 2
Summary of results for clinical-based predictive models across data sets. For each data set, we present the total number of preterm infants in each class: growth faltering (GF) or normal growth (GN) as well as three different performance metrics: sensitivity, accuracy and area under the ROC curve (AUC-ROC) for classifiers using clinical data at three different time-periods: *Birth*, *Two weeks* and *One month*.

Dataset	# of infants		Performance metrics								
	GF	GN	Sensitivity			Accuracy			AUC-ROC		
			Birth	*Two weeks*	*One month*	*Birth*	*Two weeks*	*One month*	*Birth*	*Two weeks*	*One month*
Training	91	195	0.68	0.72	0.71	0.61	0.66	0.62	0.64	0.73	0.69
Test	20	51	0.70	0.80	0.80	0.70	0.66	0.68	0.75	0.72	0.76
Validation	73	62	0.66	0.56	0.64	0.68	0.55	0.65	0.76	0.59	0.75

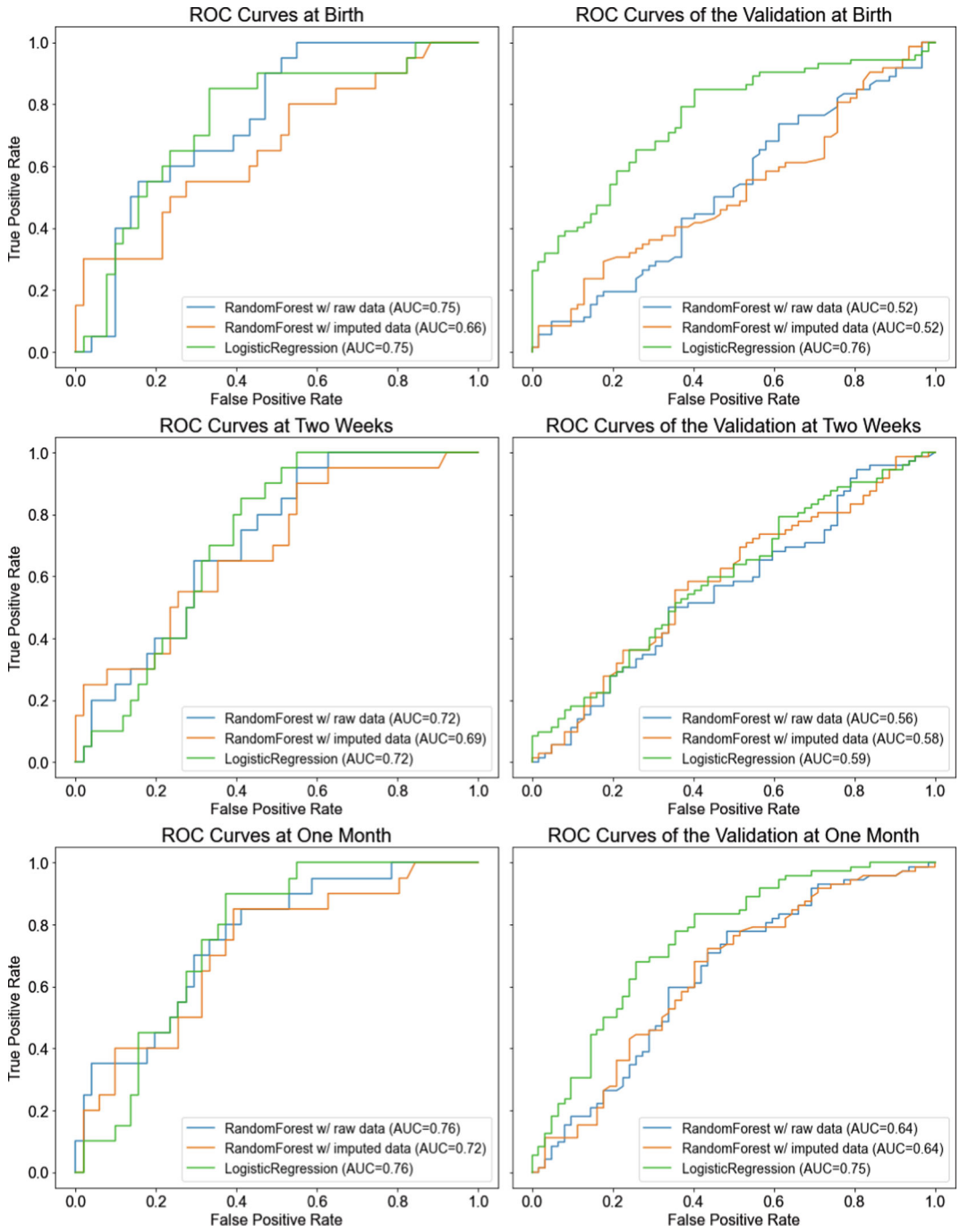

Fig. 2. Performance comparison of clinical-based predictive models for three time-windows. ROC curve and area under the curve (AUC) for the three different classifiers at each time window: *Birth*, *Two weeks*, and *One month*. Left: Figures show the performance results on a held-out test set based on the same sites as the training data. Right: Figures show the performance results on an independent validation cohort from a different site. For two of the three time windows the results for logistic regression on the independent cohort are as good as the results on the training cohorts indicating good generalizability.

and microbiome data outperformed the DMM baseline (AUC: 0.68 vs. 0.61) as well as the clinical data only baseline (AUC: 0.68 vs. 0.64). Furthermore, the combined clinical-microbiome HMM slightly improved on the HMM that only used microbiome data (AUC: 0.68 vs. 0.67). Finally, the extended IO-HMM further improved predictive performance over the combined HMM approach (AUC: 0.70 vs. 0.68).

Reducing the set of 444 taxa to 15 (Methods) can further improve the performance of our methods (AUC: 0.68 vs. 0.66, Fig. S8). The top list of taxa suggests some potentially useful microbiome biomarkers for discriminating between GF and GN infants (e.g., *E. coli* in the learned GF model for GCT 4 or *B. breve* in the learned GN model for GCT 5 Fig. S9 and Supporting Results).

3.5. Two-stage hierarchical model improves overall accuracy

Our results indicate that microbiome data can help improve prediction accuracy, but only for a limited set of infants. We have thus tested a hierarchical strategy which attempts to initially predict GF using clinical data and uses microbiome data only for those infants for which the clinical data does not provide clear prediction. Fig. S10 shows the distribution of predicted scores (x-axis) for the best clinical classifier based on logistic regression (LR). This result suggests that preterm

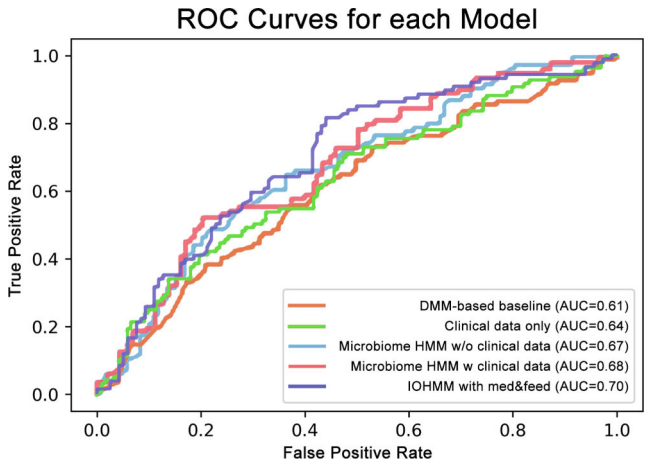

Fig. 3. Results for integrating clinical and microbiome data. For each method, the figure presents the ROC curve and area under the curve (AUC). Baseline microbiome method uses clustering results based on a previously published clustering approach [37] while clinical data only method is based on a logistic regression model trained on the same set of preterm infants.

infants with LR scores $s \leq 0.40$ or $s \geq 0.70$ can be accurately discriminated by the LR classifier (AUC = 0.67) as well as our HMM approach (AUC = 0.71). On the other hand, the remaining preterm infants (i.e., $0.40 < s < 0.70$) are more difficult to predict by the LR model (AUC = 0.59), whereas our proposed HMM approach on this subset leads to better performance (AUC = 0.66). The proposed two-stage hierarchical approach (AUC = 0.73) outperformed both models: LR (AUC = 0.64) and HMM (AUC = 0.68) over the full set of infants (Fig. 4).

3.6. Predicting interventions to reduce GF rate

An advantage of using the IO-HMM approach is the ability to separate the impact of interventions (for example, feeding types) from observations (for example, weight or microbiome composition) in a supervised framework. We tested two potential intervention strategies: (1) feeding-based interventions - we explored the

impact of changing the feeding type to another type (e.g., maternal breastmilk to formula), and (2) medication-based interventions. It is worth noting that the UK site (n_3) does not use donated breastmilk as part of the standard of care, thus, we combined maternal breastmilk and donated breastmilk into a broader type: *human milk*. Figs. 5 and S11 summarize predicted impacts for feeding- and medication-based intervention strategies. The IO-HMM method predicts that for selected preterm babies (based on their microbiome profile) switching to human milk helps increase the likelihood of GF.

4. Discussion

Despite many advances in nutritional care, achieving optimal postnatal growth and nutrient accretion in the NICU remains a major challenge. To date, parenteral nutrition and enteral fortification have been the focus of interventions, but the straightforward strategy of increasing nutrient provision has proven to be inadequate. Early identification of the infants most at risk for GF remains a pediatric health priority.

We assembled a large unique dataset for predicting GF in preterm infants. Our data included longitudinal clinical and microbiome data for 357 infants (259 of which with microbiome samples) from the UK and the US. We tested different classifiers using the clinical data and developed a new classification framework based on Hidden Markov Models (HMMs) to combine the clinical and microbiome data. Our HMM framework combines several desirable features. First, it treats time as a key variable and does not assume independence between time points as other classifiers do [38]. Second, since it is probabilistic it can directly handle noise and missing values. Finally, it can be extended to include not just observations but also interventions using our IO-HMM model.

For our clinical based models, we found that LR with imputation performs best and that a subset of the features provides adequate accuracy. We also observed that as the infants grow older, feeding information (i.e., feeding type and volume) plays a central role in improving model accuracy. In contrast, we found that the predictive ability of clinical attributes such as gestational age, post-menstrual age and birth z-score decreases over time. We have further tested our models on data from an independent cohort. The fact that the accuracy for prediction for this cohort is comparable to the accuracy of our held-out test set is a strong indicator that the model can indeed generalize to unseen cases.

Microbiome information can help to further improve prediction accuracy. We found that using the top 15 species for each sample (a total of

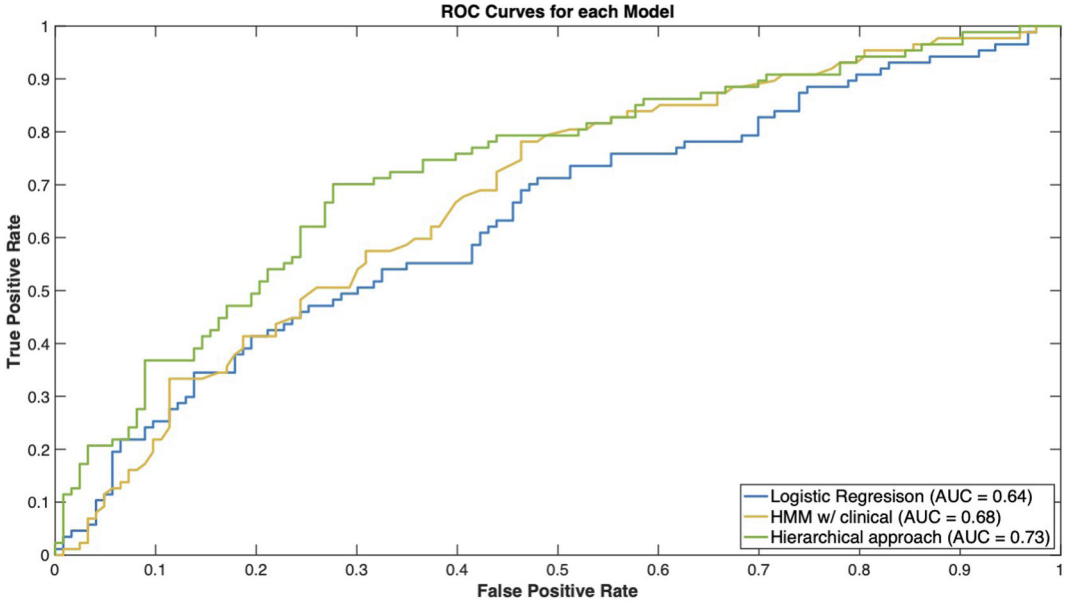

Fig. 4. Hierarchical classification improves prediction accuracy. Figure shows the ROC curve and area under the curve (AUC) for the two-stage hierarchical classifier, as well as both of the corresponding individual clinical- and microbiome-based classifiers.

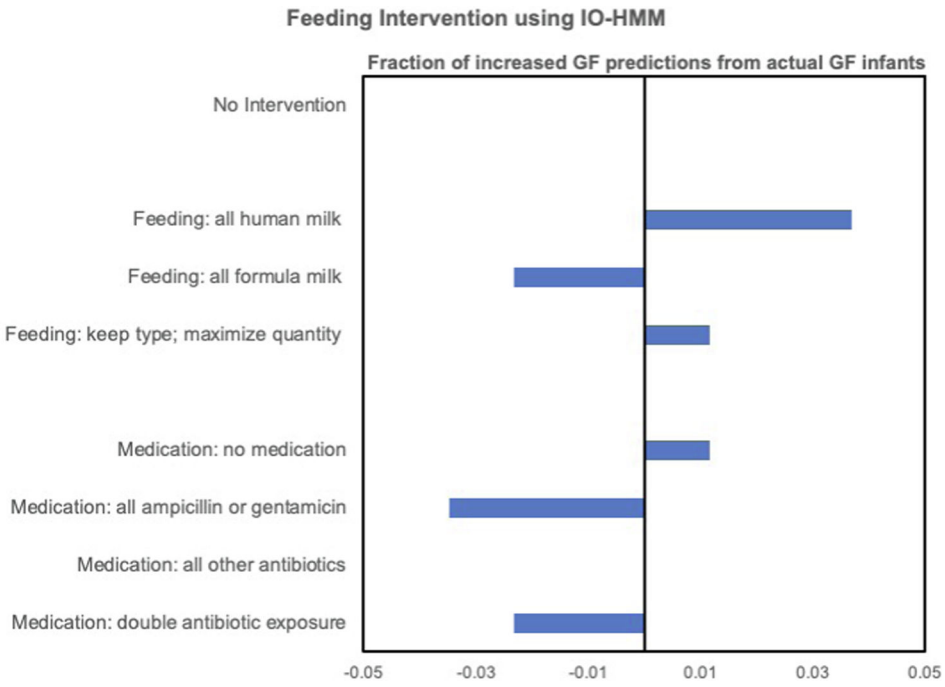

Fig. 5. Summary of feeding- and medication-based interventions predictions based on the Input/Output HMM (IO-HMM). For each intervention type, we list the effect of the intervention strategy as the fraction of preterm infants predicted to be at risk of growth failure after intervention over the observed number of infants at risk of growth faltering. Values to the left of center indicate reduced risk while values to the right indicate elevated risk.

17 of the 444 identified for all samples) is sufficient to accurately discriminate between GF and GN.

Our IO-HMM model allows for *in silico* analysis to determine the impact of various clinical decisions and interventions. Our systematic analysis of intervention strategies using the learned IO-HMM suggests that for certain preterm infants switching to formula milk helps reduce the risk of GF. Parallel to this result, our analysis of feeding interventions using LR (solely based on clinical data, Figs. S12-S13) suggests that for selected preterm babies switching from donated and even maternal breastmilk to formula significantly reduces GF. These results may be driven by the fact that most of the preterm GN infants (61%: 103 out of 168) received formula at least once, whereas only a small fraction of GF infants (26%: 21 out of 81) received formula at least once. Unlike IOHMMs that can distinguish between interventions and observations, the clinical LR model treats both equally and so it strongly associates the formula feeding type with reduced risk of growth faltering.

5. Conclusion

In summary, while predicting GF risk remains a nuanced and challenging task, we demonstrate predictive accuracy with sophisticated models combining clinical and microbiome data. We presented a hierarchical model that can improve accuracy while being cost effective requiring additional data only for those for which results are inconclusive based on initial data.

Data availability statement

Data and code used in this study will be available upon acceptance.

Author contributions

T.W. and Z.B.-J. conceived the project. J.L.-M., S.X., A.T., D.G., T.W. and Z.B.-J. contributed to the design of experiments. J.L.-M. and S.X. developed the computational methods. S.X., J.L.-M. and Z.B.-J. analyzed the data and wrote the first draft of the manuscript. J.L. and D.G. curated multiple data sets for training and validation purposes. C.J.S., J.E.B., N. D.E., G.Y., K.E.G. and M.G. provided clinical and microbiome data for the development of computational models. L.A.P. and J.N. provided data for independent validation of clinical classifier. D.G., A.T., T.W. and Z.B.-J. supervised the project development. All authors contributed to the review and writing of the manuscript.

Funding

Work partially funded by a grant from Astarte Medical to Z.B.-J.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Astarte Medical.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104031.

References

[1] H. Blencowe, S. Cousens, M.Z. Oestergaard, D. Chou, A.-B. Moller, R. Narwal, A. Adler, C. Vera Garcia, S. Rohde, L. Say, J.E. Lawn, National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications, The Lancet. 379 (9832) (2012) 2162–2172, https://doi.org/10.1016/S0140-6736(12)60820-4.

[2] C.K. Shapiro-Mendoza, E.M. Lackritz, Epidemiology of late and moderate preterm birth, Semin. Fetal. Neonatal. Med. 17 (3) (2012) 120–125, https://doi.org/10.1016/j.siny.2012.01.007.

[3] N.-H. Morken, Preterm birth: new data on a global health priority, The Lancet. 379 (9832) (2012) 2128–2130, https://doi.org/10.1016/S0140-6736(12)60857-5.

[4] L. Trasande, P. Malecha, T.M. Attina, Particulate Matter Exposure and Preterm Birth: Estimates of U.S. Attributable Burden and Economic Costs, Environ. Health Perspect. 124 (12) (2016) 1913–1918, https://doi.org/10.1289/ehp.1510810.

[5] M.S. Harrison, R.L. Goldenberg, Global burden of prematurity, Predict. Prev. Preterm. Birth Sequelae. 21 (2) (2016) 74–79, https://doi.org/10.1016/j.siny.2015.12.007.

[6] A.S. Bryant, A. Worjoloh, A.B. Caughey, A.E. Washington, Racial/ethnic disparities in obstetric outcomes and care: prevalence and determinants, Am. J. Obstet. Gynecol. 202 (4) (2010) 335–343, https://doi.org/10.1016/j.ajog.2009.10.864.
[7] H.A. Frey, M.A. Klebanoff, The epidemiology, etiology, and costs of preterm birth, Predict. Prev. Preterm. Birth Sequelae. 21 (2) (2016) 68–73, https://doi.org/10.1016/j.siny.2015.12.011.
[8] L. Sices, D. Wilson-Costello, N. Minich, H. Friedman, M. Hack, Postdischarge growth failure among extremely low birth weight infants: Correlates and consequences, Paediatr. Child Health 12 (1) (2007) 22–28.
[9] M.J. Johnson, S.A. Wootton, A.A. Leaf, A.A. Jackson, Preterm Birth and Body Composition at Term Equivalent Age: A Systematic Review and Meta-analysis, Pediatrics 130 (3) (2012) e640, https://doi.org/10.1542/peds.2011-3379.
[10] Horbar JD, Ehrenkranz RA, Badger GJ, et al. Weight growth velocity and postnatal growth failure in infants 501 to 1500 grams: 2000–2013. https://doi.org/10.1542/peds.2015-0129.
[11] M.M.S. Rover, C.S. Viera, R.C. Silveira, A.T.B. Guimarães, S. Grassiolli, Risk factors associated with growth failure in the follow-up of very low birth weight newborns, J. Pediatr. (Rio J) 92 (3) (2016) 307–313, https://doi.org/10.1016/j.jped.2015.09.006.
[12] I.J. Griffin, D.J. Tancredi, E. Bertino, H.C. Lee, J. Profit, Postnatal growth failure in very low birthweight infants born between 2005 and 2012, Arch. Dis. Child – Fetal. Neonatal. Ed. 101 (1) (2016) 50–55, https://doi.org/10.1136/archdischild-2014-308095.
[13] N. Wells, T.A. Stokes, K. Ottolini, C.H. Olsen, A.R. Spitzer, C.E. Hunt, Anthropometric trends from 1997 to 2012 in infants born at ⩽28 weeks' gestation or less, J. Perinatol. 37 (5) (2017) 521–526, https://doi.org/10.1038/jp.2016.244.
[14] L. Simon, M. Hanf, A. Frondas-Chauty, D. Darmaun, V. Rouger, G. Gascoin, C. Flamant, S. Nusinovici, J.-C. Rozé, U. Simeoni, Neonatal growth velocity of preterm infants: The weight Z-score change versus Patel exponential model, PLoS ONE 14 (6) (2019) e0218746, https://doi.org/10.1371/journal.pone.0218746.
[15] J.H. Chou, S. Roumiantsev, R. Singh, PediTools Electronic Growth Chart Calculators: Applications in Clinical Care, Research, and Quality Improvement, J. Med. Internet. Res. 22 (1) (2020) e16204, https://doi.org/10.2196/16204.
[16] S. Aneja, P. Kumar, T.S. Choudhary, A. Srivastava, R. Chowdhury, S. Taneja, N. Bhandari, A. Daniel, P. Menon, H. Chellani, R. Bahl, M.K. Bhan, Growth faltering in early infancy: highlights from a two-day scientific consultation, BMC Proc. 14 (S12) (2020), https://doi.org/10.1186/s12919-020-00195-z.
[17] T.R. Fenton, B. Cormack, D. Goldberg, R. Nasser, B. Alshaikh, M. Eliasziw, W. W. Hay, A. Hoyos, D. Anderson, F. Bloomfield, I. Griffin, N. Embleton, N. Rochow, S. Taylor, T. Senterre, R.J. Schanler, S. Elmrayed, S. Groh-Wargo, D. Adamkin, P. S. Shah, "Extrauterine growth restriction" and "postnatal growth failure" are misnomers for preterm infants, J. Perinatol. 40 (5) (2020) 704–714, https://doi.org/10.1038/s41372-020-0658-5.
[18] J. Rotteveel, M.M. van Weissenbruch, J.W.R. Twisk, H.A. Delemarre-Van de Waal, Infant and childhood growth patterns, insulin sensitivity, and blood pressure in prematurely born young adults, Pediatrics 122 (2) (2008) 313–321, https://doi.org/10.1542/peds.2007-2012.
[19] M.B. Belfort, S.L. Rifas-Shiman, T. Sullivan, C.T. Collins, A.J. McPhee, P. Ryan, K. P. Kleinman, M.W. Gillman, R.A. Gibson, M. Makrides, Infant Growth Before and After Term: Effects on Neurodevelopment in Preterm Infants, Pediatrics 128 (4) (2011) e899–e906, https://doi.org/10.1542/peds.2011-0282.
[20] G.F. Kerkhof, R.H. Willemsen, R.W.J. Leunissen, P.E. Breukhoven, A.C.S. Hokken-Koelega, Health Profile of Young Adults Born Preterm: Negative Effects of Rapid Weight Gain in Early Life, J. Clin. Endocrinol. Metab. 97 (12) (2012) 4498–4506, https://doi.org/10.1210/jc.2012-1716.
[21] K.A. Bell, L.G. Matthews, S. Cherkerzian, C. Palmer, K. Drouin, H.L. Pepin, D. Ellard, T.E. Inder, S.E. Ramel, M.B. Belfort, Associations of Growth and Body Composition with Brain Size in Preterm Infants, J. Pediatr. 214 (2019) 20–26.e2, https://doi.org/10.1016/j.jpeds.2019.06.062.
[22] D.N. Rendina, G.R. Lubach, M. Lyte, G.J. Phillips, A. Gosain, J.F. Pierre, R. M. Vlasova, M.A. Styner, C.L. Coe, Proteobacteria abundance during nursing predicts physical growth and brain volume at one year of age in young rhesus monkeys, FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 35 (6) (2021), https://doi.org/10.1096/fj.202002162R.
[23] H.-W. Chang, N.P. McNulty, M.C. Hibberd, et al., Gut microbiome contributions to altered metabolism in a pig model of undernutrition, Proc. Natl. Acad. Sci. U. S. A. 118 (21) (2021), https://doi.org/10.1073/pnas.2024446118.
[24] J. Santos, S.E. Pearce, A. Stroustrup, Impact of hospital-based environmental exposures on neurodevelopmental outcomes of preterm infants, Curr. Opin. Pediatr. 27 (2) (2015). https://journals.lww.com/co-pediatrics/Fulltext/2015/04000/Impact_of_hospital_based_environmental_exposures.18.aspx.
[25] X. Cong, J. Wu, D. Vittner, W. Xu, N. Hussain, S. Galvin, M. Fitzsimons, J. M. McGrath, W.A. Henderson, The impact of cumulative pain/stress on neurobehavioral development of preterm infants in the NICU, Early Hum. Dev. 108 (2017) 9–16, https://doi.org/10.1016/j.earlhumdev.2017.03.003.
[26] B. Brooks, M.R. Olm, B.A. Firek, R. Baker, B.C. Thomas, M.J. Morowitz, J. F. Banfield, Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome, Nat. Commun. 8 (1) (2017), https://doi.org/10.1038/s41467-017-02018-w.
[27] M. Reyman, M.A. van Houten, D. van Baarle, A.A.T.M. Bosch, W.H. Man, M.L.J. N. Chu, K. Arp, R.L. Watson, E.A.M. Sanders, S. Fuentes, D. Bogaert, Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life, Nat. Commun. 10 (1) (2019), https://doi.org/10.1038/s41467-019-13014-7.
[28] W.A. Bowes, The role of antibiotics in the prevention of preterm birth, F1000 Med Rep. 1 (2009) 22, https://doi.org/10.3410/M1-22.
[29] A.J. Gasparrini, T.S. Crofts, M.K. Gibson, P.I. Tarr, B.B. Warner, G. Dantas, Antibiotic perturbation of the preterm infant gut microbiome and resistome, Gut Microbes. 7 (5) (2016) 443–449, https://doi.org/10.1080/19490976.2016.1218584.
[30] Z.-H. Zou, D. Liu, H.-D. Li, D.-P. Zhu, Y.u. He, T. Hou, J.-L. Yu, Prenatal and postnatal antibiotic exposure influences the gut microbiota of preterm infants in neonatal intensive care units, Ann. Clin. Microbiol. Antimicrob. 17 (1) (2018), https://doi.org/10.1186/s12941-018-0264-y.
[31] H.-Y. Chang, J.-S. Chiang Chiau, Y.-H. Ho, J.-H. Chang, K.-N. Tsai, C.-Y. Liu, C.-H. Hsu, C.-Y. Lin, M.-J. Ko, H.-C. Lee, Impact of Early Empiric Antibiotic Regimens on the Gut Microbiota in Very Low Birth Weight Preterm Infants: An Observational Study, Front. Pediatr. 9 (2021), https://doi.org/10.3389/fped.2021.651713.
[32] A. Kakaroukas, M. Abrahamse-Berkeveld, J.E. Berrington, R.J.Q. McNally, C. J. Stewart, N.D. Embleton, R.M. van Elburg, An Observational Cohort Study and Nested Randomized Controlled Trial on Nutrition and Growth Outcomes in Moderate and Late Preterm Infants (FLAMINGO), Front. Nutr. 8 (2021), https://doi.org/10.3389/fnut.2021.561419.
[33] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinforma. Oxf. Engl. 17 (6) (2001) 520–525, https://doi.org/10.1093/bioinformatics/17.6.520.
[34] J. Lugo-Martinez, D. Ruiz-Perez, G. Narasimhan, Z. Bar-Joseph, Dynamic interaction network inference from longitudinal microbiome data, Microbiome 7 (1) (2019) 54, https://doi.org/10.1186/s40168-019-0660-3.
[35] D. Ruiz-Perez, J. Lugo-Martinez, N. Bourguignon, K. Mathee, B. Lerner, Z. Bar-Joseph, G. Narasimhan, T. Korem, Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data, mSystems 6 (2) (2021), https://doi.org/10.1128/mSystems.01105-20.
[36] W. Awada, T.M. Khoshgoftaar, D. Dittman, R. Wald, A. Napolitano, A review of the stability of feature selection techniques for bioinformatics data, in: , 2012, pp. 356–363. https://doi.org/10.1109/IRI.2012.6303031.
[37] I. Holmes, K. Harris, C. Quince, J.A. Gilbert, Dirichlet multinomial mixtures: generative models for microbial metagenomics, PLoS ONE 7 (2) (2012) e30126, https://doi.org/10.1371/journal.pone.0030126.
[38] Z. Bar-Joseph, A. Gitter, I. Simon, Studying and modelling dynamic biological processes using time-series gene expression data, Nat. Rev. Genet. 13 (8) (2012) 552–564.