Statistical Modelling xxxx; **xx(x)**: 1–20

A Bayesian hierarchical model for predicting rates of oxygen consumption in mechanically ventilated intensive care patients

Luke Hardcastle[1], Samuel Livingstone[1], Claire Black[2], Federico Ricciardi[3] and Gianluca Baio[1]

[1]Department of Statistical Science, University College London, UK

[2]Therapies and Rehabilitation, University College London Hospitals NHS Foundation Trust, London, UK

[3]Owlstone Medical, Cambridge, UK

Abstract: Patients who are mechanically ventilated in the Intensive Care Unit participate in exercise as a component of their rehabilitation to ameliorate the long-term impact of critical illness on their physical function. The effective implementation of these programmes is limited, however, as clinicians do not have access to a patient's $\dot{V}O_2$ values, a physiological measure that quantifies an individual patient's exercise intensity level in real-time. In this work we have developed a Bayesian hierarchical model with temporally correlated latent Gaussian processes to predict $\dot{V}O_2$ using readily available physiological data, providing clinicians with information to personalise rehabilitation sessions in real-time. The model was fitted using the Integrated Nested Laplace Approximation and validated using posterior predictive checks, and the impact of alternate specifications of the latent process was examined. Assessed using leave-one-patient-out cross-validation, we show that the ability to provide probabilistic statements describing classification uncertainty gives the model favourable predictive power compared to a state-of-the-art comparator based on the oxygen uptake efficiency slope, with a more than seven-fold increase in accuracy in identifying when a patient is at risk of over-exertion.

Key words: Clinical prediction tools, Gaussian processes, Integrated Nested Laplace Approximation, Critical illness; Exercise rehabilitation

Received 10 October 2023; **revised** 10 January 2024; **accepted** 6 February 2024

1 Introduction

Patients who are mechanically ventilated in the Intensive Care Unit (ICU) as a result of critical illness are often left with a range of impairments, due to the pathological effects of critical illness and its treatments on nerve, muscle cardiac and respiratory function (Guarneri, Bertolini, and Latronico, 2008). This phenomenon is known as post-Intensive Care syndrome (Myers, Smith, Allen, and

Address for correspondence: Luke Hardcastle, Department of Statistical Science, Faculty of Mathematical and Physical Sciences, University College London, Gower Street, WC1E 6BT, London, UK
E-mail: luke.hardcastle.20@ucl.ac.uk

10.1177/1471082X241238810

Kaplan, 2016) and can include a decline in physical, psychological and/or cognitive status. More than 224 000 patients are admitted to Intensive Care in the UK every year (Intensive Care National Audit and Research Centre, 2019) and only 50% of those previously employed will have returned to work one year after admission due to ongoing health issues (Kamdar, Suri, Suchyta, Digrande, Sherwood, Colantuoni, Dinglas, Needham, and Hopkins, 2020). Reducing the impact of post-Intensive Care syndrome is therefore a crucial issue.

Rehabilitation sessions while the patient is still receiving mechanical ventilation in the ICU involve progressing patients through activities such as sitting over the edge of the bed and standing and walking, and are considered to be the best way to ameliorate the impact of critical illness and its associated treatments on physical function.

The current approach to ICU rehabilitation is based on the default assumption that the metabolic cost of individual rehabilitation activities does not differ across patients and is similar to that of healthy individuals, resulting in patients receiving a broad one-size-fits-all approach to rehabilitation. This assumption, however, was shown to be invalid by Black, Grocott, and Singer (2020), who found high levels of variation in absolute exercise intensity both between patients and within patients over time. This suggests that under the current approach to rehabilitation individuals with very different physiological profiles receive similar, and often sub-optimal, exercise programmes.

Consequently, patients are often being under- or over-exercised, in the former case receiving minimal benefit from rehabilitation and in the latter being subjected to severe physiological stress. To address these issues, a scientific method for quantifying the exercise intensity level of an individual patient in real-time is required.

Exercise load during rehabilitation can be quantified by measuring a patient's rate of oxygen consumption ($\dot{V}O_2$), but these measurements are not available in many ICU. In addition, the equipment required to measure $\dot{V}O_2$ is costly and requires technical expertise, meaning it is unlikely to be introduced as a regular feature in critical care in the foreseeable future (Black, Grocott, and Singer, 2015).

While $\dot{V}O_2$ cannot be easily measured in the ICU, data are collected on a number of other physiological covariates as standard, many of which have known relationships to $\dot{V}O_2$.

The primary contribution of this article is to develop and validate a first-of-its-kind prediction model of $\dot{V}O_2$ for mechanically ventilated Intensive Care patients, through the development of a Bayesian hierarchical model with temporally correlated latent Gaussian processes. This model can provide clinicians with predictions of patients' levels of absolute exercise intensity allowing rehabilitation sessions to be personalised in real-time—both maximising the benefit and minimising potential harm to the patient.

The structure of this article is as follows. We begin with a presentation of the data in Section 2, followed by a discussion of the model development process in Section 3. In Section 4 we present our results before assessing predictive performance in Section 5. We conclude with a discussion in Section 6.

2 Background and data

Rehabilitation in Intensive Care is considered fundamental to improving physical function outcomes for patients post-ICU. During rehabilitation, patients progress through a number of activities, including sitting up in bed, sitting over the edge of the bed and standing up. A key role of the clinician

Table 1 Physiological variables recorded in the dataset.

Variable	Description
$\dot{V}O_2$	Rate of oxygen consumption. Measured as the volume of oxygen consumed per minute and standardised by a patients weight. Values are not readily available for most Intensive Care patients and are measured through BBGEA.
V_T	Tidal volume, the amount of air moved in and out of the lungs during a single breath.
RR	Respiratory rate, the number of breaths per minute.
$\dot{V}_E$	Minute ventilation, the amount of air breathed per minute, defined as the product of tidal volume and respiratory rate.
$P_{ET}CO_2$	End-tidal CO_2, the amount of CO_2 in exhaled air on a breath-by-breath basis.
VCO_2	Volume of exhaled carbon dioxide.

in these sessions is to ensure the patient is receiving the optimal level of exercise load at a given time and is not being over- or under-exerted. For example, they may finish an activity early if they believe a patient is being exposed to too much stress, or move onto more strenuous activities if a patient appears to be receiving no benefit from an exercise.

This is a challenging task, however, as due to the heterogeneous nature of the Intensive Care population the exercise load of each activity can vary dramatically from patient to patient, and for a given patient over their stay in the ICU (Black et al., 2020).

The standard measure of exercise load outside of critical care is rate of oxygen consumption expressed as $\dot{V}O_2$ (see Table 1), which provides an indicator of current exertion levels on a breath-by-breath basis regardless of activity. In an ideal world a clinician would have access to this measurement, and would be able to reactively adjust rehabilitation plans in real-time in response to these values.

To measure $\dot{V}O_2$ breath-by-breath gas exchange analysis (BBGEA) is used, a common technique for healthy individuals but one that has only recently been applied to the critical care environment (Black et al., 2015; Sommers et al., 2019). This method requires a high level of technical expertise and is difficult and costly to implement consistently (Black et al., 2020). It is therefore unlikely to become available as standard in Intensive Care in the near future.

This motivates the primary objective of this work—to model and provide predictions of $\dot{V}O_2$ based on measurements which are readily accessible in Intensive Care. These predictions can then be used to personalise rehabilitation sessions as they are being conducted.

In practice, breath-by-breath values of $\dot{V}O_2$ would be useful for clinicians, but challenging to interpret while also leading a rehabilitation session. Therefore, while $\dot{V}O_2$ values are modelled directly, predicted $\dot{V}O_2$ is classified as being at rest ($\dot{V}O_2 < 3.5$), low ($3.5 \leq \dot{V}O_2 < 5$), medium ($5 \leq \dot{V}O_2 < 7.5$) or high ($\dot{V}O_2 \geq 7.5$), with at rest corresponding to a patient experiencing no exercise load, and high corresponding to a patient being over-exerted. These classifications were developed in consultation with clinicians working on the study and represent current clinical best practice.

2.1 Data

The data are from the observational study conducted by Black et al. (2020), which used BBGEA to take measurements as mechanically ventilated ICU patients participating in various rehabilitation activities. All patients in the study had been persistently critically ill and/or mechanically ventilated for more than 7 days. BBGEA records measurements on a breath-by-breath basis, resulting in high frequency data consisting of 74 332 measurements from 37 patients and 103 rehabilitation sessions. Individual measurements are therefore indexed by a continuous time index t. They are naturally hierarchical in nature with repeated measurements within sessions and often multiple sessions per patient.

Available physiological covariates are defined in Table 1. The data are dominated by lower values of $\dot{V}O_2$, with 45.83% and 38.11% of observations classified as rest and low $\dot{V}O_2$ and 2.72% of observations classified as high. Additionally standard baseline characteristics of patients were measured such as body mass index (BMI), sex, age and the patient's pre-admission physical activity level given by the General Practice Physical Activity Questionnaire (GPPAQ). The Sequential Organ Failure Assessment (SOFA) score, a measure of organ system performance, was re-assessed before each session.

Finally the time since the start of the session was recorded, as well as the quality of session as assessed by the clinician running the study (classified as good, reasonable or poor), with poorer quality sessions being potentially more susceptible to measurement error in any of the physiological covariates, typically due to instability of the ventilator delivered oxygen levels and patients coughing.

2.1.1 Exploratory analysis

Exploratory analysis revealed a number of relationships between covariates in the data. Figure 1 features example plots of these relationships for individual rehabilitation sessions. Figure 1 (A) shows us that while there is a non-linear relationship between V_T and $\dot{V}O_2$ on the natural scale, this relationship is brought closer to linearity when placed on the log-log scale (see Figure 1 (B)). Figure 1 (C) shows us that this relationship similarly holds for $\dot{V}_E$—unsurprisingly given that $\dot{V}_E$ is the product of V_T and respiratory rate.

Figure 1 (B) features V_T and $\dot{V}O_2$ measurements from two patients. Note how for each patient, the relationship between V_T and $\dot{V}O_2$ appears linear, but the gradient differs, indicating that there is between-patient heterogeneity present. Furthermore, for a small minority of sessions, these relationships were not linear even on the log-log scale as shown in Figure 1 (D). This suggests that simple clinical rules to predict $\dot{V}O_2$ directly would typically fail in practice, and more sophisticated modelling tools are required.

For both respiratory rate and $P_{ET}CO_2$ there was weak evidence for linear relationships with $\dot{V}O_2$ when placed on the log-log scale, however examining plots of their product with $\log(V_T)$ against $\dot{V}O_2$ suggested that interactions may be present. Finally an interaction between age and BMI was also identified graphically, as shown in the supplementary material.

2.1.2 Data features

We briefly discuss some additional features of the data relevant to the analysis. First, patients were liable to cough during rehabilitation sessions, potentially disrupting the relationships between covariates. The effect of coughing on $\dot{V}O_2$, V_T and $\dot{V}_E$ is circled in Figure 2.

Second, measurements were taken over time and on a breath-by-breath basis. The result is that as respiratory rate increases the rate of measurements increases and vice versa, resulting in inconsistent intervals between measurements.

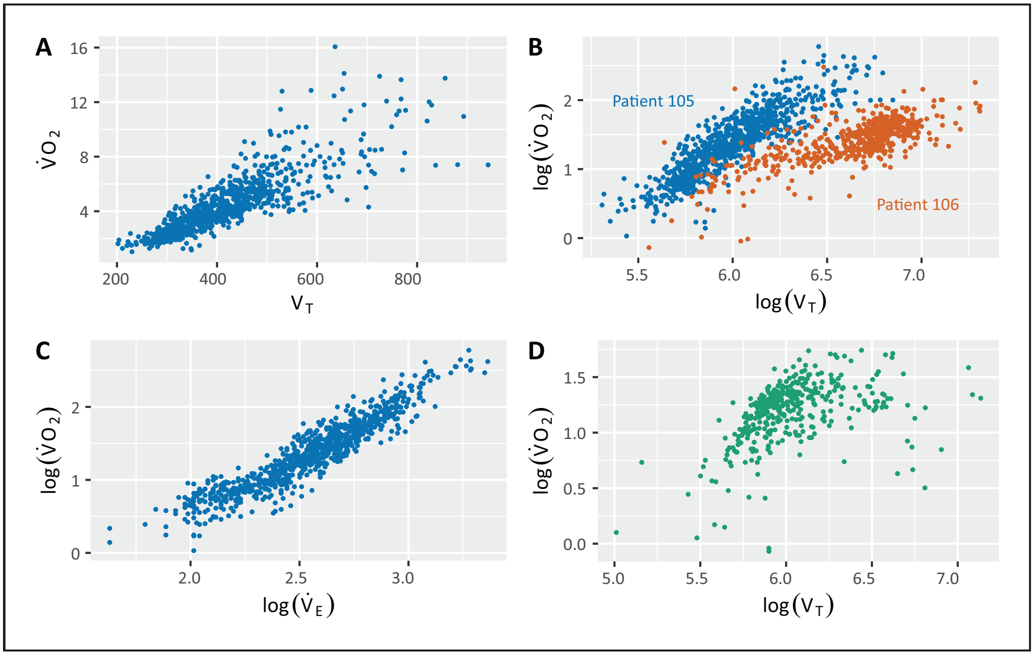

Figure 1 Relationships between $\dot{V}O_2$ and the physiological covariates in the data, plots are colour coded by patient. (A) $\dot{V}O_2$ against V_T. (B) $\dot{V}O_2$ against V_T on the log-log scale for patients 105 and 106 to show the heterogeneity in the effect of V_T on $\dot{V}O_2$. (C) $\dot{V}O_2$ against $\dot{V}_E$ on the log-log scale. (D) $\dot{V}O_2$ against V_T on the log-log scale for patient 117, where the relationship is decidedly non-linear.

The data were subject to measurement error from multiple sources. The BBGEA methodology is imperfect and may naturally lead to incorrect measurements for all the physiological covariates. The most obviously incorrect values these produced were removed from the data using set criteria, but incorrect values will still be present.

Finally, the patient population was relatively homogeneous in terms of age and BMI. Following the advice of the clinician who ran the study, two patients were removed from the analysis aged <40, as their age and physiological profiles were sufficiently dissimilar to the other patients that they could not be viewed as exchangeable with the rest of the study population. The minimum, lower quartile, median, upper quartile and maximum values for age after their removal was (50.2, 58.0, 69.7, 77.4, 86.0). Following their exclusion all patients were over 50 and had been mechanically ventilated for ≥ 7 days, providing a sufficiently exchangeable population for which the model should be restricted to in practice.

2.2 Existing physiological relationships

Documented relationships already exist between $\dot{V}O_2$ and the physiological measures introduced in the preceding sections. These have helped motivate this work and so for completeness we provide a brief overview of them here. As these relationships only hold in specific instances (Ward, 2007), however, they were not directly used in developing the model.

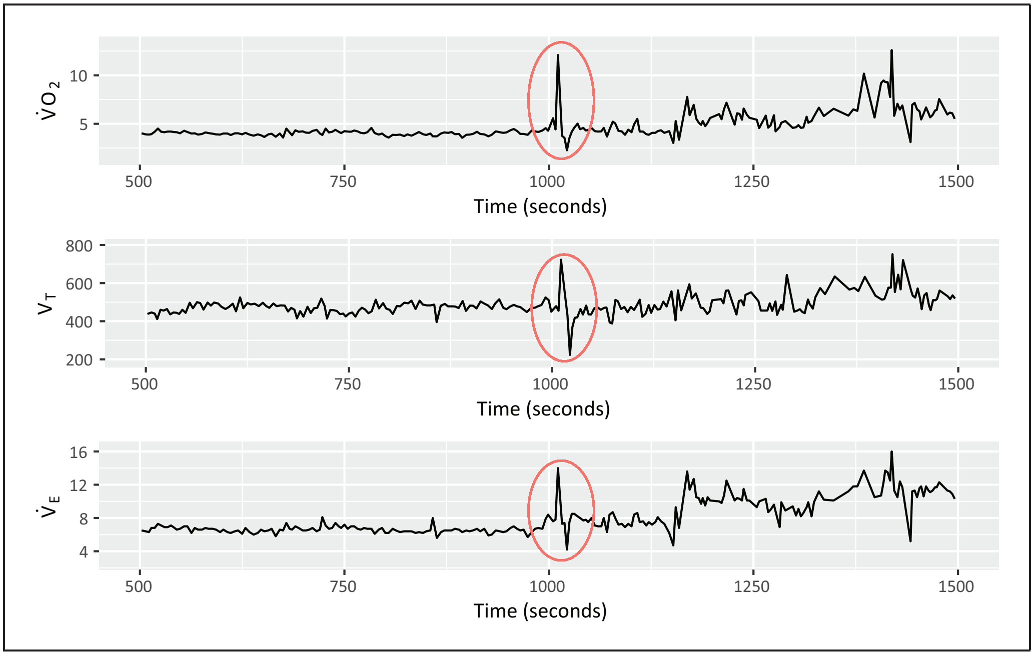

Figure 2 Plots showing $\dot{V}O_2$, $\dot{V}_E$ and V_T over time for session 73, with disruption caused by a cough circled.

The relationship that has received the most attention in previous research is that between $\dot{V}O_2$ and minute ventilation ($\dot{V}_E$) (Ramos, Alencar, Treptow, Arbex, Ferreira, and Neder, 2013). An existing relationship between the two quantities is the Oxygen Uptake Efficiency Slope (OUES) (Baba, Nagashima, Goto, Nagano, Yokota, Tauchi, and Nishibata, 1996), where oxygen uptake efficiency is the amount of gas an individual needs to breathe in and out in order to utilise a given volume of oxygen.

The OUES is defined by the equation $\dot{V}O_2 = a \times \log_{10}(\dot{V}_E) + b$, and is simply the change in $\dot{V}O_2$ as $\log(\dot{V}_E)$ increases. The OUES is important in the exercise intensity literature as a measure of physical fitness independent of a patient's motivation to exercise (Baba et al., 1996). This can also be extended to further physiological covariates as minute ventilation is defined as the product of tidal volume (V_T) and respiratory rate (RR). Additional established relationships indirectly used in this work are detailed in the supplementary material.

3 Model development

We have developed a hierarchical model to predict $\dot{V}O_2$ with temporally correlated latent Gaussian processes to account for the structure and heterogeneity present in the data. The model has been fitted using a Bayesian approach. This is particularly advantageous in this setting as prior regularisation suppresses poor estimates of group or effect variance (see for example Chung, Rabe-Hesketh, Dorie, Gelman, and Liu (2013)), and prior information can be used to encode preference for clinically plausible effect sizes. Further, for ease of clinical use, the model provides classifications of $\dot{V}O_2$ to clinicians. This approach allows for uncertainty around $\dot{V}O_2$ predictions to be easily propagated through the model and used to quantify uncertainty about classifications. Code to fit the model is available at: https://github.com/LkHardcastle/ExerciseIntensityExample/

3.1 The model

We indicate the value of $\log(\dot{V}O_2)$ taken at time t for the i^{th} patient's j^{th} rehabilitation session as y_{ijt}, where t is the time in seconds since the start of the session. We then assume that

$$y_{ijt} \mid \mu_{ijt}, \tau \sim \text{Normal}(\mu_{ijt}, \tau), \tag{3.1}$$

where τ indicates the precision.

We then specify μ_{ijt} as a linear predictor of the form

$$\begin{aligned} \mu_{ijt} = \quad & \alpha_{ij} + \beta_{1i} \log((V_T)_{ijt}) + \beta_2 \log((P_{ET}CO_2)_{ijt}) + \beta_3 \log(RR_{ijt}) + \\ & \beta_4 \log((V_T)_{ijt}) \times \log((P_{ET}CO_2)_{ijt}) + \beta_5 \log((V_T)_{ijt}) \times \log(RR_{ijt}) + \\ & [\ldots] + s_{ijt} \, . \end{aligned} \tag{3.2}$$

Figure 3 shows a directed acyclic graph representing the modelling assumptions. Note that in the above [...] is used for brevity to indicate inclusion of session and patient-level covariates in the linear predictor — in this case age, BMI, GPPAQ and SOFA score, as well as an interaction between age and BMI as identified previously. Further, α_{ij} and β_{1i} denote hierarchical effects and s_{ijt} denotes a temporal effect, discussed later in this section.

Covariates were included based on empirical findings and their clincial relevance as determined by a clinician. Although, as shown in Figure 1 (C), $\dot{V}_E$ has a strong linear relationship with $\dot{V}O_2$, we have chosen to decompose $\dot{V}_E$ into V_T and respiratory rate to maximise the information provided to the model.

All continuous covariates and $\dot{V}O_2$ are placed on the log-scale and centered around their sample means to aid interpretation and for numerical stability. Further, pairwise interactions between V_T and both $P_{ET}CO_2$ and respiratory rate are included as motivated by exploratory analysis findings. The unstructured coefficients, that is, those without a hierarchical prior structure, were given minimally informative Normal(0, 0.1) priors, corresponding to a prior belief that effect sizes of over 6.3 were highly unlikely.

The remainder of this section is focused on defining hierarchical and temporal effects and their corresponding prior distributions to account for the structured nature of the data.

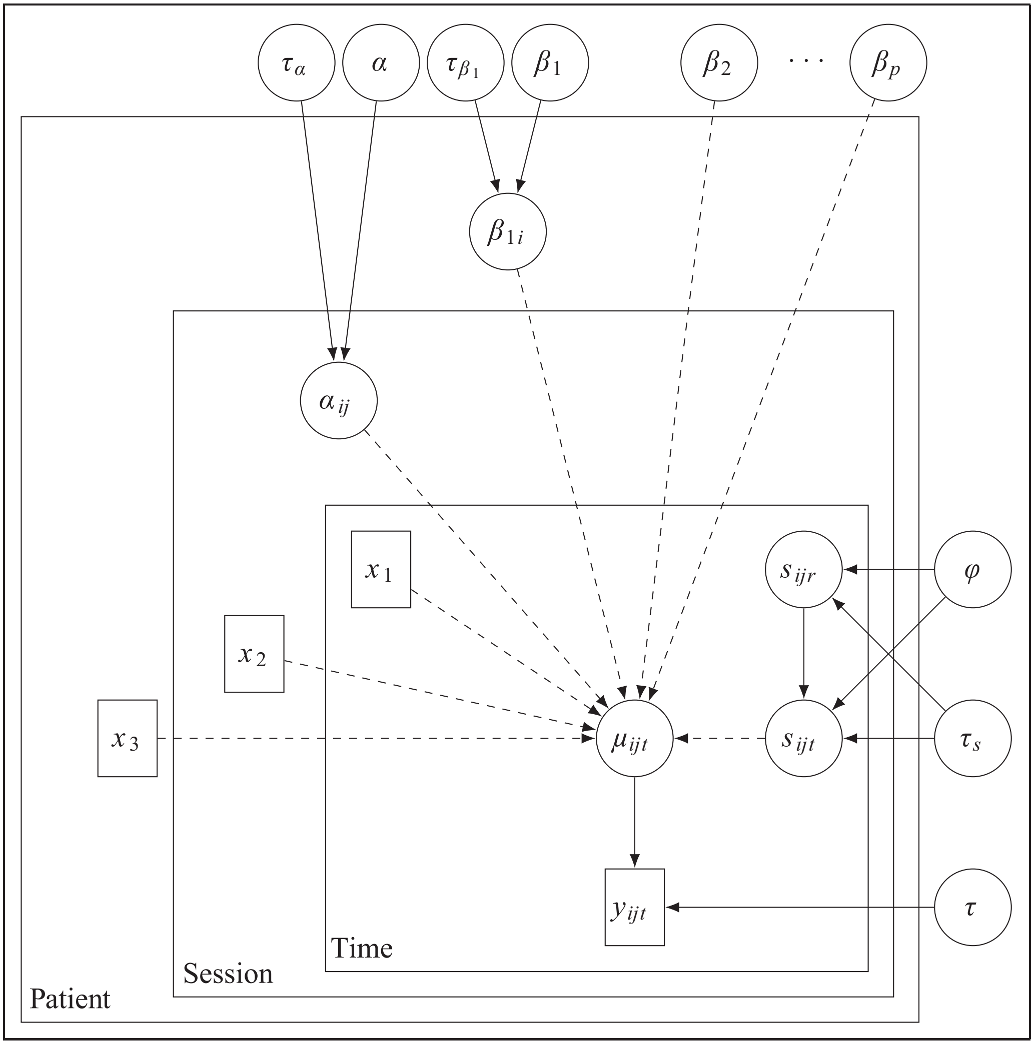

Figure 3 A directed acyclic graph representing the model assumptions outlined in Section 3.1. Solid and dashed lines represent distributional and deterministic relationships respectively, and x_1, x_2 and x_3 represent vectors of covariates.

3.1.1 Structured effects

The between-patient and between-session heterogeneity in the model is accounted for through two terms in (3.2) — a session-level varying intercept term and a patient-level varying coefficient for the effect of $\log(V_T)$.

These individual coefficients would be unable to be inferred in a practical implementation of this model without available prior training data at the individual level. Their inclusion, however, ensures that the model accurately quantifies the uncertainty in marginal $\dot{V}O_2$ predictions and marginal inference for the population-level coefficients described in the previous section.

The structured coefficients are defined as

$$\alpha_{ij} \mid \mu_\alpha, \tau_\alpha \sim \text{Normal}(\mu_\alpha, \tau_\alpha), \tag{3.3}$$

$$\beta_{1i} \mid \mu_{\beta_1}, \tau_{\beta_1} \sim \text{Normal}(\mu_{\beta_1}, \tau_{\beta_1}), \tag{3.4}$$

with Normal(0, 0.1) priors for μ_α and μ_{β_1} and non-informative, proper log-Gamma priors for τ_α and τ_{β_1}.

The structured intercept term in (3.3) is specified at the session-level. Figure 4 shows plots of predicted $\log(\dot{V}O_2)$ values against observed $\log(\dot{V}O_2)$ for models with the structured intercepts at the patient- and session-level, with observations highlighted by session. After accounting just for between-patient heterogeneity there is still high between session variation. This motivates the decision to capture this through a hierarchical prior structure at the session-level, rather than a simpler patient-level structure.

Figure 1 provides evidence for a heterogeneous relationship between V_T and $\dot{V}O_2$, which motivates the inclusion of individual patient-level coefficients, (3.4). To examine further whether individual coefficients were required, we re-fit the model separately for each patient and then examined the resulting marginal posteriors for the V_T coefficient. The effect of V_T on $\dot{V}O_2$ had a high level of heterogeneity between patients, indicating we need to account for a heterogeneous effect of V_T at the patient-level. Graphical summaries of these posteriors can be found in the supplementary material.

3.1.2 Temporal effects

In (3.2) s_{ijt} is a temporal error term. We model this using an Ornstein-Uhlenbeck process (O-U process) (Øksendal, 2003). This is a Markov process such that if the previous realisation of the process was at time $r < t$, we have

$$s_{ijt} \mid \phi, \tau_s, s_{ijr} \sim \text{Normal}(\mu_*, \tau_*),$$

with

$$\mu_* = s_{ijr} \exp(-\phi|t-r|), \quad \tau_* = \tau_s \Big(1 - \exp\left(-2\phi|t-r|\right)\Big)^{-1}.$$

Further, if t is the first observation time for the j^{th} session, then

$$s_{ijt} \sim \text{Normal}(0, \tau_s^{-1}).$$

Here τ_s is the precision of the process at stationarity and ϕ is a mean-reversion parameter, which intuitively represents the level of similarity between consecutive observations, with smaller values of ϕ indicating higher levels of similarity. A vague Normal(0,0.2) prior was used for $\log(\phi)$ and a weakly informative log-gamma(50,1) was used for τ_s, corresponding to a prior belief that there is a 2% probability of precision values higher than 200.

Temporal dependence between observations is more commonly characterised using the discrete-time analogue to the O-U process — the $AR(1)$ process (Blangiardo, Cameletti, Baio, and Rue,

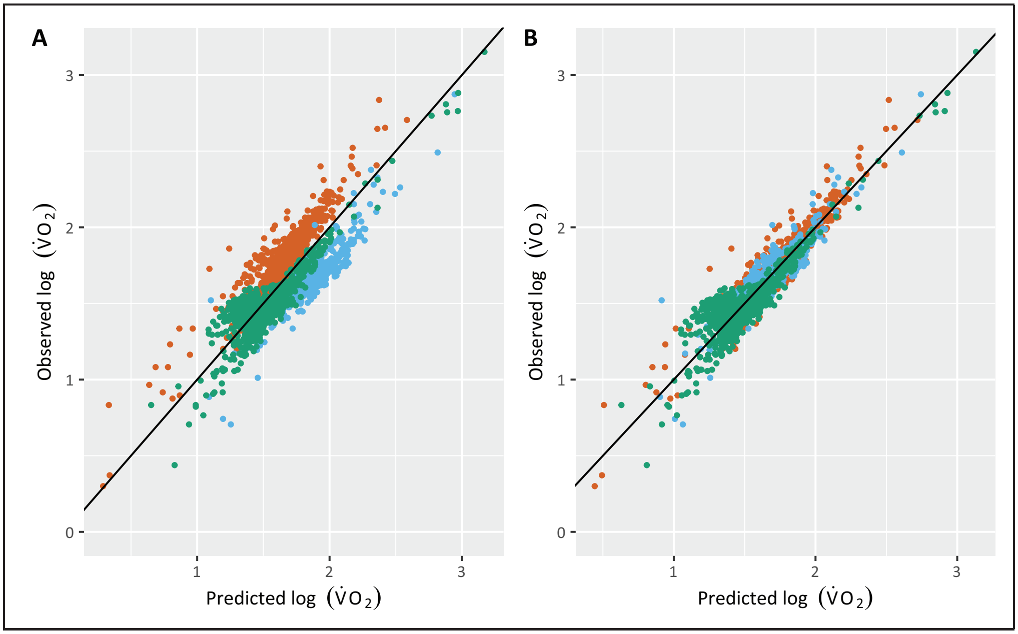

Figure 4 Plots of predicted $\dot{V}O_2$ against actual $\dot{V}O_2$ for patient 112, coloured by session for models with individual intercepts at the patient (A) and session (B) level. There is clear between session heterogeneity in (A) which is only resolved with the session-level intercept in (B).

2013). The primary advantage of using the O-U process in this setting is that it naturally accounts for irregular intervals between observations, which arise from breath-by-breath measurements.

The specified model is subject to measurement-level variance from both the temporal effects and the iid error terms. Special cases of the model can be derived by: dropping the temporal effects and assuming observations within the same session are independent conditional on their covariates; setting $\tau = e^{15}$ as seen in Wang, Yue, and Faraway (2019), forcing all the observational error to appear through the temporal error term, and assuming that the measurement-level variance of $y_{ij,t+1} \mid y_{ijt}$ vanishes as $t \to 0$.

We will refer to the model including both the O-U process and Gaussian error as the full model and the special cases as the O-U only and iid error models. We consider all three models at the validation stage of the analysis, in Section 5.4.

3.2 Inference

There are two primary inferential challenges presented by this model. First, the varying intercept and V_T coefficient terms naturally introduce non-linearity skewing the posterior, and second, due to the presence of temporal effects, the dimension of the posterior grows with the number of observations, and there is often (by construction) strong posterior dependence between the high-dimensional latent variable and the hyperparameters associated with it. Both of these challenges can hamper the performance of traditional MCMC methods. The model presented in (3.2) was therefore fitted using the Integrated Nested Laplace Approximation (INLA), with the `R-INLA` package (Rue, Riebler, Sørbye, Illian, Simpson, and Lindgren, 2017).

INLA approximates the marginal posterior distributions of interest via repeated, nested Laplace approximations and numerical integration over a low-dimensional hyperparameter space. In contrast to simulation based methods, the comparatively low cost of the Laplace approximations means the dimension of certain posteriors (such as those induced by the inclusion of temporal effects as outlined in Section 3.1.2) can grow with the number of observations with only a minor impact on computation time.

Numerical intergration over the hyperparameters means that complex dependencies between parameters and hyperparameters can be dealt with effectively albeit at the cost of enforcing that the dimension of the hyperparameters is low.

3.3 Oxygen uptake efficiency slope model

The previous state-of-the-art approach to this problem was a model developed based on the OUES (Section 2.2) and fitted using the `lme4` R package (Bates, Mächler, Bolker, and Walker, 2015) by Black, Singer, and Grocott (2017). We reproduce this model here as a baseline comparator for our approach.

Let y_{it} be the $\dot{V}O_2$ value for the i^{th} patient at time t. The OUES model is then defined as

$$y_{it} = \alpha_i + \beta_i \log_{10}((\dot{V}_E)_{it}) + \epsilon_{it}, \tag{3.5}$$
$$\alpha_i \sim \text{Normal}(0, \tau_\alpha), \quad \beta_i \sim \text{Normal}(0, \tau_\beta).$$

where α_i and β_i are random effects accounting for between patient heterogeneity. This naturally corresponds with the notion that the coefficients in the OUES are unique for each individual, but given that we are considering a relatively homogeneous population there should not be drastic differences in coefficient values.

4 Results

Table 2 shows posterior summaries for hyperparameters and the fixed effects for the full model, model with only the O-U process and the model with only the iid error terms. The posteriors for the coefficients of the physiological covariates indicate that the effect of each covariate on $\dot{V}O_2$ is positive, as expected. Further, given the posteriors for the interactions, the effect of $\log(V_T)$ increases as respiratory rate increases and decreases as $P_{ET}CO_2$ increases in the full model. All session and patient-level covariate effects are in the expected direction, or their posteriors are centered close to 0. The posterior summaries of τ_α and τ_β both indicate that there is non-negligible heterogeneity in the model. Notably, the finding of high between session heterogeneity aligns with the findings of Black et al. (2020).

The fixed effect marginal posteriors for the full and O-U only models are very similar, however the effect sizes of $\log(V_T)$, $\log(P_{ET}CO_2)$ and $\log(RR)$ are noticeably smaller in the model without the O-U process. The inclusion of the O-U process naturally results in an increase in τ, as it introduces a second source of measurement-level variance, and also sees an increase in τ_α. The marginal

Table 2 Posterior summaries for the fixed effects coefficients and hyperparameters of the model defined in (3.2) and its variations discussed in Section 3.1.2, with posterior means and 95% credible intervals.

	Estimates (mean and 95% credible intervals)		
	Full model	O-U only	iid error only
Intercept	1.33 (1.11, 1.55)	1.32 (1.14, 1.50)	1.35 (1.17, 1.52)
$\log(V_T)$	1.52 (1.48, 1.57)	1.56 (1.52, 1.61)	1.31 (1.26, 1.36)
$\log(P_{ET}CO_2)$	1.91 (1.90, 1.92)	1.92 (1.91, 1.93)	1.32 (1.30, 1.33)
$\log(RR)$	1.09 (1.09, 1.10)	1.11 (1.10, 1.11)	0.87 (0.87, 0.88)
$\log(V_T) \times \log(P_{ET}CO_2)$	−0.11 (−0.13, −0.11)	−0.20 (−0.22, −0.17)	0.14 (0.10,0.17)
$\log(V_T) \times \log(RR)$	0.30 (0.29, 0.31)	0.33 (0.32, 0.34)	0.33 (0.32, 0.34)
SOFA	−0.01 (−0.05, 0.04)	−0.01 (−0.05, 0.03)	−0.01 (−0.04, 0.03)
GPPAQ = 2	−0.01 (−0.11, 0.11)	−0.01 (−0.11, 0.09)	−0.04 (−0.14, 0.05)
GPPAQ = 3	0.01 (−0.11, 0.14)	0.01 (−0.09, 0.11)	−0.02 (−0.12, 0.08)
GPPAQ = 4	0.29 (0.11, 0.48)	0.31 (0.16, 0.46)	0.16 (0.01, 0.03)
Sex	−0.07 (−0.18, 0.04)	−0.08 (−0.17, 0.00)	−0.09 (−0.17, 0.00)
$\log(age)$	0.34 (0.05, 0.64)	0.35 (0.11, 0.59)	0.17 (−0.07, 0.40)
$\log(BMI)$	−1.12 (−1.33, −0.91)	−1.13 (−1.31, −0.96)	−0.96 (−1.13, −0.79)
$\log(age) \times \log(BMI)$	−2.06 (−3.59, −0.53)	−2.12 (−3.39, −0.96)	−1.47 (−2.69, −0.24)
τ	256.72 (253.05, 260.44)	–	64.04 (63.38, 64.70)
τ_α	48.39 (33.04, 64.62)	45.75 (32.74, 57.68)	36.50 (14.80, 59.23)
τ_β	31.07 (23.68, 37.83)	44.89 (37.42, 50.63)	35.93 (26.47, 49.79)
O-U τ	41.00 (36.05, 47.43)	46.75 (45.17, 47.90)	–
O-U ϕ	0.04 (0.03, 0.05)	0.09 (0.09, 0.10)	–

posteriors for τ_α are similar in both the full model and the model without the iid error terms, however the full model has lower values of τ_β.

4.1 Additional model checks

It was noted in Section 2.1.2 that some of the covariates were potentially subject to an unknown degree of measurement error, with the clinician running the study able to classify rehabilitation sessions as poor, reasonable or good depending on the potential for this error to occur. If this measurement error is large enough, it may adversely effect the predictions of the model.

In the supplementary material we re-fit the model to a subset of the data excluding poor quality sessions and another smaller subset excluding both poor and reasonable quality sessions. A similar technique is commonly seen in study meta-analysis (Valentine, 2009). The resulting posteriors show no noticeable changes in the locations of the posteriors for the covariate effects or hyperparameters.

5 Validation

In this section we internally validate the predictive ability of the model. This is done, first, through in time posterior predictive checks, to understand how the model performs given maximal information, and then second, using leave-one-patient-out cross-validation to understand how the model performs in realistic conditions.

5.1 Posterior predictive checks

To perform posterior predictive checks in light of hierarchical and temporal structure we re-fit the model except for data from session i, where we only use data observed before time $t = 1\,000$. We then use the model to predict $\dot{V}O_2$ for $t > 1\,000$ for session i. Here $t = 1\,000$ was chosen so that enough data were present to ensure model stability for the new patient and to allow for an initial calibration phase at the start of each session during which patient $\dot{V}O_2$ values were largely in the rest category.

The outcome of these checks for patient 137 are shown in Figure 5, with observed (black) and predicted (red) values of $\log(\dot{V}O_2)$.The majority of model heterogeneity comes at the session and patient-levels, meaning that credible intervals are relatively small here. In all cases, the model closely matches the shape of the $\dot{V}O_2$ curve, although the iid only model performs slightly worse in this regard. This aligns with the marginal posteriors for the fixed effects being similar for the first two models, but having reduced effects in the model without the O-U process as noted in Section 4.

Further, we note that the O-U only model has lower predictive uncertainty compared to the other models, which we believe is due to reduced flexibility when excluding the iid error term. We discuss the impact of model specification on posterior predictive uncertainty further in the Section 5.4.

In 2.1.2 we noted that patients coughing induces spikes in $\dot{V}O_2$ values. One such spike occurs after 1 450 seconds in Figure 5. In light of this a pre-analysis decision was made to consider a model fit to smoothed data using a three-value rolling average for the physiological covariates. Figure 5 shows, however, that the model adapts very well to this spike matching the rise and subsequent drop in $\dot{V}O_2$ very closely. The analysis using the smoothed data is available in the supplementary material, providing similar or worse predictive ability and no improvement during coughs compared to the non-smoothed data models.

5.2 Predictions in practice

The remainder of this section is dedicated to assessing how the model would perform under realistic conditions.

In practice, a clinician would require $\dot{V}O_2$ predictions in real-time (as the model can provide) at each stage of the rehabilitation session. They would be needed at the start of the session to ensure a baseline low level of activity; progressing to medium $\dot{V}O_2$ values to establish an exercise response; during the session to ensure the patient's $\dot{V}O_2$ values do not enter the high classification indicating over-exertion; and during breaks in the session or at the end of the session to ensure the patient's $\dot{V}O_2$ values have returned to a resting level. We therefore focus on assessing how the model performs observation by observation, rather than trying to predict a single event (e.g., time to first high observation).

5.3 Leave-one-patient-out cross-validation

To conduct cross-validation observations were left out at the patient-level, to account for similarity between observations within patients and prevent leakage. This resulted in 35 sets of observations to assess on re-fitted models. These were fit using the same specification as previously used.

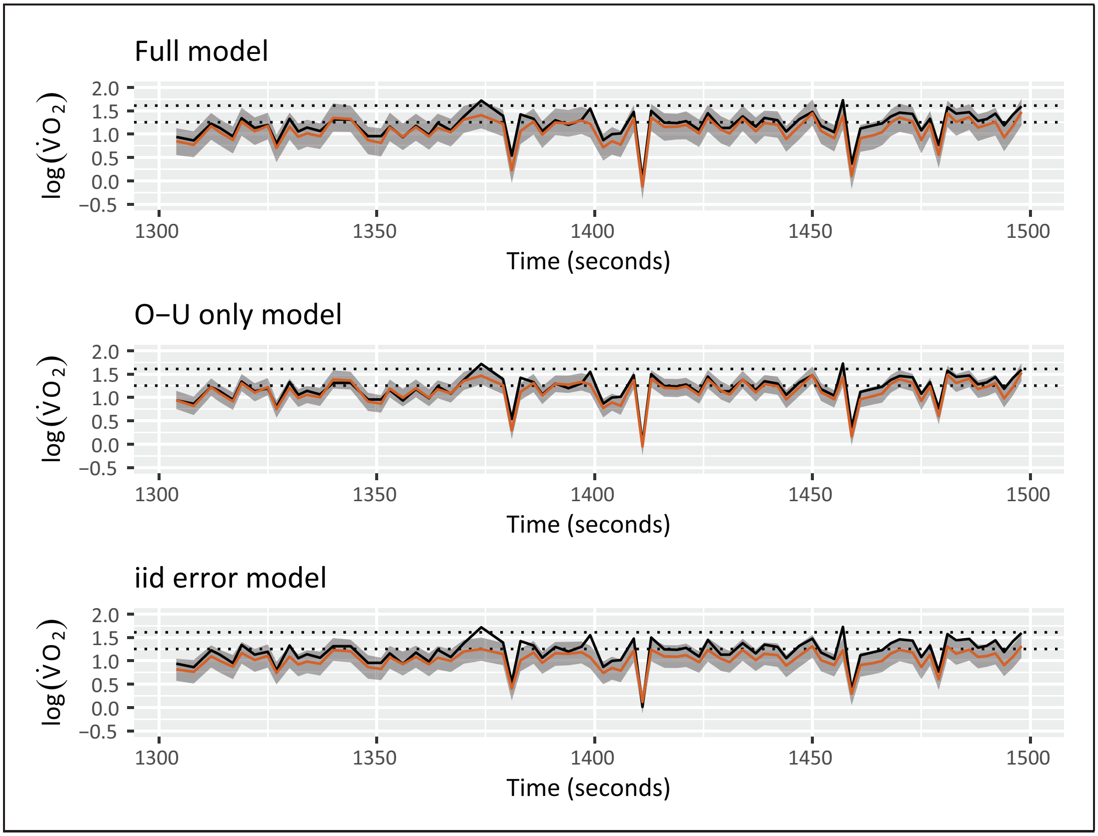

Figure 5 An example of posterior predictive checks using cross-validation in time for patient 137, for the three specifications of the latent Gaussian process. Orange lines indicate predicted $\dot{V}O_2$ values with corresponding predictive intervals in grey. Black lines show the observed $\dot{V}O_2$ values. Dotted lines represent the boundaries between rest and low, and low and medium exercise intensity classifications. Predictions are worse in the model excluding the O-U process and the inclusion of the iid error increases the predictive uncertainty.

Here the model is unable to learn individual coefficients, increasing the uncertainty in the classifications. Given that clinicians would not have access to $\dot{V}O_2$ values during regular rehabilitation sessions, this aligns with how the model would be implemented in practice.

To obtain the posterior predictive densities for each point, 1 000 samples were generated from the posterior of the model fit to data excluding each patient in turn. These samples were then used to make predictions for the excluded patient via the posterior predictive distribution, fully quantifying the uncertainty in the predictions.

Probabilities for each category were then easily generated using the proportion of samples from the posterior predictive density that fell into that category and, as this is the information that would be available to clinicians in practice, we classified the observations based on the category with the highest probability assigned to it.

Table 3 Confusion matrices for the predictive accuracy of the full, O-U error only, iid error only and OUES models as assessed using leave-one-patient-out cross-validation. $\mathcal{L}$ = the loss function defined in Section 5.4.

	Predicted							
	Full model				O-U error only			
Observed	Rest	Low	Med.	High	Rest	Low	Med.	High
Rest	0.74	0.21	0.05	<0.01	0.73	0.23	0.04	<0.01
Low	0.37	0.38	0.23	0.01	0.36	0.43	0.21	0.01
Medium	0.10	0.24	0.45	0.21	0.09	0.26	0.47	0.18
High	0.01	0.03	0.19	0.77	<0.01	0.03	0.22	0.74
Specificity	0.70	0.83	0.89	0.98	0.72	0.81	0.90	0.98
	$\mathcal{L}$ = 0.399				$\mathcal{L}$ = 0.412			
	iid error only				OUES			
Observed	Rest	Low	Med.	High	Rest	Low	Med.	High
Rest	0.75	0.24	<0.01	0.00	0.57	0.38	0.05	0.00
Low	0.37	0.53	0.10	<0.01	0.25	0.54	0.21	<0.01
Medium	0.09	0.35	0.50	0.05	0.07	0.40	0.52	<0.01
High	<0.01	0.05	0.42	0.52	0.01	0.17	0.72	0.10
Specificity	0.71	0.80	0.95	>0.99	0.80	0.60	0.86	0.99
	$\mathcal{L}$ = 0.424				$\mathcal{L}$ = 0.560			

5.4 Validation results

Table 3 presents the results of the cross-validation for the three Bayesian models and the current state-of-the-art OUES model. The performance of the models is quantified using a loss function $\mathcal{L}$ which averages across observed classifications and then across patients. This quantifies a model's predictive ability across all classifications, regardless of their underlying prevalence in the data. This is particularly useful in this setting as $\mathcal{L}$ reflects the clinical importance of accurately predicting classified high $\dot{V}O_2$ values, when $< 3\%$ of observed values meet this threshold.

All three Bayesian models outperform the OUES model in terms of the loss function, with the fully specified model the best of the three. The models incorporating the O-U process have the best performance in terms of the loss function and noticeably higher accuracy for observed high observations in comparison to the iid error only model and the OUES model, with the drop-off in accuracy in the second model being particularly dramatic (10% vs. 77% in the full model).

We note that the OUES model has slightly higher accuracy than the Bayesian models for observed low values, however the specificity for low classifications is only 60% compared to 80–83% in the Bayesian models. This suggests this increase in accuracy is due to the model predicting low more often rather than the model having an improved ability to predict low $\dot{V}O_2$ values. Further, it may initially appear from Table 3, that all the models have a tendency to under-predict exercise intensity, however, this is simply an artefact resulting from the distribution of values of $\dot{V}O_2$ present in the data being positive-skewed, and therefore typically lying closer to the lower end of the exercise intensity classifications.

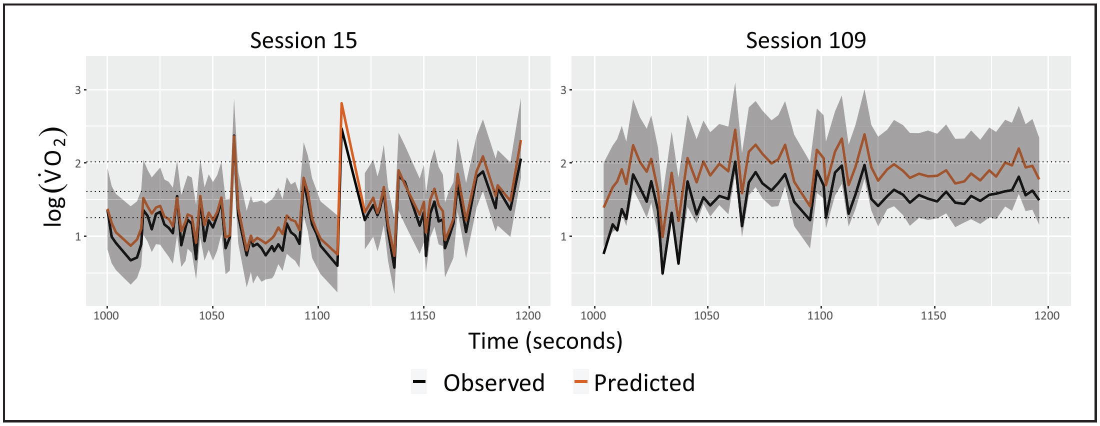

Figure 6 Compares predicted (with 95% credible intervals) values of $\dot{V}O_2$ from the full model and observed values for two example sessions, one where the model performs well (session 15) and one where it performs poorly (session 109). Dashed lines indicate the boundaries between exercise intensity classifications.

5.4.1 Predictions in time

To further understand model performance we directly examined the predictions for $\dot{V}O_2$. Figure 6 shows predicted $\dot{V}O_2$ values over time for two sessions, for the full Bayesian model. In both examples the models predict the shape of the $\dot{V}O_2$ curves extremely well (as expected from the posterior predictive checks). The plots for session 109 highlight, however, that there is high predictive uncertainty in the scale of the curve stemming from relatively small estimated values of τ_α and τ_β. The result is that in some cases (e.g., session 109) simple point predictions are insufficient, but true values of $\dot{V}O_2$ are still contained within 95% posterior predictive intervals. We explore this idea further in the following section.

5.5 Probabilistic predictions

Until now, we have primarily assessed the single classifications provided by the model. Probabilistic statements about the classifications are also given, however, and these may carry important information for clinicians. In Figure 7 we assess the usefulness of this information by showing the number of observations within each category that the model assigns probabilities above a certain threshold. Heuristically, this is a way of assessing when the model considers the correct classification to be 'plausible', based on a set probability threshold for plausibility.

This is a weaker condition than seen in Table 3 and we naturally see a larger number of observations assigned as plausible. For example, at the 20% threshold for plausible observations, 84.3%, 81.7% and 83.0% of observations are plausible in the full, O-U and iid models respectively when averaging over classifications. This benefit is most notable for the Low and Medium categories, where over 80% of observations are assigned as plausible for the 20% threshold.

For resting and low observations the iid error only model marginally outperforms the full model and both models perform better than the model with only the O-U process. For medium observations, the full model performs best, but all three models have similar performance. Crucially, for high observations the full model, and to an extent the O-U only model outperform the iid only model.

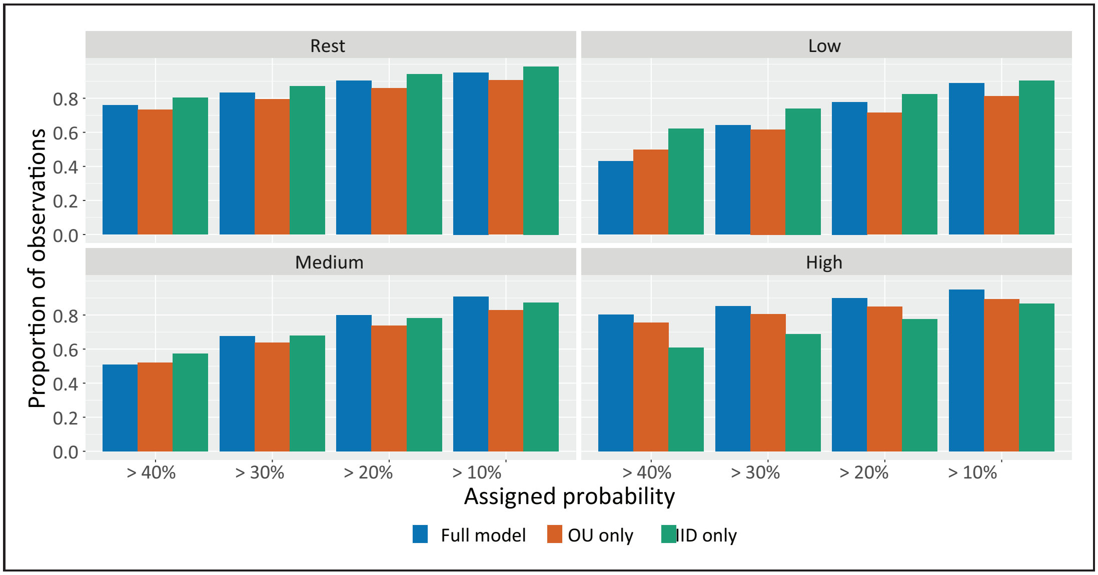

Figure 7 Graphs by observed exercise intensity category for the full model (blue) model with iid errors (red) and the iid error only (green) showing the proportion of observations for which the model considers the correct classification plausible, based on different plausibility thresholds.

To see how this information would be used in practice we can consider how a clinician would apply the output of the model. To avoid risks of over-exertion, a clinician may in practice apply an informal decision rule to lower a patient's exercise load once the probability they are at high intensity crosses a certain threshold.

This aligns exactly with the information showcased in Figure 7 where, given the rarity of high $\dot{V}O_2$ values, the 20% threshold would be a reasonable choice. At this threshold these results indicate that the full model would be able to identify 90.0% of high values, correctly guiding the clinician in the majority of cases.

This is only one example of how the full information from this model may be utilised and in practice a variety of rules may be implemented. It does indicate, however, that the full potential of the model is only realised once probabilistic statements about classifications are fully utilised.

6 Discussion

We have developed and internally validated a first-of-its-kind prediction model for exercise intensity in mechanically ventilated Intensive Care patients, using Bayesian hierarchical modelling and covariates that are readily available in the majority of Intensive Care settings.

The model revealed a high level of heterogeneity between exercise rehabilitation sessions, meaning that resulting classifications based on using point estimates alone had an accuracy of 58.4%. The usefulness of the model dramatically increases in practical situations when the full information from the posterior distribution is utilised, in the form of probabilistic statements about classifications, allowing clinicians to make the correct decision around 90.0% of the time if appropriate decision rules are adopted.

We have presented three feasible Bayesian models based on various assumptions on the measurement-level variance, the best performing of which is the fully specified model with both an O-U process and iid Gaussian error. Without this temporal structure, the model is unable to reliable classify high values of $\dot{V}O_2$, which is particularly clinically important in ensuring that patients are not over-exerted during rehabilitation sessions. All models were compared to the current state-of-the-art model based on the oxygen uptake efficiency slope, and were found to have noticeably superior performance.

In practice application of the method presented here should be restricted to patients who are sufficiently similar to the sample used to develop the model, specifically patients over 50 years of age who have been mechanically ventilated for at least 7 days.

The ultimate aim of the modeling work presented here is the creation of a practical tool that can be widely implemented in critical care environments (although further data collection and external validation in clinical settings would be needed for this). The model here returns live predicted classifications of a patient's level of exercise intensity on a breath-by-breath basis, allowing clinicians to use this tool to provide personalised rehabilitation sessions by reactively increasing or decreasing the level of exercise, therefore optimising the benefit received from these sessions.

Acknowledgements

We would like to thank the editor, associate editor and two anonymous referees whose comments lead to a much improved manuscript. We would like to thank the UCL CHIMERA group, and in particular Professor Christina Pagel and Professor Rebecca Shipley for their invaluable comments and feedback.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

LH acknowledges support from EPSRC grant number EP/W523835/1 during part of this work. This work forms part of the research activity of the ESPRC UCL CHIMERA hub for Mathematics in Health Care (Grant number EP/T017791/1). CB was funded by a National Institute for Health Research Clinical Academic Training Fellowship. This report presents independent research funded by the NIHR. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Supplementary materials

The supplementary materials include additional exploratory data analysis, an extended description of established physiological relationships between covariates in the data, model results from additional models fit during model development and the results of the analysis using a smoothed version of the data. They are available from the journals page at https://journals.sagepub.com/home/smj.

References

Baba R, Nagashima M, Goto M, Nagano Y, Yokota M, Tauchi N and Nishibata K (1996) Oxygen uptake efficiency slope: A new index of cardiorespiratory functional reserve derived from the relation between oxygen uptake and minute ventilation during incremental exercise. *Journal of the American College of Cardiology*, **28**, 1567–1572. doi: 10.1016/S0735-1097(96)00412-3.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. doi: 10.18637/jss.v067.i01.

Black C, Grocott MPW and Singer M (2015) Metabolic monitoring in the intensive care unit: a comparison of the Medgraphics Ultima, Deltatrac II, and Douglas bag collection methods. *BJA: British Journal of Anaesthesia*, **114**, 261–268. doi: 10.1093/BJA/AEU365. URL https://academic.oup.com/bja/article/114/2/261/295706.

Black C, Singer M and Grocott M (2017) Estimating oxygen consumption from minute ventilation (VE) during rehabilitation in mechanically ventilated patients recovering from critical illness. *American Journal of Respiratory and Critical Care Medicine*, pages A1776–A1776. doi: 10.1164/ajrccm-conference.2017.195.1_MeetingAbstracts.A1776. URL https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2017.195.1_MeetingAbstracts.A1776.

Black C, Grocott M and Singer M (2020) The oxygen cost of rehabilitation interventions in mechanically ventilated patients: an observational study. *Physiotherapy (United Kingdom)*, **107**, 169–175. doi: 10.1016/j.physio.2019.06.008.

Blangiardo M, Cameletti M, Baio G and Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, **4C**, 33–49. doi: 10.1016/j.sste.2012.12.001.

Chung Y, Rabe-Hesketh S, Dorie V, Gelman A and Liu J (2013) A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, **78**, 685–709.

Guarneri B, Bertolini G and Latronico N (2008) Long-term outcome in patients with critical illness myopathy or neuropathy: the Italian multicentre CRIMYNE study. *J Neurol Neurosurg Psychiatry*, **79**, 838–841. URL `PM: 18339730`. DA - 20080618.

Intensive Care National Audit and Research Centre (2019). Annual quality report 2018/19 for adult critical care. URL https:// onlinereports.icnarc.org/Reports/2019/12/annual-quality-report-201819-for-adult-critical-care.

Kamdar BB, Suri R, Suchyta MR, Digrande KF, Sherwood KD, Colantuoni E, Dinglas VD, Needham DM and Hopkins RO (2020) Return to work after critical illness: a systematic review and meta-analysis. *Thorax*, **75**, 17–27. doi: 10.1136/thoraxjnl-2019-213803.

Myers EA, Smith DA, Allen SR and Kaplan LJ (2016) Post-ICU syndrome: Rescuing the undiagnosed. *Journal of the American Academy of PAs*, **29**, 34–37.

Ramos RP, Alencar MCN, Treptow E, Arbex F, Ferreira EMV and Neder JA (2013) Clinical Usefulness of Response Profiles to Rapidly Incremental Cardiopulmonary Exercise Testing. *Pulmonary Medicine*, **2013**, 25. doi: 10.1155/2013/359021. URL /pmc/articles/PMC3666297/pmc/articles/

PMC3666297/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3666297/.

Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP and Lindgren FK (2017) Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, **4**, 395–421. doi: 10.1146/annurev-statistics-060116-054045. URL https://doi.org/10.1146/annurev-statistics-060116-054045.

Sommers J, Klooster E, Zoethout SB, van den Oever HLA, Nollet F, Tepaske R, Horn J, Engelbert RHH and van der Schaaf M (2019) Feasibility of exercise testing in patients who are critically ill: A prospective, observational multicenter study. *Arch Phys Med Rehabil*, **100**, 239–246.

Valentine JC (2009) Judging the quality of primary research. *The handbook of research synthesis and meta-analysis*, **2**, 129–146.

Wang X, Yue Y and Faraway J (2019) Bayesian regression modeling with INLA. Boca Raton: CRC Press., vol 75. doi: 10.1111/biom.13128.

Ward S (2007) Discriminating features of responses in cardiopulmonary exercise testing. *European Respiratory Monograph*, **40**, 36–68. doi: 10.1183/1025448x.00040002.

Øksendal BK (2003) *Stochastic differential equations : an introduction with applications / Bernt Øksendal.* Springer, Berlin; London, 6th ed.