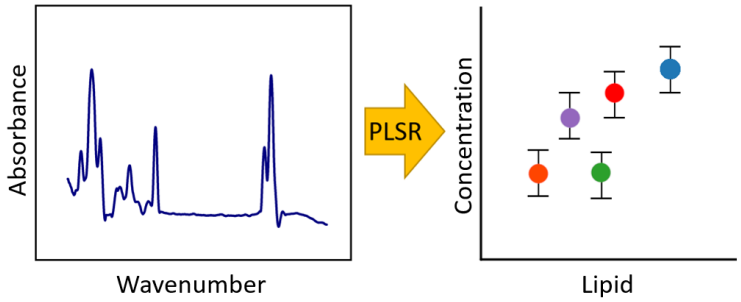
# Graphical Abstract

**Towards Quantifying Biomarkers for Respiratory Distress in Preterm Infants: Machine Learning on Mid Infrared Spectroscopy of Lipid Mixtures**

Waseem Ahmed, Aneesh Vincent Veluthandath, Jens Madsen, Howard W. Clark, Ahilanandan Dushianthan, Anthony D. Postle, James S. Wilkinson, Ganapathy Senthil Murugan



# Highlights

**Towards Quantifying Biomarkers for Respiratory Distress in Preterm Infants: Machine Learning on Mid Infrared Spectroscopy of Lipid Mixtures**

Waseem Ahmed, Aneesh Vincent Veluthandath, Jens Madsen, Howard W. Clark, Ahilanandan Dushianthan, Anthony D. Postle, James S. Wilkinson, Ganapathy Senthil Murugan

- Comprehensive calibration of PLSR models using physiological concentrations of lung surfactant lipids
- Prediction intervals for quantified uncertainty in PLSR models
- Use of SHAP values to explain strength of AI model features with a view to optimising a spectroscopic point of care platform.

# Towards Quantifying Biomarkers for Respiratory Distress in Preterm Infants: Machine Learning on Mid Infrared Spectroscopy of Lipid Mixtures

Waseem Ahmed[a], Aneesh Vincent Veluthandath[a], Jens Madsen[b], Howard W. Clark[b], Ahilanandan Dushianthan[c], Anthony D. Postle[d], James S. Wilkinson[a], Ganapathy Senthil Murugan[a]

[a]*Optoelectronics Research Centre, University of Southampton, Southampton, SO17 1BJ, Hampshire, UK*
[b]*Neonatology, Faculty of Population Health Sciences, EGA Institute for Women's Health, University College London, London, WC1E 6AU, London, UK*
[c]*Perioperative and Critical Care Theme, NIHR Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust, Southampton, SO16 6YD, Hampshire, UK*
[d]*Academic Unit of Clinical & Experimental Sciences, Faculty of Medicine, Southampton General Hospital, Southampton, SO16 6YD, Hampshire, UK*

**Abstract**

Neonatal respiratory distress syndrome (nRDS) is a challenging condition to diagnose which can lead to delays in receiving appropriate treatment. Mid infrared (IR) spectroscopy is capable of measuring the concentrations of two diagnostic nRDS biomarkers, lecithin (L) and sphingomyelin (S) with the potential for point of care (POC) diagnosis and monitoring. The effects of varying other lipid species present in lung surfactant on the mid IR spectra used to train machine learning models are explored. This study presents a lung lipid model of five lipids present in lung surfactant and varies each in a systematic approach to evaluate the ability of machine learning models to predict the lipid concentrations, the L/S ratio and to quantify the uncertainty in the predictions using the jackknife+-after-bootstrap and variant bootstrap methods. We establish the L/S ratio can be determined with an uncertainty of approximately $\pm 0.3$ moles/mole and we further identify the 5 most prominent wavenumbers associated with each machine learning model.

*Email address:* waseem.ahmed@soton.ac.uk (Waseem Ahmed)

*Keywords:* ATR-FTIR, Machine, Learning, PLSR, Lipid, SHAP values, nRDS

## 1. Introduction

Mid-infrared (mid-IR) spectroscopy has been posited as a platform for rapid, point of care diagnostic tool which can interrogate samples in the functional group and fingerprint spectral regions[1, 2] . This allows for both qualitative and quantitative determination of biomarkers to provide a clinician with useful information to consider during diagnosis and prognostication. Attenuated total reflectance Fourier Transform infrared spectroscopy (ATR-FTIR) is one method by which such information may be obtained. Spectroscopic analysis of biological samples followed by machine learning of collected spectra provides an opportunity to apply a more robust multivariate analysis, using all relevant information from multiple spectral peaks, to give quantitative estimates of biomarker concentrations without the need for an expert end-user.

Neonatal respiratory distress syndrome (nRDS) is a condition affecting pre-term neonates due to immature lungs and surfactant deficiency. The resulting increased surface tension within the alveoli means that there is a greater effort required to breathe[3], and intervention is required to reduce it. Treatment usually consists of exogenous surfactant replacement and, if required, mechanical ventilation[4]. While some neonates respond positively to such treatment, there is a subset that do not. Early treatment for positive responders is linked with better survival[5], so differentiating between these two groups is key to ensuring the best patient outcome[6]. Moreover, repeated doses of exogenous surfactant may be required in some neonates but currently there are no established rapid methods to assess ongoing surfactant deficiency.

Lung surfactant is a complex mix of proteins (approx. 10%) and lipids (approx. 90%) and its ability to lower the surface tension in lungs is related to its composition[7, 8]. The compositions of lung surfactant for both healthy babies and neonates suffering from nRDS have been previously compiled and reported[9, 10, 11, 12, 13] and show that there is a proportionately lower ratio of phosphatidylcholine (PC) component (also known as lecithin) to sphingomyelin (L/S ratio) present in cases of nRDS. Lecithin (between approximately 63 to 80% of the total phospholipids) is the primary constituent responsible for reducing lung surface tension, the largest proportion

of which is dipamitoylphosphatidylcholine (DPPC) (between 40 to 55% of the PC fraction[14]), with the next most prolific being palmitoyloleoylphosphatidylcholine (POPC) [15] which comprises around 12% of the PC fraction. The DPPC concentration is known to increase with increasing lung maturity[16], while the concentration of another lipid, sphingomyelin (S), remains relatively constant. In late gestation, surfactant phosphatidylglycerol (PG) (present at concentrations around 10%) is observed to increase with a concomitant decrease in phosphatidylinositol (PI) [13] (present in smaller amounts), heralding the onset of maturity. Cholesterol (Chol) is present at concentrations up to 15% [17] of the total lipid concentration and comprises the majority of the neutral lipid component.

A diagnosis of nRDS may be made on the basis of the ratio of these biomarkers, so giving rise to the use of the lecithin/sphingomyelin ratio (L/S ratio). To discern between the healthy and nRDS states it is necessary to identify a particular L/S ratio which is considered to be the "cut-off" value below which the baby is considered to have nRDS and requires appropriate treatment. This cut-off is often decided by analysing and balancing the sensitivity (true positive rate), and specificity (false negative rate) for the ability of the biomarker to predict the disease[18, 19]. The value of L/S ratio cut-off has been the subject of discussion in the literature [20, 21, 22], but has been reported as 2.2 in a previously reported study performed using mid-IR spectroscopy[23]

Previously reported studies have assessed fetal lung maturity by using thin layer chromatography (TLC)[24] to measure the lecithin/sphingomyelin (L/S) ratio and for the presence of PG, while others have measured the L/S ratio of phospholipids extracted from bronchoalveolar lavage using mass spectrometry (MS)[20] . However, none of these are suitable for use as point of care (PoC) tests as they require time and skill to carry out and interpret and there are no current devices available to assist with diagnosis. Measurement of dried gastric aspirate using attenuated total reflectance coupled with Fourier transform infrared spectrometry (ATR-FTIR)[23] has been reported and shows promise to be realized in a PoC device. However, although the use of dried samples ensures the highest concentration of biomarkers for spectral analysis, it can cause challenges for measurement. For example, variations in thickness within and between samples [25, 26], and crack formations occur as an artefact from the drying process [27] and both of these phenomena will result in spectra which have additional sources of error that can increase the uncertainty of component concentration predictions. Our approach [28]

makes use of bulk liquid sampling to overcome these spectral issues and presents a repeatable and reproducible method by which such measurements can be made.

Our previous work[28] applied machine learning to binary liquid component mixtures of S and DPPC to establish that these lipid biomarkers are amenable to being quantified using ATR-FTIR spectra and that it was possible to establish the L/S ratio based on the spectra alone. However, lung surfactant is constituted of many more components than DPPC and S, and the impact of these additional factors is likely to complicate the analysis and impair the ability to estimate the L/S ratio using the spectra. In this work we investigate the impact of variations in concentrations of five prominent lung surfactant lipids on the ability to estimate the L/S ratio from ATR FTIR spectra, by training machine learning models to predict the lipid concentrations. This study, to our knowledge, is the first to systematically assess the effect of varying DPPC, POPC, S, PG and Chol within physiologically relevant concentration ranges and demonstrate the use of partial least squares regression models to predict the concentration of nRDS biomarkers and the L/S ratio using liquid samples. Our findings show promise for predictive models of total PC, SM and PG, all useful nRDS biomarkers, to be simultaneously used to give a lung maturity index score and provide a clinician an evidential basis for deciding treatment. These results further indicates that ATR-FTIR on liquid samples may be translated into a clinical point of care device for the diagnosis of nRDS.

## 2. Materials and Methods

### *2.1. Machine Learning Procedure*

For measurements which generate spectra, many biomarkers may contribute to the absorbance at a particular wavenumber. It is, however, difficult to summarize the contribution of each biomarker to the absorbance at a particular spectral region. In order to identify and then quantify a biomarker of interest, multivariate techniques must be resorted to reduce the complexity of the data, and allow effective modelling of the spectral information.

The design of experiments (DoE) methodology [29] is a systematic approach to experimentation for characterizing and quantifying how a set of process parameters affects some response parameter. The experiments are designed to efficiently capture information about how the response parameter

is affected by input parameter interactions/higher order effects. Identification of input parameters is usually based on prior domain knowledge and are shortlisted by a using a screening experimental design, however this requires a level of domain knowledge to identify which ones to observe. This approach provides a useful framework to generate appropriate mixtures and map out the experimental domain, and establish whether the presence of varying concentrations of surfactant lipids impinges on the ability to develop machine learning models and ultimately predict the L/S ratio. Machine learning can be used to model the underlying patterns in the collected data and simplify the data analysis without requiring an *a priori* understanding of the specific relevant parameters[30].

FTIR spectra are sensitive to the molecular environment of an analyte due to intermolecular interactions, so testing spectral variation by varying a single lipid at a time would not permit modelling of the interactions that occur between them. A general full-factorial [31] approach testing 5 components at 5 different concentration levels requires 3125 samples for a single run. While this would comprehensively answer how prediction models are affected by the variation in each component, the number of runs required is prohibitive in time and cost and is wasteful in terms of exploring an unrealistic physiological model. An alternative approach is to pursue a mixture design with constraints on each component to limit them closer to reported physiological ranges. This has the advantage of reducing the number of tests required by focusing on realistic values for each of the components, and is known as an extreme vertices design.

For this study, five lipids (DPPC, S, POPC, PG and Chol) were chosen to generate the mixtures. A constraint applied on the mixtures was that the component parts, when combined should give a total proportion of 100% to avoid exploration of mixtures that stray from the physiologically relevant concentration and the proportions were directly relatable to reported values in the literature. The dataset necessarily contained imbalances with respect to the number of levels tested with each lipid, but as the primary aim of this experiment was to establish sample L/S ratios and to explore the extent to which the presence of other lipids in a physiological range can interfere with its determination, this remained unmodified and there were no treatments to the dataset to address this imbalance.

Partial least squares regression (PLSR) was chosen to build prediction models for the lipids test because it can be used to reduce the dimensionality of the dataset and handle multicollinear data. This builds from our

previous work following the same pre-processing methods and where it was established that PLSR models of the second derivative dataset provided the best performance for predicting DPPC and S concentrations in similar, albeit less complex mixtures[28].

The data pathway used is shown in fig. 1 showing the methodology employed to calibrate the PLSR models and shows how the data is split and processed after collection. The PLSR models were trained on the training data using a cross validation approach to establish the model parameters. The optimal parameters for each lipid PLSR model are used and trained using the whole training dataset giving a final set of models and uncertainty calculations. K-fold[32] cross-validation (CV) was used to split the training data such that for each evaluation the CV test dataset contained around 1% of the total training set data. To establish the optimal parameters for the model (number of latent variables (LVs), models with increasing numbers of LVs were generated and trained using each of the 100 CV train datasets and then assessed against the CV test datasets. The maximum average $R^2$/minimum average mean squared error (MSE) for the LVs was then used to select the optimal parameters for the model. The final model was retrained on the full training dataset.

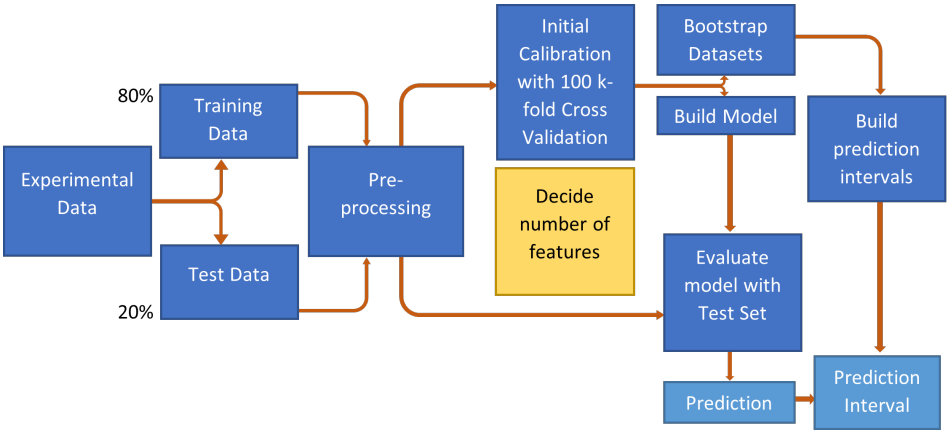



Figure 1: Machine learning data flow pathway used to generate the PLSR models

Reporting uncertainties in estimates provided by a point of care measurement [33] is useful in giving clinicians additional information about how much weight to give the predicted lipid concentration when reaching a diagnosis. If the prediction interval encompasses the cut-off region of L/S 2.2,

then it should necessarily be treated with more caution than one in which the prediction intervals were far from the cut-off region. This diagnostic 'gray' zone[34] is the region between where the test results are not definitive, and are distinguished from those test results that indicate a clear positive or negative result.

The uncertainties related to the model predictions in this study were obtained using two similar approaches, a "jackknife+-after-bootstrap"[35] method implemented in the MAPIE library and a variant bootstrap method[36] implemented in the Doubt library. The jackknife+-after-bootstrap incorporates theoretical guarantees on the prediction intervals such that, in the worst case, they will always provide a coverage rate of 1-2$\alpha$ (where $\alpha$ is the required level of uncertainty) while placing the mild requirement of "exchangeability" on the modelled data. The variant bootstrap also incorporates theoretical guarantees on the prediction intervals of a coverage rate of 1-$\alpha$, and has been shown to provide better coverage in circumstances where the model outputs have a high variance, while requiring input data fulfilling the slightly more stringent independent and identically distributed condition. Both methods are expected to provide coverage close to 1-$\alpha$ under most conditions and are suitable for analyzing models without the constraint of requiring a normally distributed output data; something which is not guaranteed when generating predictive regression models.

### *2.2. Experimental Lung Surfactant Lipid Model*

This study required the use of a generated lung surfactant lipid model so that the concentrations of each lipid in the mixture could be controlled and its effect on the spectra evaluated. Using patient derived samples would require a large pool of individuals, with no guarantee that the whole physiological range would be covered and was therefore deemed inappropriate in this case. Using a generated model permits the application of a designed approach to explore the physiological concentrations of each lipid and map out the physiological space. The lipids in this study were dissolved in dichloromethane in the state that they would be in post liquid-liquid extraction which separates the phospholipid fraction from other components present in the surfactant sample. The surfactant sample, when taken from a neonate, is in an aqueous form. While it would be ideal from the the perspective of a point of care device to measure directly in this form, it does invite complexity, such as the presence of proteins and strong absorption by water in the mid infrared region, which can make measuring small concentrations of lipids more

difficult.

The lung surfactant lipid model in this study (given in table 1, below) was generated by including the more prominent phospholipids (DPPC and POPC), neutral lipids (cholesterol) and those lipids considered to be biomarkers for nRDS (S and PG). The concentration of DPPC and POPC together formed 53% to 69% of the total lipid fraction (58% to 76% of the total phospholipid fraction) which encompasses the range of concentrations reported in healthy control and nRDS conditions[11, 13] . The concentrations for each lipid were chosen by combining the reported values for each tested lipid in both healthy and nRDS states, and checking to see that the regions above and below L/S ratio 2.2 were covered so that the impact of varying each of the lipids could be established for measuring the L/S ratio, by ATR-FTIR, on liquid samples. The levels studied denote the unique concentrations for each lipid within the sample mixtures due to the constraints enforced by the extreme vertices [31] approach and the lipid concentration ranges tested were in the order of those that have been previously reported as recovered from pharyngeal aspirates of neonates[12].

| **%** | **DPPC** | **POPC** | **S** | **PG** | **Chol** |
|---|---|---|---|---|---|
| **Max** | 48 | 24 | 30 | 12.4 | 6 |
| **Min** | 38 | 19.3 | 9.6 | 0.3 | 4 |
| **Levels** | 44 | 5 | 84 | 41 | 5 |
| **Concentration Range (mM)** | 1.035 - 1.308 | 0.508 - 0.632 | 0.273 - 0.853 | 0.008 - 0.333 | 0.207 - 0.310 |

Table 1: Lipid composition limits used in Minitab® for the Extreme Vertices experimental design and concentration ranges for the lipids tested.

*2.3. Sample Generation*

Measured masses of purified synthetic DPPC (1,2-dipalmitoyl-sn-glycero-3-phosphocholine), POPC (1-palmitoyl-2-oleoyl-glycero-3-phosphocholine), PG (1,2-dipalmitoyl-sn-glycero-3-phospho-(1'-rac-glycerol)) (Avanti® Polar Lipids) and SM (N-Palmitoyl-D-sphingomyelin) and cholesterol (Merck) were dissolved in methanol (Arcos) and dichloromethane (Arcos) and used to prepare homogeneous stock solutions, with vortex mixing as required. Methanol was used to increase solubility of the lipids at room temperature so that each mixture was homogeneous, as PG in particular does not easily dissolve in

dichloromethane. Samples were generated by pipetting required amounts of each stock lipid solution into a vial and then dried on a hot plate set to 63 °C under nitrogen. The proportions of each lipid for each sample were constrained to sum to 100 when generating the samples, which was done using the Extreme Vertices design implemented in Minitab®, so as to cover a physiologically relevant experimental space, while still exploring the impact of different L/S ratios on the spectra (further detail in Supplementary Material). The sample vials were capped with a PTFE lined screw cap until measurement. Immediately prior to measurement samples were redissolved in 3 ml of dichloromethane, vortex mixed for 30 s and observed to be visibly homogenous and fully dissolved. If required, heat was applied by placing the closed vial on a hotplate for 30 s and further vortex mixing applied until visibly homogeneous and fully dissolved. The complete list of lipid concentrations can be found in the supplementary materials table S2.

### *2.4. FTIR Spectra Collection and Preprocessing*

Prior to generating optimised prediction models spectra were collected and preprocessed to remove sources of noise that were attributed to parameters other than changes in the target analyte. Data used to assess the model performance should not influence nor inform the model optimisation process, so the data should be appropriately split to prevent such data leakage from occurring.

Measurements were performed on an Agilent Cary 670 FTIR instrument equipped with a potassium bromide (KBr) beam splitter and a deuterated triglycine sulphate (DTGS) detector. The resolution was set to 4$cm^{-1}$ with 32 co-added scans. Nitrogen purging was set to 8 liters per minute and the sample presented to the spectrometer by means of a 10-bounce Pike® zinc selenide (ZnSe) horizontal ATR (HATR) accessory with a solvent lid. 0.5 ml of sample was used for each measurement and tested in a random order. Between each sample the ATR crystal was cleaned and a new background scan performed as per our previously published[28] protocol. Nine spectra were collected consecutively for each sample.

The spectra collected were initially preprocessed using Peak® Spectroscopy (Operant LLC) software where the spectra were baseline offset corrected, the spectra were truncated to 3500 $cm^{-1}$ to 850 $cm^{-1}$ and regions where dichloromethane and carbon dioxide exhibit strong absorbances were removed from the spectra. This region was chosen because the strong absorbance below 850 $cm^{-1}$ of dichloromethane results in a poor signal-to-

noise ratio, while spectra above 3500 $cm^{-1}$ contain little additional information that would be useful for a PLSR model (OH stretch modes exist above 3500 $cm^{-1}$, but this is . The data were converted into csv files and further processed in Python using Pandas and Scikit-Learn[37] to calibrate PLSR models to predict the concentration of each lipid component in the sample. The spectra were grouped by sample and split into a training and test dataset. Additional preprocessing was separately performed on these sets in Python by applying a Savitzky-Golay derivative filter (15-point window, 3rd order polynomial, second derivative) to obtain smoothed second derivative spectra. The test set spectra were not considered for any model building and only used to assess the model performance after optimal parameters obtained through cross-validation had already been obtained. Once the PLSR models had been generated, they were evaluated using the test set spectra.

## 3. Results and Discussion

### *3.1. Model Optimization*

Figure 2 shows the normalised ATR-FTIR spectra of each of the individual lipids (originally at 1 mM concentration) that comprised the model lung surfactant. PG and both lecithins (DPPC and POPC) share peaks in the 1730 $cm^{-1}$ region which is associated with the ester carbonyl bond present on these molecules. Table 2 assigns the more prominent peaks and shows that DPPC and POPC share many spectral features but are distinguishable from other lipids present. Chol shares the 2800 - 3000 $cm^{-1}$ spectral region with other lipids but its spectrum is distinct which is expected to be a useful property to develop effective predictive PLSR models (in proceeding section).

In total 1062 spectra were collected from 118 different samples (9 spectra for each sample) and randomly split by sample into a training (846 spectra) and evaluation test set (216 spectra). The training set spectra were further split by sample into 100 different CV training sets (containing 837 spectra) and CV test sets (containing 9 spectra) corresponding to a 'leave one out' k-fold approach with respect to the sample number. The CV test sets were used to evaluate the initial models calibrated using the CV training sets to discover the optimal number of LVs required for the final PLSR models.

The average $R^2$ values for the 100 cross-validation models for LVs from 1 to 39 are shown in fig 3. For each of the PLSR lipid models, the LV corresponding to the maximum average $R^2$ value for the CV test set was chosen as the optimal number of LVs for the prediction model. The highest

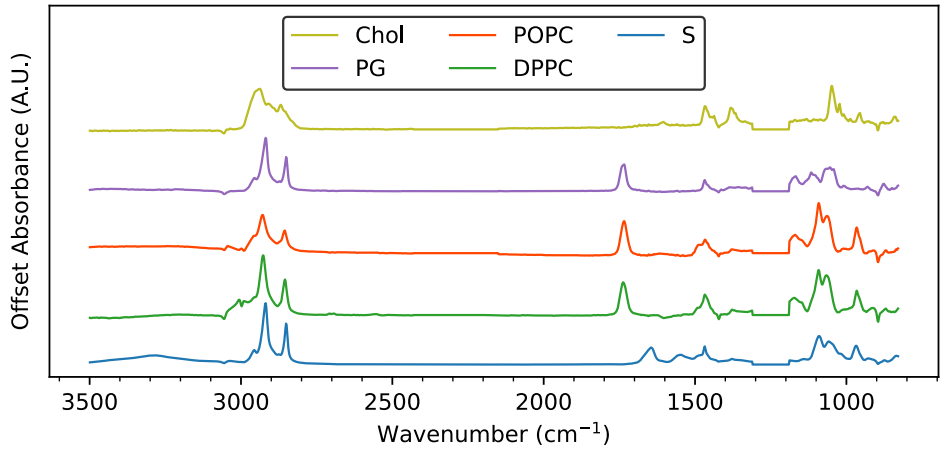



Figure 2: Normalised ATR-FTIR Spectra of the lipids constituting the lung surfactant model in this study at 1 mM concentration, with regions, where $CO_2$ and dichloromethane absorbs strongly, removed.

peak $R^2$ values for the CV test data were observed from PG (fig 3 C, $R^2 = 0.934$, MSE $= 0.0007$ mM$^2$) and S (fig 3 A, $R^2 = 0.928$, MSE $= 0.0013$ mM$^2$) indicating that a high degree of variance in the spectra was explained by their respective PLSR models. The DPPC model performed less well ($R^2 = 0.758$), and those for POPC and Chol performed poorly, with lower CV test $R^2$ values (0.612 and 0.448 respectively) indicating that much less of the variance in the data was captured by the model. The reason for such may be due to the fewer levels, smaller concentrations and smaller sample interval ranges tested for POPC and Chol (between 0.10 to 0.27 mM concentrations), as opposed to the levels and ranges in the better performing models (between 0.33 to 0.58 mM concentrations). The DPPC model is thought to be affected by the molecular similarity between DPPC and POPC which share many features in their respective mid-IR spectra so that differentiating between them is more difficult. The addition of other lipids into the mixtures, as expected, makes it more difficult to distinguish between similarly structured molecules.

Due to the poorer performance of the POPC and DPPC models individually, an additional model for PC was developed, which was trained on the total PC concentration in the sample (sum of DPPC and POPC) and used to generate a predictive model. The CV test performance for this model (fig 3, $R^2 = 0.876$) was better than either of the two lipids individually. The larger

| **Wavenumber ($cm^{-1}$)** | **Bond Vibration** |
|---|---|
| 2853-2962 | C-H stretch ($CH_2$ and $CH_3$) |
| 1720-1745 | C=O stretch of ester carbonyl |
| 1650-1640 | Amide I (C=O stretch) |
| 1445-1480 | $CH_2$ bend |
| 1160-1190 | C-O stretch |
| 1100-1200 | C-C stretch |
| 1085-1110 | P-$O_2^-$ stretch |
| 1060 | C-O-P stretch |
| 970 | C-N stretch (choline) |

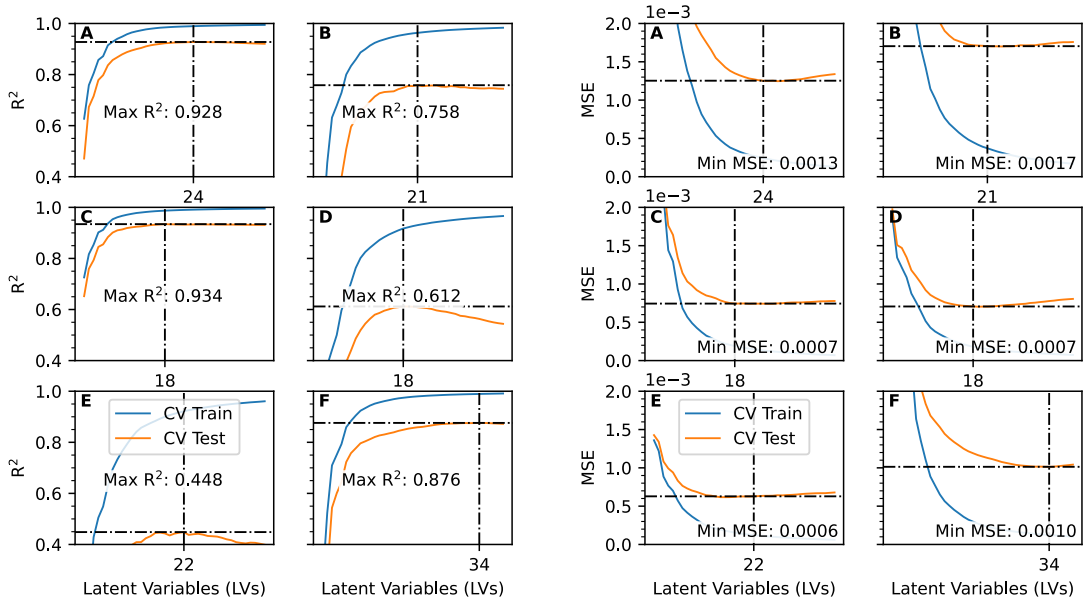
Table 2: Selected wavenumbers associated with bond vibrations present within the lipids tested [38, 39]

sample PC concentration range (0.396 mM, ranging from 1.543 to 1.939 mM) and additional levels was likely to have contributed a larger variance to the data, providing a better basis upon which to calibrate a predictive model, and match the variance of other better performing models. In a physiological setting this approach may be equally valid, as the largest PC component is made up of DPPC, so long as PCs as a group can be differentiated from other lipids present.

The performance of the PG model shows additional promise for the use of IR spectroscopy to diagnose nRDS as PG increases in concentration in surfactant towards the end of neonatal gestation and has been considered a late-stage lung maturity biomarker[40]. By quantifying multiple diagnostic biomarkers, it is potentially possible to develop a lung maturity index which considers other inputs, including (for example) gestational age and the quantity of phosphatidylinositol (PI – which decreases in concentration towards the end of neonatal gestation), to give additional information to a clinician upon which to base a diagnosis.

### *3.2. Test Set Lipid Concentration Determination*

The final PLSR models for each lipid were generated by training on the full training set using the optimal number of LVs established from the cross validation. The 'test set' was used to evaluate the performance of these models for predicting concentrations of each of the lipids (shown in fig 4), and the




(a) Average cross-validation $R^2$

(b) Average cross-validation MSE

Figure 3: Graphs showing the maximum CV $R^2$ (left) and minimum CV mean squared error (MSE) (right) obtained for each PLSR model (**A**: S, **B**: DPPC, **C**: PG, **D**: POPC, **E**: Chol and **F**: PC) for both the CV train datasets (blue line) and the CV test datasets (orange line). The maximum $R^2$/minimum MSE and corresponding LV for the CV test dataset is indicated for each lipid model and was used as the number of LVs for each of the optimized PLSR lipid models.

prediction intervals generated using both the jackknife+-after-bootstrap and the variant bootstrap methods. The prediction performance for each model against the 'test set' showed similar trends to the CV test performances in terms of the $R^2$ values obtained. The models for S, PG and PC ($R^2$ = 0.817, 0.846 and 0.846 respectively) were able to determine lipid concentrations within the physiological ranges tested. The model for DPPC ($R^2$ = 0.592) performed less well than these models but still better than those for Chol ($R^2$ =-0.082) and POPC ($R^2$ =-0.334). While the model performance for Chol may be explained by the fewer levels tested, as well as the limited concentration range, this does not explain why the DPPC and POPC performance are both poor. It is more likely that this is a result of the molecular similarities between the two lipids which results in similar changes in the spectra when changing the concentration in either one. The prediction intervals for the

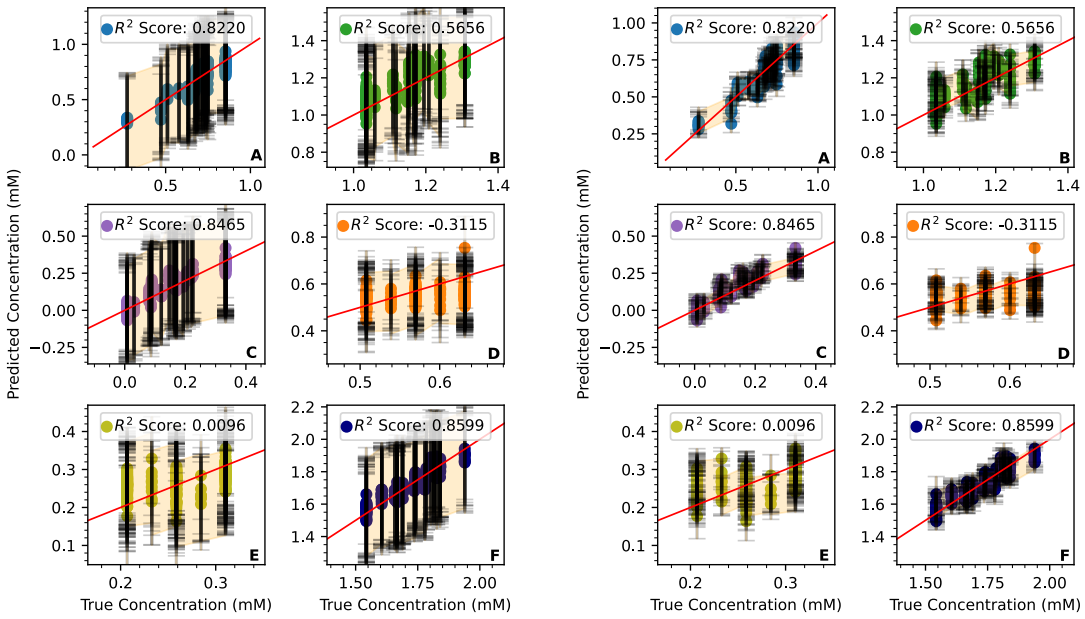
jackknife+-after-bootstrap and variant bootstrap give the uncertainty in the model predictions, and are indicated in figure 3 by the error bars. While the prediction intervals generated using the jackknife+-after-bootstrap method cover a narrower range, their utility is suboptimal due to their optimistic nature (fig 4, summary in supplementary information table **??**) for the 1-$\alpha$ region. The coverage probability for these intervals (see table 3) was observed to be between 0.602 (POPC) and 0.935 (PC), all less than the expected 0.95 for an $\alpha$ value of 0.05. The variant bootstrap method is more conservative, with wider prediction intervals, and in all but one case the coverage probability for the predictions with their respective intervals fully encloses the true lipid concentration for the sample. The coverage probability for the theoretical guaranteed region for the jackknife+-after-bootstrap, encompassing the 1-2$\alpha$ region, was valid for all models apart from the POPC model, for which the coverage probability was 0.945. While this is close to the required coverage probability it does suggest an ill-fit POPC model, which may also relate to the initial assumptions on the data of exchangeability in the context of highly similar lipids.

| Lipid | Jackknife+-after-bootstrap 95% Prediction Interval Coverage Probability | | Variant Bootstrap 95% Prediction Interval Coverage Probability |
|---|---|---|---|
| | 1-$\alpha$ | 1-2$\alpha$ | 1-$\alpha$ |
| S | 0.787 | 0.991 | 1.000 |
| DPPC | 0.806 | 1.000 | 1.000 |
| POPC | 0.602 | 0.944 | 0.995 |
| PG | 0.815 | 0.995 | 1.000 |
| Chol | 0.773 | 0.991 | 1.000 |
| PC | 0.935 | 0.995 | 1.000 |

Table 3: Prediction interval coverage probabilities for the jackknife+-after-bootstrap and variant bootstrap methods for each of the lipid models tested.

### *3.3. Test Set L/S Ratios*

For the final evaluation of the L/S ratio, the mean of the 9 spectra from each sample were used for generating averaged spectra (giving 24 different spectra) which were used to make predictions for PC and S. The prediction



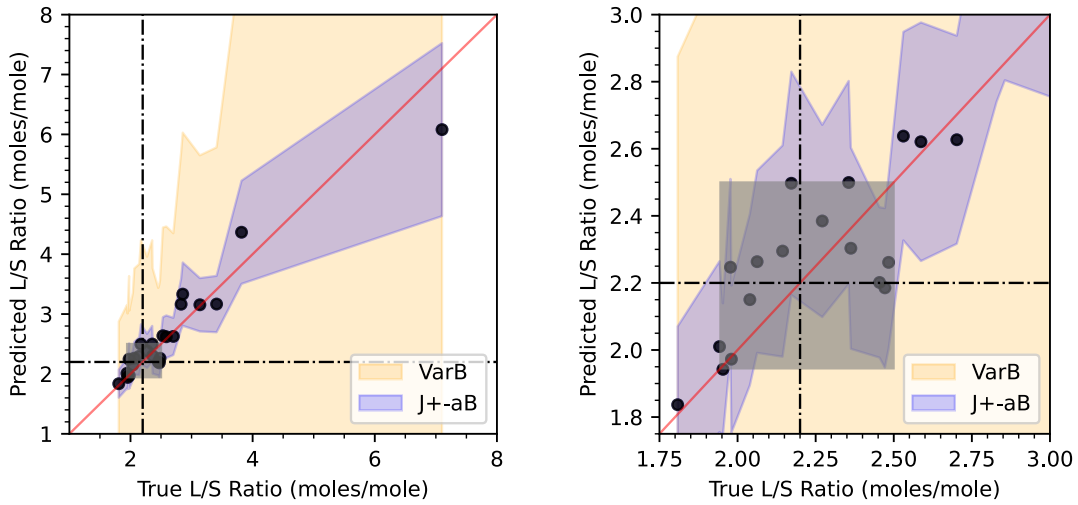

(a) Variant bootstrap prediction intervals

(b) Jackknife+ after bootstrap prediction intervals

Figure 4: Lipid Predictions (**A**: S, **B**: DPPC, **C**: PG, **D**: POPC, **E**: Chol and **F**: PC) from each of the models on the Test Set along with the $R^2$ values for each lipid with the 95% prediction intervals generated by the variant bootstrap (left) and the jackknife+-after-bootstrap (right) methods. The red line in each of the graphs corresponds to the expected prediction value while the shaded region corresponds to the prediction with the largest prediction interval range for that concentration.

intervals for these L/S ratios were determined by combing the uncertainties obtained from the prediction of S and PC (see supplementary data equation **??**). The largest error between the upper and lower prediction interval was used in the error propagation calculations and used as a symmetric prediction interval around the prediction. Fig 5 shows the final L/S ratio predictions for the averaged spectra of the test set. The prediction intervals given by the variant bootstrap method were extremely wide, and while they cover the true L/S ratio, they would not be useful in a clinical setting due to their width. The prediction interval provided by the jackknife+-after-bootstrap method was much smaller and in all but three cases included the true L/S ratio. In those cases (right, fig 5) the error in the the prediction interval was within L/S 0.05 moles/mole of the nearest prediction interval limit. The

largest errors for all three were due to larger errors in the model predicting S concentrations, which was the component at the lower concentration in this calculation. Predictions which included the L/S 2.2 moles/mole region within the prediction uncertainty were deemed to be within the diagnostic gray area and these were found to fall within a region of L/S ±+0.30 moles/mole for the averaged test set predictions. On this basis, a sample with an L/S ratio of 2.2 would in 95% of cases provide a reading of less than 2.5 moles/mole.




(a) Full Range

(b) Zoomed in between L/S 1.75 to 3.00 moles/mole

Figure 5: The predicted L/S ratios based on the PC and S models, with prediction intervals generated by combing and propagating the largest errors in the models. The full test set region is shown on the left, and zoomed in on the L/S 2.2 region is shown on the right.

### *3.4. Comparison with Reported Clinical Data*

While use in a clinical scenario requires the development of different prediction models trained on surfactant samples obtained from patients, it is useful to understand what a 'gray' region of ±0.3 moles/mole either side of the diagnostic L/S ratio of 2.2 (which itself is still the subject of discussion) would mean in a clinical scenario. Data on the L/S ratio of neonates from Verder et al.[23] were reanalyzed in the context of this gray region to see how many neonates might benefit from a point of care diagnostic device with the performance of the current lipid models (see fig 6). The number of patients

in the gray zone represent approximately 10% of the total 136 patients that were part of the study. Approximately 35% of the cohort would have been correctly diagnosed as requiring surfactant replacement therapy on the basis of the L/S ratio, and a further 33% would have been correctly diagnosed as not requiring surfactant replacement. The remaining 22% were incorrectly diagnosed using the L/S ratio as a biomarker for nRDS, and would likely have been subject to further investigation based on the presenting symptoms. In the absence of any current method to diagnose nRDS at the point of care, a device with a similar performance to the lipid models in this investigation would appear to be a beneficial pursuit, and would likely improve prognoses for a majority of preterm neonates with relevant symptoms.

One strategy to deploy the models similar to those developed in this study would be to take measurements from many patient samples and build a calibration library. However, this may be impacted by the non-availability of suitable concentrations to develop a model that works across the full physiological range. Using a hybrid approach, which takes the data collected from this study and combines it with data from patient derived samples, is likely to be an efficient method to build a robust model for clinical use and has been planned as future work.

### *3.5. Opportunities for Miniaturization: Model Feature Importance's*

This section considers how a point of care device might be realized in a small form factor that is easy to place in a clinical setting by identifying opportunities to reduce device complexity. While a commercial ATR-FTIR system would fulfil the paradigm of a PoC device, further miniaturisation opportunities exist for devices based on quantum cascade lasers (QCLs). FTIR spectrometers require the use of an interferometer to generate an interferogram which is Fourier transformed into a spectrum but QCL based spectrometers can directly scan the spectrum within a specified bandwidth. Thus, by dispensing with the requirement for an interferometer, smaller devices may be realized. It is then also possible to reduce the time required to produce a spectrum with a QCL system by concentrating on those regions which the PLSR models have previously identified as providing the most information for the determination of the analyte concentration. By identifying and focussing on specific wavenumbers of interest it is possible to reduce increase the signal-to-noise ratio by increasing the number of co-added scans within a given time period.

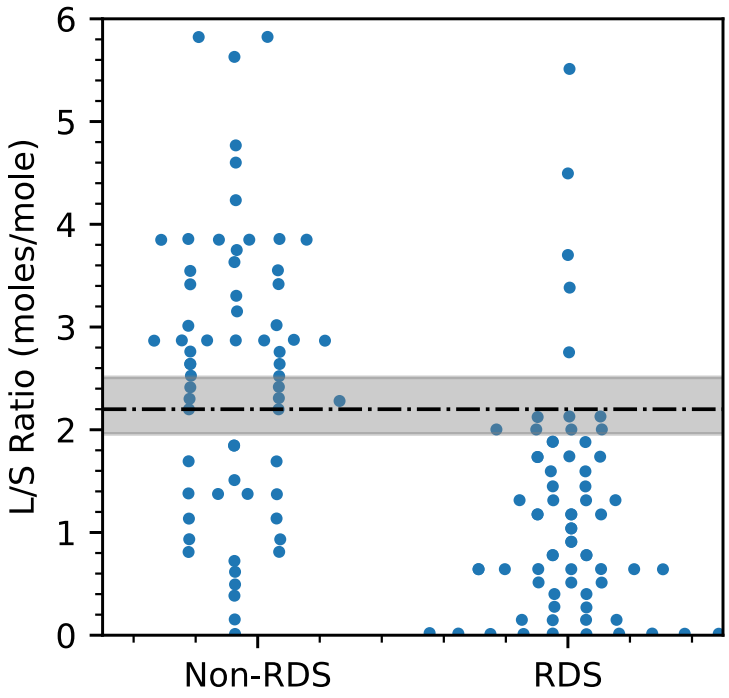



Figure 6: Effect of gray diagnostic region on previously reported dataset from a study seeking to diagnose nRDS using an L/S ratio of 2.2. Datapoints beyond the viewing frame are not displayed, but were considered as part of the analysis. Reproduced, with permission, from Verder et al[23].

SHAP (SHapley Additive exPlanations) values[41] can be used to explain the strength of the wavenumber features used by a model when providing a prediction in a model-agnostic manner, and therefore inform decisions on which regions of the spectrum to focus on. One issue with comprehensively generating SHAP values, especially for large datasets and complex models is that there is a large computational overhead associated with it. As a result, in this study a subset of 100 spectra from the test set was used to generate the SHAP values to balance between sufficient data for the provision of some insight while at the same time presenting a manageable computational overhead.

The wavenumber features which contributed the strongest features, as represented by the features comprising the five largest mean SHAP values for each model are shown in fig 7. Wavenumbers 1041 and 1043 $cm^{-1}$ ap-

pear as strong features in the models that predicts S, DPPC and PC, which corresponds to the wavenumbers associated with the side of a peak centered at 1061 $cm^{-1}$ in the original spectra. It is also interesting to note that the wavenumber associated with the ester carbonyl bond on PCs (on the side of a peak near 1736 $cm^{-1}$) was a strong feature in the S model while at the same time a wavenumber associated with the amide I vibration in S (on the side of a peak near 1650 $cm^{-1}$) was a strong feature in the PC model.

The mean SHAP values show that wavenumber features in the region 1600 – 1800 $cm^{-1}$ are in the top 5 features that contribute to the models predicted output and this is also a region where the impact of water vapor can degrade spectra in both the training and the test sets. While it is expected that the PLSR algorithm will generate models that maximise the covariance between the wavenumber features and the applied lipid concentrations, this will necessarily be affected by differences in water vapor, particularly if it affects a region deemed important for the model. A method to improve the spectra collected, which can be used in a clinical setting, will permit better models to be generated which is expected to positively influence their ability to predict the correct concentrations. We are developing a method to overcome the issue of vapour interference in spectra so that high quality spectra can be provided both during training as well as to correct spectra obtained in a point of care setting.

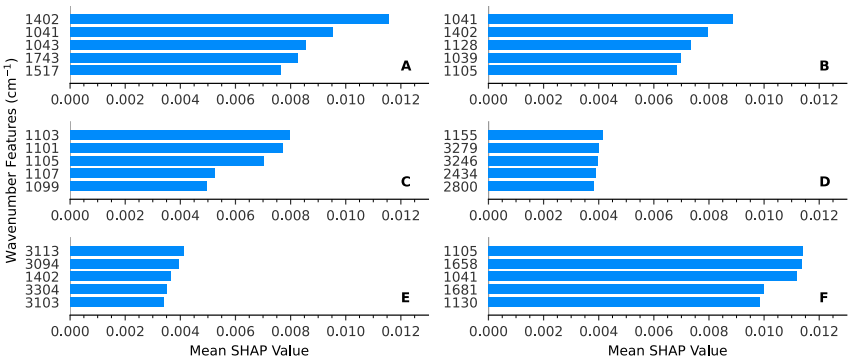



Figure 7: Top five SHAP Training outputs for each PLSR model (**A**: S, **B**: DPPC, **C**: PG, **D**: POPC, **E**: Chol and **F**: PC). The features have been rounded to the nearest wavenumber.

## 4. Summary and Future Study

In summary the present investigation used a physiologically informed model of lipids in lung surfactant to investigate the effect of varying lipid composition on the FTIR spectra when used to generate models to predict the concentration of each lipid in a mixture. Based on an extreme vertices design, 118 liquid samples were systematically generated with varying compositions. Each sample was measured using ATR-FTIR to collect 9 spectra which were used to generate and further test predictive models of lipid concentrations. The dataset was split into a training and test dataset, and the training dataset used to define the preprocessing and optimal model parameters required. The spectra were preprocessed by removing irrelevant data (peaks due to carbon dioxide and dichloromethane), baseline offset correction and by taking the second derivative spectra for modelling. Models generated for predicting S, PC and PG performed the best while those for DPPC, POPC and Chol performed poorly in predicting the test set. Prediction intervals for each model were generated using the variant bootstrap and jackknife+-after-bootstrap methods. The coverage probability of the variant bootstrap is too conservative, while the jackknife+-after-bootstrap is slightly too aggressive. A future study interest will be to understand the best way to obtained well balanced prediction intervals. The SHAP values for each model show that the five most informative features in the models includes regions on the side of peaks that vary most with concentration, indicating possible regions of interest for a small, compact spectrometer to concentrate design effort on. The L/S ratio was generated using the best performing 'L' model (PC) and the S model and propagating the errors obtained from the predictions for each individual component. The overly conservative variant bootstrap prediction intervals provide no meaningful output as they provide an interval that is too wide, while the jackknife+-after-bootstrap provides more balanced intervals, giving a prediction interval of +/- 0.3 moles/mole in the region of L/S ratio of 2.2 which is considered to be the diagnostic cut-off for nRDS.

## Acknowledgement

The authors acknowledge the funding support received from the UK EPSRC grant EP/S03109X/1

The authors thank Jody Clarke from Gilson® for her assistance procuring pipette tips at the height of the COVID19 pandemic.

## Data Availability Statement

All data supporting this study are available from the University of Southampton repository at: https://doi.org/10.5258/SOTON/D2532.

## Competing Interests Statement

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

## CRediT Authorship Contribution Statement

**W. Ahmed:** Data acquisition, Investigation, Methodology, Writing – original draft. **A. V. Veluthandath:** Methodology, Writing – original draft, Writing – review & editing, **A. Dushianandan:** Writing – review & editing, **J. Madsen:** Writing – review & editing, **H. W. Clark:** Writing – review & editing, **A. D. Postle:** Conceptualization and Writing – review & editing, **J. S. Wilkinson:** Conceptualization, Supervision, Writing – review & editing, **G. S. Murugan:** Conceptualization, Funding acquisition, Supervision, Resources, Project administration and Writing – review & editing

## References


[1] S. De Bruyne, M. M. Speeckaert, J. R. Delanghe, Applications of mid-infrared spectroscopy in the clinical laboratory setting, Critical Reviews in Clinical Laboratory Sciences 55 (1) (2018) 1–20. doi:10.1080/10408363.2017.1414142.

[2] D. Finlayson, C. Rinaldi, M. J. Baker, Is Infrared Spectroscopy Ready for the Clinic?, Analytical Chemistry 91 (19) (2019) 12117–12128. doi:10.1021/acs.analchem.9b02280.

[3] M. E. Avery, J. Mead, Surface Properties in Relation to Atelectasis and Hyaline Membrane Disease, Archives of Pediatrics & Adolescent Medicine 97 (5 (Part I)) (1959) 517. doi:doi:10.1001/archpedi.1959.02070010519001.

[4] N. Seger, R. Soll, Animal derived surfactant extract for treatment of respiratory distress syndrome, Cochrane Database of Systematic Reviews (Apr. 2009). doi:10.1002/14651858.CD007836.

[5] F. L. Bahadue, R. Soll, Early versus delayed selective surfactant treatment for neonatal respiratory distress syndrome, Cochrane Database of Systematic Reviews (Nov. 2012). doi:10.1002/14651858.CD001456.pub2.

[6] R. Soll, C. J. Morley, Prophylactic versus selective use of surfactant in preventing morbidity and mortality in preterm infants, in: The Cochrane Collaboration (Ed.), Cochrane Database of Systematic Reviews, John Wiley & Sons, Ltd, Chichester, UK, 2001, p. CD000510. doi:10.1002/14651858.CD000510.

[7] M. Griese, Pulmonary surfactant in health and human lung diseases: State of the art, European Respiratory Journal 13 (6) (1999) 1455–1476. doi:10.1183/09031936.99.13614779.

[8] U. Pison, R. Herold, S. Schürch, The pulmonary surfactant system: Biological functions, components, physicochemical properties and alterations during lung disease, Colloids and Surfaces A: Physicochemical and Engineering Aspects 114 (1996) 165–184. doi:10.1016/0927-7757(96)03572-8.

[9] M. Hallman, T. A. Merritt, M. Pohjavuori, L. Gluck, Effect of Surfactant Substitution on Lung Effluent Phospholipids in Respiratory Distress Syndrome: Evaluation of Surfactant Phospholipid Turnover, Pool Size, and the Relationship to Severity of Respiratory Failure, Pediatric Research 20 (12) (1986) 1228–1235. doi:10.1203/00006450-198612000-00008.

[10] M. Hallman, R. Spragg, J. H. Harrell, K. M. Moser, L. Gluck, Evidence of lung surfactant abnormality in respiratory failure. Study of bronchoalveolar lavage phospholipids, surface activity, phospholipase activity, and plasma myoinositol., Journal of Clinical Investigation 70 (3) (1982) 673–683. doi:10.1172/JCI110662.

[11] M. Hallman, B. H. Feldman, E. Kirkpatrick, L. Gluck, Absence of Phosphatidylglycerol (PG) in Respiratory Distress Syndrome in the New-

born, Pediatric Research 11 (6) (1977) 714–720. doi:10.1203/00006450-197706000-00003.

[12] C. F. Poets, A. Arning, W. Bernhard, C. Acevedo, H. Von Der Hardt, Active Surfactant in Pharyngeal Aspirates of Term Neonates: Lipid Biochemistry and Surface Tension Function, European Journal of Clinical Investigation 27 (4) (1997) 293–298. doi:10.1046/j.1365-2362.1997.1050655.x.

[13] M. Hallman, M. Kulovich, E. Kirkpatrick, R. G. Sugarman, L. Gluck, Phosphatidylinositol and phosphatidylglycerol in amniotic fluid: Indices of lung maturity, American Journal of Obstetrics and Gynecology 125 (5) (1976) 613–617. doi:10.1016/0002-9378(76)90782-1.

[14] R. Veldhuizen, K. Nag, S. Orgeig, F. Possmayer, The role of lipids in pulmonary surfactant, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1408 (2-3) (1998) 90–108. doi:10.1016/S0925-4439(98)00061-1.

[15] A. D. Postle, E. L. Heeley, D. C. Wilton, A comparison of the molecular species compositions of mammalian lung surfactant phospholipids, Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 129 (1) (2001) 65–73. doi:10.1016/S1095-6433(01)00306-3.

[16] L. Gluck, M. V. Kulovich, R. C. Borer, P. H. Brenner, G. G. Anderson, W. N. Spellacy, Diagnosis of the respiratory distress syndrome by amniocentesis, American Journal of Obstetrics and Gynecology 109 (3) (1971) 440–445. doi:10.1016/0002-9378(71)90342-5.

[17] J. Goerke, Pulmonary surfactant: Functions and molecular composition, Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1408 (2-3) (1998) 79–89. doi:10.1016/S0925-4439(98)00060-X.

[18] I. Unal, Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach, Computational and Mathematical Methods in Medicine 2017 (2017) 1–14. doi:10.1155/2017/3762651.
URL https://www.hindawi.com/journals/cmmm/2017/3762651/

[19] F. Habibzadeh, P. Habibzadeh, M. Yadollahie, On determining the most appropriate test cut-off value: The case of tests with continuous results, Biochemia Medica 26 (3) (2016) 297–307. doi:10.11613/BM.2016.034.

[20] H.-S. Kwak, H.-J. Chung, Y. S. Choi, W.-K. Min, S. Y. Jung, Prediction of fetal lung maturity using the lecithin/sphingomyelin (L/S) ratio analysis with a simplified sample preparation, using a commercial microtip-column combined with mass spectrometric analysis, Journal of Chromatography B 993–994 (2015) 81–85. doi:10.1016/j.jchromb.2015.05.012.

[21] E. B. Olson, S. N. Graven, R. D. Zachman, Amniotic Fluid Lecithin to Sphingomyelin Ratio of 3.5 and Fetal Pulmonary Maturity, Pediatric Research 9 (2) (1975) 65–69. doi:10.1203/00006450-197502000-00002.

[22] L. Campanella, M. Sammartino, M. Tomassetti, G. Visco, Chemometric investigation of some analytical methods used for the chemical test of foetal lung maturity, Journal of Pharmaceutical and Biomedical Analysis 8 (8-12) (1990) 743–747. doi:10.1016/0731-7085(90)80115-6.

[23] H. Verder, C. Heiring, H. Clark, D. Sweet, T. E. Jessen, F. Ebbesen, L. J. Björklund, B. Andreasson, L. Bender, A. Bertelsen, M. Dahl, C. Eschen, J. Fenger-Grøn, S. F. Hoffmann, A. Höskuldsson, M. Bruusgaard-Mouritsen, F. Lundberg, A. D. Postle, P. Schousboe, P. Schmidt, H. Stanchev, L. Sørensen, Rapid test for lung maturity, based on spectroscopy of gastric aspirate, predicted respiratory distress syndrome with high sensitivity, Acta Paediatrica 106 (3) (2017) 430–437. doi:10.1111/apa.13683.

[24] L. Sharma, A. Desai, A. Sharma, A thin layer chromatography laboratory experiment of medical importance, Biochemistry and Molecular Biology Education 34 (1) (2006) 44–48. doi:10.1002/bmb.2006.49403401044.

[25] C. Hughes, M. Brown, G. Clemens, A. Henderson, G. Monjardez, N. W. Clarke, P. Gardner, Assessing the challenges of Fourier transform infrared spectroscopic analysis of blood serum: Assessing the challenges of FTIR spectroscopic analysis of blood serum, Journal of Biophotonics 7 (3-4) (2014) 180–188. doi:10.1002/jbio.201300167.

[26] E. Goormaghtigh, V. Raussens, J.-M. Ruysschaert, Attenuated total reflection infrared spectroscopy of proteins and lipids in biological membranes, Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes 1422 (2) (1999) 105–185. doi:10.1016/S0304-4157(99)00004-0.

[27] J. M. Cameron, H. J. Butler, D. S. Palmer, M. J. Baker, Biofluid spectroscopic disease diagnostics: A review on the processes and spectral impact of drying, Journal of Biophotonics 11 (4) (2018) e201700299. doi:10.1002/jbio.201700299.

[28] W. Ahmed, A. V. Veluthandath, D. J. Rowe, J. Madsen, H. W. Clark, A. D. Postle, J. S. Wilkinson, G. S. Murugan, Prediction of Neonatal Respiratory Distress Biomarker Concentration by Application of Machine Learning to Mid-Infrared Spectra, Sensors 22 (5) (2022) 1744. doi:10.3390/s22051744.

[29] S. A. Weissman, N. G. Anderson, Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications, Organic Process Research & Development 19 (11) (2015) 1605–1633. doi:10.1021/op500169m.

[30] S. Nikita, R. Sharma, J. Fahmi, A. S. Rathore, Process optimization using machine learning enhanced design of experiments (DOE): Ranibizumab refolding as a case study, Reaction Chemistry & Engineering 8 (3) (2023) 592–603. doi:10.1039/D2RE00440B.

[31] RG. Brereton, Chemometrics: Data Analysis for the Laboratory and Chemical Plant, Wiley, 2003.

[32] T. Hastie, R. Tibshirani, J. Friedman, Model Assessment and Selection, Springer New York, New York, NY, 2009, Ch. Model Assessment and Selection, pp. 219–259. doi:10.1007/978-0-387-84858-7_7.

[33] J. IntHout, J. P. A. Ioannidis, M. M. Rovers, J. J. Goeman, Plea for routinely presenting prediction intervals in meta-analysis, BMJ Open 6 (7) (2016) e010247. doi:10.1136/bmjopen-2015-010247.

[34] J. Coste, J. Pouchot, A grey zone for quantitative diagnostic and screening tests, International Journal of Epidemiology 32 (2) (2003) 304–313. doi:10.1093/ije/dyg054.

[35] B. Kim, C. Xu, R. F. Barber, Predictive Inference Is Free with the Jackknife+-after-Bootstrap (Nov. 2020). arXiv:2002.09025.

[36] C. Mougan, D. S. Nielsen, Monitoring Model Deterioration with Explainable Uncertainty Estimation via Non-parametric Bootstrap (Nov. 2022). arXiv:2201.11676.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[38] S. E. Park, H. Y. Yu, S. Ahn, Development and validation of a simple method to quantify contents of phospholipids in krill oil by fourier-transform infrared spectroscopy, Foods 11 (1) (2022). doi:10.3390/foods11010041.

[39] U. P. Fringeli, H. H. Günthard, Membrane Spectroscopy, Springer Berlin Heidelberg, Berlin, Heidelberg, 1981, Ch. Infrared Membrane Spectroscopy, pp. 270–332. doi:10.1007/978-3-642-81537-9_6.

[40] A. Bent, J. Gray, E. Luther, M. Oulton, L. Peddle, Assessment of fetal lung maturity: Relationship of gestational age and pregnancy complications to phosphatidylglycerol levels, American Journal of Obstetrics and Gynecology 142 (6) (1982) 664–669. doi:10.1016/S0002-9378(16)32438-3.

[41] S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, Advances in Neural Information Processing Systems 2017-Decem (Section 2) (2017) 4766–4775. arXiv:1705.07874.