






## ORIGINAL ARTICLE

## Experimental Models of Allergic Disease

WILEY

# Personalized prediction of daily eczema severity scores using a mechanistic machine learning model

Guillem Hurault<sup>1</sup>  | Elisa Domínguez-Hüttinger<sup>2</sup>  | Sinéad M. Langan<sup>3</sup>  |  
Hywel C. Williams<sup>4</sup>  | Reiko J. Tanaka<sup>1</sup> 

<sup>1</sup>Department of Bioengineering, Imperial College London, London, UK

<sup>2</sup>Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, México, México

<sup>3</sup>Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

<sup>4</sup>Centre of Evidence Based Dermatology, University of Nottingham, Nottingham, UK

**Correspondence**

Reiko J. Tanaka, Department of Bioengineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.  
Email: r.tanaka@imperial.ac.uk

**Funding information**

British Skin Foundation, Grant/Award Number: 005/R/18; Wellcome Senior Research Fellowship in Clinical Science, Grant/Award Number: (205039/Z/16/Z)

**Abstract**

**Background:** Atopic dermatitis (AD) is a chronic inflammatory skin disease with periods of flares and remission. Designing personalized treatment strategies for AD is challenging, given the apparent unpredictability and large variation in AD symptoms and treatment responses within and across individuals. Better prediction of AD severity over time for individual patients could help to select optimum timing and type of treatment for improving disease control.

**Objective:** We aimed to develop a proof of principle mechanistic machine learning model that predicts the patient-specific evolution of AD severity scores on a daily basis.

**Methods:** We designed a probabilistic predictive model and trained it using Bayesian inference with the longitudinal data from two published clinical studies. The data consisted of daily recordings of AD severity scores and treatments used by 59 and 334 AD children over 6 months and 16 weeks, respectively. Validation of the predictive model was conducted in a forward-chaining setting.

**Results:** Our model was able to predict future severity scores at the individual level and improved chance-level forecast by 60%. Heterogeneous patterns in severity trajectories were captured with patient-specific parameters such as the short-term persistence of AD severity and responsiveness to topical steroids, calcineurin inhibitors and step-up treatment.

**Conclusions:** Our proof of principle model successfully predicted the daily evolution of AD severity scores at an individual level and could inform the design of personalized treatment strategies that can be tested in future studies. Our model-based approach can be applied to other diseases with apparent unpredictability and large variation in symptoms and treatment responses such as asthma.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Clinical & Experimental Allergy* published by John Wiley & Sons Ltd

## 1 | INTRODUCTION

Atopic dermatitis (synonymous with atopic eczema or just eczema;<sup>1</sup> AD) is the most common inflammatory skin disease, and is characterized by inflamed, dry and itchy skin<sup>2</sup> leading to substantial quality of life impairment and significant socio-economic impact.<sup>3</sup> AD typically has a fluctuating course characterized by inflammatory disease flares followed by periods of remission. Treatment with topical corticosteroids or calcineurin inhibitors during disease flares is aimed at controlling symptoms and skin signs, and emollients are typically used to counteract the dry skin associated with AD.

However, successful control of AD symptoms has been challenging as responses to AD treatments vary considerably between patients. Personalized treatment strategies may be more beneficial to individual patients rather than a “one-size-fits-all” approach to therapy.<sup>4,5</sup> A first step towards developing personalized treatment strategies is to better predict the consequences of possible treatments at an individual level, rather than at population level, to deal with the variability across patients.

Prediction of the consequences of treatments at an individual level is challenging also because of dynamic and sudden fluctuations of AD symptoms. It can be difficult to identify reliable treatment responses, especially if a single end-point is considered, since the responses to a treatment can vary each time even for the same patient. Analysing the dynamic responses to the repeated application of treatment can help identify consistent treatment effects for each patient<sup>6</sup> and ultimately predict whether the chosen treatment is effective and whether the disease is adequately controlled at an individual level.

Machine learning has been successfully applied for prediction tasks. However, typical machine learning models such as artificial neural networks are often black-boxes, lacking interpretability or relying on *post hoc* explanations that are not guaranteed to match the algorithm's true decision process.<sup>7,8</sup> “Black-box models” may fail to be accepted by the medical community and AD patients. Existing

regulations such as the European Union general data protection regulation also highlight the “pressing importance of human interpretability in algorithm design.”<sup>9</sup>

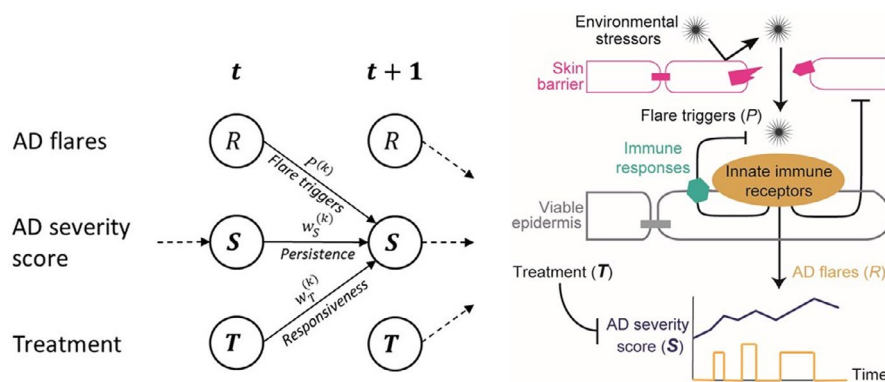
Here, we aimed to develop a biologically interpretable mechanistic machine learning model that can predict daily evolution of AD severity scores at an individual level. We applied a model-based machine learning approach,<sup>10</sup> which allowed us to develop Bayesian machine learning models that can be tailored to the particular context of a given study and the available dataset, and include biologically interpretable mechanistic knowledge. Bayesian machine learning approach has already been applied to a birth cohort data on allergic sensitization to uncover latent atopy classes<sup>11</sup> or to estimate asthma misclassification and risk factors in yearly questionnaire data.<sup>12</sup> However, it has not been applied to predict *daily* changes in disease outcome or in the field of AD.

We hypothesized that it is possible to decipher the apparent unpredictable dynamics of AD severity scores from each patient's data. We previously published a mechanistic model of AD pathogenesis which provided a coherent mechanistic explanation of the dynamic onset, progression and prevention of AD, as a result of interactions between skin barrier, immune responses and environmental stressors.<sup>13,14</sup> Our aim was therefore to adapt the structure of the published mechanistic model to real patient data (Figure S1), and to develop a mechanistic Bayesian model tailored to each individual that can predict the next day's AD severity score given their score and treatments used on that day.

## 2 | METHODS

### 2.1 | General approach

Using the longitudinal data from two published clinical studies<sup>15,16</sup> (example raw data shown in Figure S1), we developed and validated a mechanistic Bayesian model that can predict the next day's AD



**FIGURE 1** Mechanistic Bayesian model of atopic dermatitis (AD) severity dynamics. A: A schematic diagram of the probabilistic model. The arrows depict the relationships between state variables included in the model. B: A schematic diagram of the published mechanistic model of AD pathogenesis<sup>13</sup> from which the structure of the proposed model was adopted. Flare triggers ( $P$ ) and AD flares ( $R$ ) are latent variables, and AD severity score ( $S$ ) and treatment applied ( $T$ ) are the measured variables. The variable,  $T$ , corresponds to the daily binary stepping-up variables in the Flares dataset, and to the combination of the binary variables for the use of stepping-up, topical corticosteroids and calcineurin inhibitors in the SWET dataset.

severity score for each patient. Our mechanistic Bayesian model explicitly described within-patients uncertainties in disease outcomes using probability distributions, and between-patient heterogeneity in severity trajectory and treatment responses by patient-dependent parameters.

To develop the model, we firstly defined the underlying processes that could generate the data as a probabilistic graphical model (Figure 1A), which adopted the structure of a previously published mechanistic model of AD pathogenesis<sup>13,14</sup> (Figure 1B). The model was tailored to the context of the clinical studies in which the data were collected. We then trained the model (fitted to the data) using Bayesian inference, that is updating the probability distributions of the unknown (latent) variables and model parameters through Bayes' theorem, and validated the model by assessing its predictive performance in a forward-chaining setting, where the model was trained with the first week's data and tested on the second week's data, then re-trained on the first two weeks' data and tested on the third week's data, and so on (Figure S2). The first dataset was used for model development and internal validation, and the second dataset to test whether a similar predictive performance could be achieved with a different cohort of patients.

## 2.2 | Data

We chose two datasets that included daily recording of symptoms and treatments over a moderately long period (details in Supplementary A).

The first dataset, which we refer to as "Flares dataset", is a part of the data collected in an observational study that aimed to identify the triggers of AD flares for 59 children.<sup>15</sup> The Flares dataset included *daily* categorical "bother" scores over 6 to 9 months, totalling 6536 patient-day observations, graded from 0 ("no bother at all") to 10 ("the most bother you can imagine") as a response to the question "how much bother did your eczema cause today?". 38.8% of the bother score was missing in Flares dataset (Figure S3). The Flares dataset also included daily binary "stepping-up" variables, that is the answers to the question "have you had to step-up your treatment today because your eczema was worse?". What constituted "stepping-up" treatment was defined for each child at the study outset.

The second dataset, which we refer to as "SWET dataset", is a part of the data collected in a randomized controlled trial that evaluated the effects of use of ion-exchange water softeners for AD control (the softened water eczema trial or SWET) for 334 children.<sup>16</sup> The SWET dataset included the individual child's daily categorical bother score over 16 weeks with only 1.9% of the bother score missing (Figure S4) for a total of 35 854 patient-day observations. The SWET dataset additionally contained information on potential risk factors or confounders, such as the presence of filaggrin mutations, white skin type, age (in years), gender and whether the patient slept away from home. It also included details of the treatment used, such as the type of treatment modalities used each day (topical corticosteroids, calcineurin inhibitors and stepping-up

treatment), the estimated average dose used for each type of topical corticosteroids (mild, moderate, potent or very potent) and calcineurin inhibitors (mild or moderate) over the study period, together with the patient's confidence in the estimated average dose ("not at all sure", "not sure", "sure", or "very sure"). We used all the available information in SWET dataset and evaluated the contribution of each factor on daily evolution of the bother score at an individual level.

## 2.3 | Mechanistic Bayesian models

We developed a mechanistic Bayesian model that predicts the AD severity score ( $S_k(t+1)$ ) for the  $k$ -th patient at day  $t+1$ , given two observables, the previous day's score ( $S_k(t)$ ) and the treatment applied ( $T_k(t)$ ) (Figure 1A).

Our model assumed that AD severity ( $S_k(t+1)$ ) is determined by the temporal accumulation of inflammation caused by AD flares ( $R_k(t)$ ), which result from the activation of innate immune receptors by flares triggers ( $P^{(k)}$ ), and is modified by the treatment applied ( $T_k(t)$ ) (Figure 1B). Flare triggers ( $P^{(k)}$ ) and the resulting flares ( $R_k(t)$ ) were modelled as latent variables. They depend on the complex interactions between the skin barrier, immune responses and environmental stressors.  $P^{(k)}$  for the  $k$ -th patient was assumed to be constant for the duration of the data collection.

We first modelled the severity score measurement process by assuming that a continuous latent severity score,  $\hat{S}_k(t) \in [0, 10]$ , is rounded to the nearest integer to derive the discrete severity score reported by patients,  $S_k(t) = \text{Round}(\hat{S}_k(t))$ . We then described the dynamics of  $\hat{S}_k(t)$  by an exponentially modified Gaussian distribution truncated between 0 and 10,  $\hat{S}_k(t+1) \sim N_{[0,10]}(w_S^{(k)} \hat{S}_k(t) + w_T^{(k)} T_k(t) + R_k(t) + b_S, \sigma_S^2)$ . The distribution of  $\hat{S}_k(t+1)$  follows a Gaussian autoregressive process perturbed by exponentially distributed AD flares,  $R_k(t) \sim \text{Exp}(\beta = P^{(k)})$ , which reflects the assumption that flares occur more frequently in the presence of the flare triggers.

The autoregression is characterized by the patient-dependent autocorrelation or persistence of the severity score ( $w_S^{(k)}$ ), patient-dependent responsiveness to treatment ( $w_T^{(k)}$ ), and population-level intercept ( $b_S$ ) and variance ( $\sigma_S^2$ ). The patient-dependent parameters,  $w_S^{(k)}$ ,  $w_T^{(k)}$  and  $P^{(k)}$ , are given the hierarchical priors,  $\text{logit}(w_S^{(k)}) \sim N(\mu_{w_S}, \sigma_{w_S}^2)$ ,  $w_T^{(k)} \sim N(\mu_T, \sigma_T^2)$  and  $P^{(k)} \sim N^+(0, \sigma_P^2)$ , with population mean ( $\mu_{w_S}$ ,  $\mu_T$ ) and dispersion parameters ( $\sigma_{w_S}$ ,  $\sigma_T$ ,  $\sigma_P$ ).

We also developed an extended version of the mechanistic Bayesian model for SWET dataset (details in Supplementary B). The extended model allowed us to analyse the effects of potential risk factors (the presence of filaggrin mutations, age and sleeping away from home) on the severity score, with their respective weighting parameters,  $w_{FLG}^{(k)}$ ,  $w_{Age}^{(k)}$  and  $w_{Home}^{(k)}$ . We also investigated heterogeneity of treatment responsiveness by replacing the term  $w_T^{(k)} T_k(t)$  with  $w_{SU}^{(k)} SU_k(t) + w_{CS}^{(k)} CS_k(t) + w_{CI}^{(k)} CI_k(t)$ , where  $SU_k(t)$ ,  $CS_k(t)$  and  $CI_k(t)$  are binary variables that indicate whether the  $k$ -th patient

stepped-up, applied topical corticosteroids and calcineurin inhibitors, respectively, with their respective weights,  $w_{SU}^{(k)}$ ,  $w_{CS}^{(k)}$  and  $w_{CI}^{(k)}$ . The weights,  $w_{CS}^{(k)}$  and  $w_{CI}^{(k)}$ , include dose-independent effects (intrinsic responsiveness to the treatment  $b_{CS}^{(k)}$  and  $b_{CI}^{(k)}$ ) and dose-dependent effects that are functions of the quantity and the potency of the treatment (Figure S5).

Our model did not require imputation of missing values for  $S_k(t)$ , since the absence of measurements is naturally accepted by the measurement process of  $S_k(t)$  separately modelled from the dynamics of  $\hat{S}_k(t)$ . Imputation of missing values for other covariates is described in Supplementary C. We conducted prior predictive checks to define weakly informative priors (details in Supplementary D).

## 2.4 | Model fitting

Model training was performed using the Hamiltonian Monte Carlo algorithm in the probabilistic programming language Stan.<sup>17</sup> The posterior distribution was sampled by 6 Markov chains for 3000 iterations (including 50% burn-in). Convergence of the chains was monitored by inspecting the trace plots, checking the Gelman-Rubin convergence diagnostic  $\hat{R}$ <sup>18</sup> and computing effective sample sizes. More details of the inference method are provided in Supplementary E.

## 2.5 | Model validation

The predictive performance of the model was assessed in a forward-chaining setting. Model calibration (whether forecast probabilities are accurate) was assessed by an ordinal quadratic scoring rule (ranked probability score, RPS) and local logarithmic scoring rule (log predictive density, lpd). These metrics were plotted against training day (training data size) to produce learning curves. Details on performance metrics used are described in Supplementary F.

We compared our model to four reference models: a uniform forecast,  $S_k(t+1) \sim U(0, 10)$ , where each outcome is assigned with the same probability, a historical forecast where the probability of each outcome is equal to their relative occurrence in the past, a Gaussian random walk,  $S_k(t+1) \sim N(S_k(t), \sigma^2)$ , where the next score is assumed to be around the previous score, and a mixed effect autoregressive model (our model without flares triggers),

$$\hat{S}_k(t+1) \sim N_{[0,10]} \left( w_S^{(k)} \hat{S}_k(t) + w_T^{(k)} T_k(t) + b_S, \sigma_S^2 \right).$$

# 3 | RESULTS

## 3.1 | Model fitting

The model was trained on each of the two datasets, and the convergence was checked. Population-level parameters (parameters shared across patients) were estimated with a good precision and

their 95% credible interval (in which the parameter lies with 95% probability) were narrow compared to their prior, did not include 0 and were similar for the two datasets, suggesting support for the model structure (Table S1). Three main model parameters that describe patient-dependent dynamics of the severity score are the autocorrelation parameter  $w_S^{(k)}$  for the short-term persistence of the AD severity score, the parameter  $w_T^{(k)}$  for the responsiveness to treatment and  $P^{(k)}$  for the amount of flares triggers, of the  $k$ -th patient.  $w_S^{(k)} \rightarrow 1$  or  $w_S^{(k)} \rightarrow 0$  means that the predicted severity is close to or does not depend on the previous day's severity, respectively.  $w_T^{(k)} < 0$  or  $w_T^{(k)} > 0$  implies that the patient is responsive to treatments or the treatment has an adverse effect on the patient, respectively. A larger  $P^{(k)}$  suggests more severe and frequent flares. These estimates greatly varied from one patient to another, confirming their patient dependence (Figures S6 and S7).

Posterior predictive checks demonstrated that the developed model captured diverse patterns of the dynamic trajectories of the severity score, despite the presence of missing values (representative patients' score dynamics in Figure 2). Typical trajectories observed included fluctuations of the severity score with a return to a healthier state (Figure 2A,C) or without (Figure 2B,D).

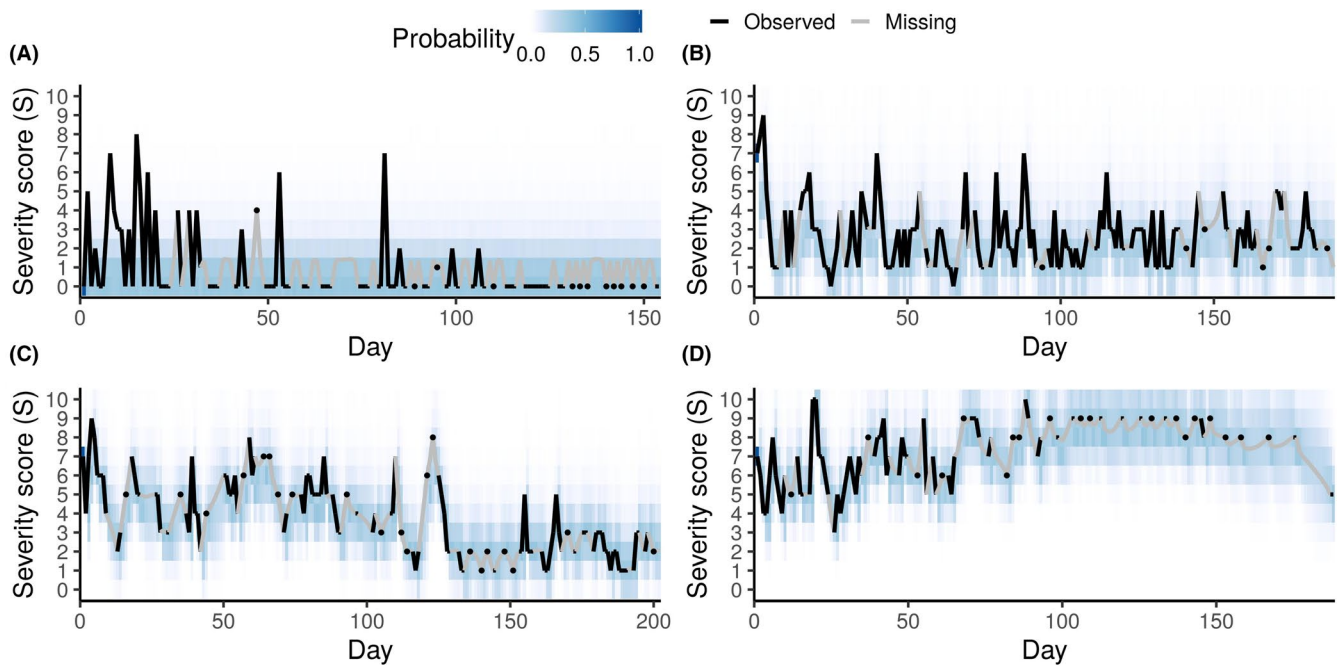
## 3.2 | Model validation

We then validated the model to assess its generalizability beyond the training data. The learning curves demonstrated an improvement in both RPS and lpd, as more data become available (Figure 3), confirming that the model learned the dynamic patterns of the severity scores from the data. Similar or better performance was achieved with the SWET dataset, compared to the Flares dataset, confirming the predictive ability of the model on multiple cohorts. Our model outperformed or performed as well as the four reference models in terms of RPS and lpd for both datasets. Our model demonstrated approximately 60% of improvement in RPS than the chance-level (uniform) forecast for both Flares and SWET datasets (Figure 3). For example, we achieved a lpd of  $\log(0.25)$  with SWET dataset, meaning that the model assigns a 25% probability to the true outcome on average, compared to 9% for a chance-level forecast. Calibration curves (Figure S8) suggested that the predicted probabilities were reasonably calibrated up to 30%-40% in Flares dataset and up to 50%-60% in SWET dataset.

Similar results were obtained for a model we developed using the daily scratch score recorded in the observational study for Flares dataset (Figure S9). The scratch score was not recorded in SWET.

## 3.3 | Effects of treatment modalities and risk factors on the predicted severity scores

The extended model with additional covariates was also successfully fit to SWET dataset (Tables S1 and S2). The posterior predictive checks confirmed that the model could capture diverse patterns of



**FIGURE 2** Fitting of the mechanistic Bayesian model. Posterior predictive distribution of atopic dermatitis (AD) severity score for four representative patients from Flares dataset. (A, C) Bother score returns to a healthier state. (B, D) Bother score does not improve. The plots show the time evolution of the posterior predictive probability mass function as a heatmap. Darker colour represents outcomes with higher probabilities. Black and grey lines show the observed scores and the posterior mean estimate for the missing scores, respectively.

the severity score trajectories, such as large and rapid fluctuations (Figure 4A), large but slow fluctuations (Figure 4B), and controlled AD (Figure 4C). The model could not predict previously unseen patterns, such as transitions of the score from 1 to 10 in a day (Figure 4D at around 70 days), as the model learned the dynamic patterns from past data.

Analysis of the model parameters suggested that older age, absence of filaggrin gene mutations and sleeping at home were associated with *greater* improvement (decrease) in severity scores at the 95% credible level (Figure 5A), as the 95% credible interval of the relevant parameters did not contain 0 and by  $w_{\text{Age}} < 0$  (older age decreases the severity score),  $w_{\text{FLG}} > 0$  (the presence of filaggrin mutations increases the severity score) and  $w_{\text{Home}} < 0$  (sleeping at home decreases the severity score). The estimated effects may appear small in absolute terms, compared to the range of the bother score (0-10), but their effects on the severity score may become practically significant as they accumulate over time. White skin type and sex were not found to be associated with changes in the severity score at the 95% credible level (Figure 5A; 95% credible interval of  $w_{\text{Sex}}$  and  $w_{\text{White}}$  in both sides of 0, suggesting that their effects on the severity score could be both negative and positive).

Further analysis of the parameters,  $w_{\text{SU}}^{(k)}$ ,  $b_{\text{CS}}^{(k)}$  and  $b_{\text{CI}}^{(k)}$ , which describe the dose-independent effects of the treatment on the severity score, demonstrated that none of the treatments appear to have a significant effect at the population level (grey shaded areas in Figure 5B spans from negative to positive values). However, the treatments could have a significant effect at a patient-level. For example, the parameter estimates for one of the patients (orange

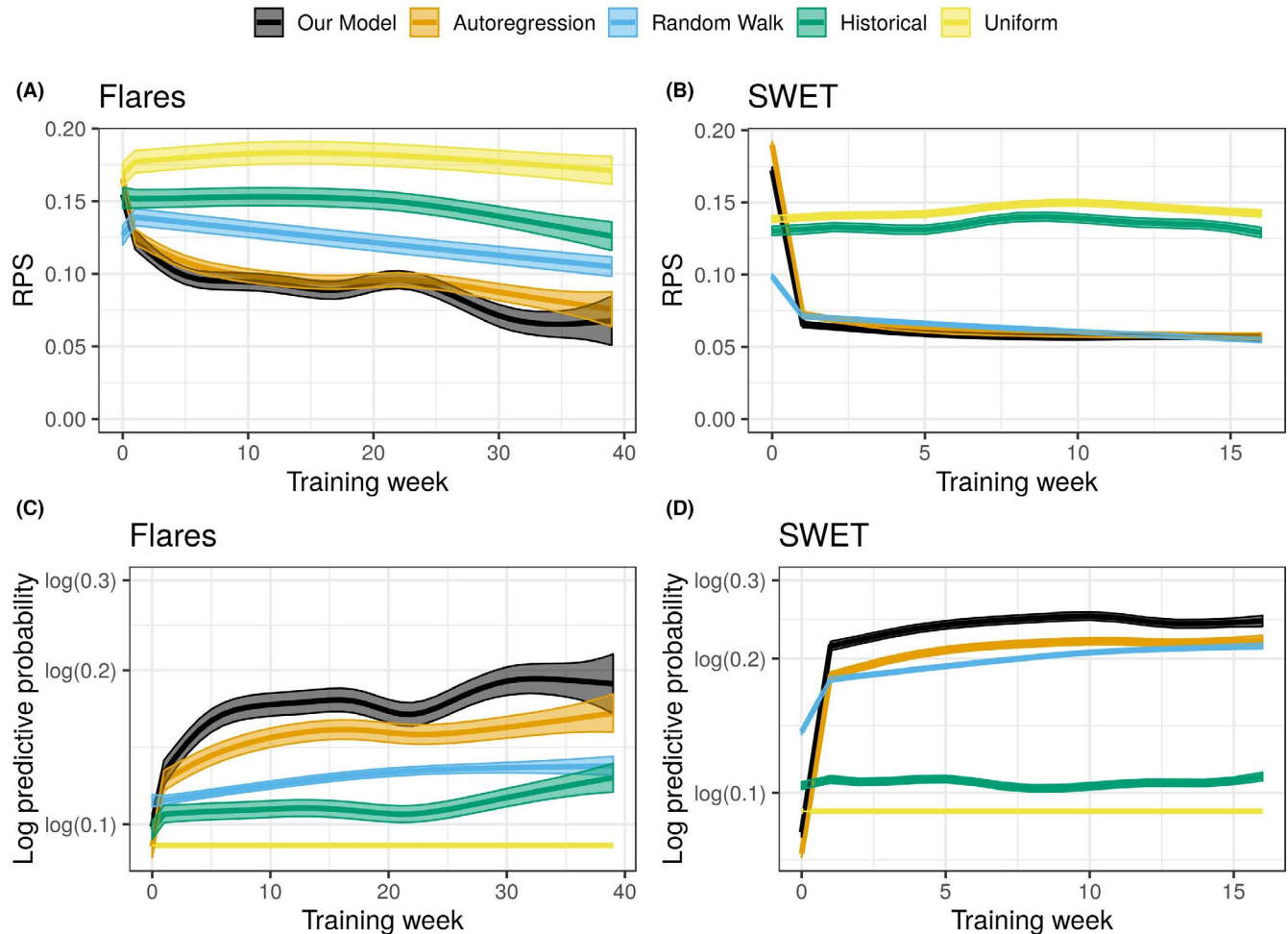
shaded areas in Figure 5B) suggest that the use of corticosteroids has a significant and consistent effect on the severity score for this patient at the 95% credible level. That is, the posterior probability for  $b_{\text{CS}}^{(k)}$  (the dose-independent responsiveness to corticosteroids) being negative (ie the use of corticosteroids reduces the severity score) is greater than 95%. Interestingly, this 95% criterion for the consistent treatment effect was not met for calcineurin inhibitors ( $b_{\text{CI}}^{(k)}$ ) and step-up ( $w_{\text{SU}}^{(k)}$ ) for the same patient. Following this criterion, we confirmed significant effect of corticosteroids in 90 individuals (out of 295 who used corticosteroids) and of step-up in 25 individuals (out of 284 who used step-up). However, we did not find evidence of an intrinsic responsiveness in any of the 92 patients who used calcineurin inhibitors, although 6 of them show a significant dose-dependent responsiveness.

## 4 | DISCUSSION

### 4.1 | Main findings

This study demonstrated a proof-of-concept that predicting the evolution of eczema severity is possible. We developed a novel mechanistic Bayesian machine learning model that can predict patient-specific daily evolution of the AD bother score. The model is biologically interpretable and describes the mechanistic assumption that the AD severity is a result of temporal accumulation of flares (Figure 1). The model learned rich, heterogeneous and dynamic patterns in the daily evolution of AD severity scores that may otherwise





**FIGURE 3** Comparison of predictive performance between the mechanistic Bayesian model (Our Model) and four reference (Uniform, Historical, Random Walk and Autoregression) models. The performance is evaluated for one-day-ahead predictions and plotted as a function of the training week. Confidence bounds correspond to  $\pm$  SE. A-B: Evolution of the ranked probability score (RPS, lower the better) for the Flares dataset (A) and the SWET dataset (B). C-D: Evolution of the log predictive probability (lpd, higher the better) for the Flares dataset (C) and the SWET dataset (D).

appear random and noisy (Figures 2 and 4). Our method extracted information on whether the chosen treatment is effective (responsiveness to treatment), and whether the AD score is persistent and susceptible to flares, at an individual level (Figure 5, Figures S6 and S7). The use of longitudinal data enabled us to look for consistent treatment responses within each patient, rather than a population average response evaluated at a single time-point. We estimated population-level risk factors associated with slower improvement of the severity score, such as the presence of a filaggrin mutation and younger age (Figure 5A). The model was validated using the data from two published clinical studies to confirm its generalizability and the possibility to learn and predict the short-term dynamics of AD severity scores from each patient's data (Figure 3).

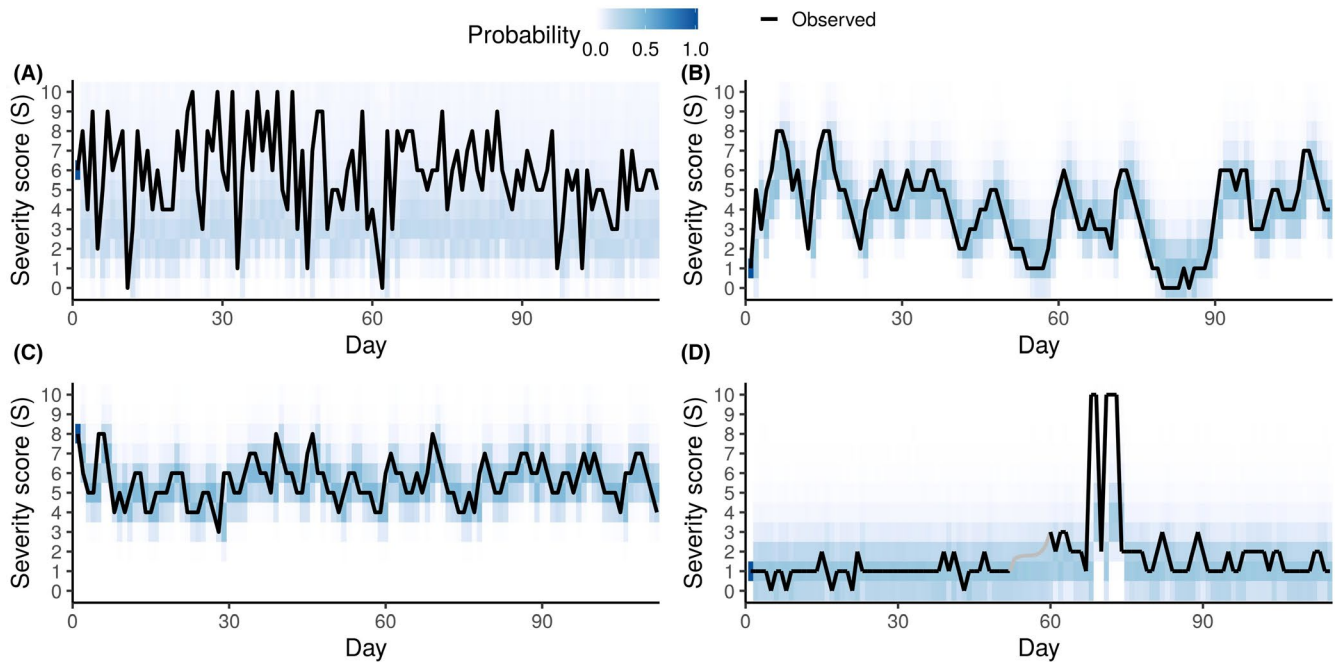
## 4.2 | Strengths of our approach

Our Bayesian approach could be useful to make predictions for new patients, outside of the two cohorts we considered. For instance, we could use the population posterior distributions of the

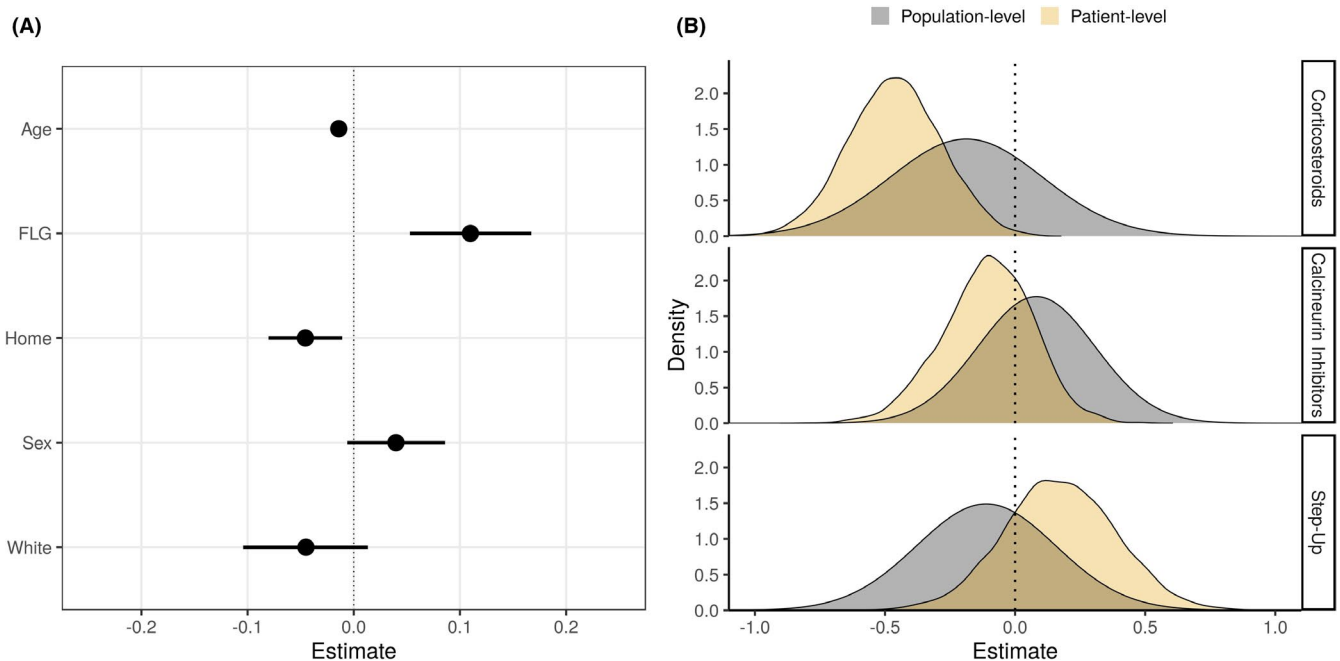
patient-dependent parameters obtained in this study as priors for new patients. The priors will then be updated as more data become available, in order to make personalized and more accurate predictions.

In addition, our model-based Bayesian approach is appropriate to develop models for clinical use, especially when the data are not as controlled as in a clinical trial. Our model explicitly describes uncertainties in disease outcomes (the severity scores) using probability distributions rather than point estimates, as well as uncertainties in the measurements. This enabled us to deal with the missing data (about 40% of scores were missing in the Flares dataset) naturally by simply assuming that the measurement process of the observed score was absent when the score is missing, while still being able to infer the dynamics of the latent severity score from the available data. This method is particularly appropriate for incomplete and partially missing data, for example when patients miss clinical visits.

The model-based approach allows us to design models by taking prior clinical and mechanistic knowledge into account, and by tailoring them to available data and study context. For example, our model was extended by incorporating the additional information (on



**FIGURE 4** Fitting of the extended model. Posterior predictive distribution of atopic dermatitis (AD) severity score for four representative patients from SWET dataset: (A) large and rapid fluctuations, (B) large but slow fluctuations, (C) controlled AD and (D) controlled but with transitions of the score from 1 to 10 in a day (at around 70 days). The plots show the time evolution of the posterior predictive probability mass function as a heatmap. Darker colour represents outcomes with higher probabilities. Black and grey lines show the observed scores and the posterior mean estimate for the missing scores, respectively.



**FIGURE 5** Estimated effects of potential risk factors and responsiveness to treatments on the severity score. A: Population-level estimates of the parameters ( $w_{Age}^{(k)}$ ,  $w_{FLG}^{(k)}$ ,  $w_{Home}^{(k)}$ ,  $w_{Sex}^{(k)}$ ,  $w_{White}^{(k)}$ ) for potential risk factors (age, presence of filaggrin mutation, sleeping at home, sex and white skin). The values represent the contribution of the relevant factor to the severity score. Negative and positive values represent a decrease and an increase in severity score (improvement and worsening), respectively, while null values suggest an absence of an effect. Black circles and the line segments represent the mean posterior and the 95% credible interval, respectively. B: Estimated distribution of the parameters for dose-independent responsiveness to different treatment modalities ( $b_{CS}^{(k)}$ ,  $b_{CI}^{(k)}$ ,  $w_{SU}^{(k)}$  for corticosteroid, calcineurin inhibitors and step-up) at a population level (grey) and for a specific patient (orange).

potential risk factors and the treatment doses) available in SWET dataset but not in Flares dataset. Similarly, our model could be expanded to include additional predictors such as environmental triggers (eg air pollution, weather), host factors (eg compliance to daily bathing, allergies) or biological markers.

These features entail that the developed model cannot be made readily available as a “plug-in” formula, as it is described by a set of context-dependent equations on probability distributions and patient-specific parameters that need to be updated to provide personalized predictions.

### 4.3 | Limitations of the study and future directions

The datasets we used in this study contained daily measurement of the bother score, a subjective global measure of distress caused by AD that has previously been used as a reference for developing asthma severity instruments<sup>19</sup> and validating AD symptom measures such as POEM.<sup>20</sup> While using objective and quantitative measurements would be preferable, this study can serve as a proof-of-concept that predicting the evolution of eczema severity is possible. When collecting daily measurements of objective severity scores becomes less challenging, similar models could be developed to predict scores such as EASI,<sup>21</sup> (o)SCORAD<sup>22</sup> or their self-assessed versions. It will allow us to evaluate the dynamics of scores that capture different aspects of AD symptoms and to compare the predictive performance for different scores. It is also possible to investigate longer time horizon with weekly (instead of daily) measurements. Appropriate evaluation of the effects of data frequency on score dynamics prediction will help designing more effective and informative clinical trials towards personalized medicine.

The predictive capabilities of the model could be potentially improved by incorporating more data, or by using better-quality data, that is with fewer missing values or more precise information about treatments. For example, our model assumes that the same quantity of treatment was applied every day, when treatment was used. This assumption might not always hold in reality and could result in a difficulty with estimating the dose-dependent responsiveness to treatments (Table S2). The daily record of the quantity of treatment applied could resolve this issue and lead to a better estimate of treatment responsiveness.

The model proposed in this paper adopted a structure that was tailored to the available datasets. The model structure was much simpler than that of the previously published mechanistic model of AD pathogenesis.<sup>13,14</sup> If the longitudinal measurement for interactions between environmental stressors, the skin barrier and immune responses becomes feasible in future, such data can be incorporated to develop a more detailed mechanistic machine learning model that provides deeper biological interpretation.

The model-based machine learning approach demonstrated here is applicable to help quantify patient responses to treatment, and may be suitable as a computational method for therapeutic stratification by identifying treatment responses for each individual.<sup>23</sup>

The prediction of daily evolution of severity scores could be further used to suggest optimal treatment strategies for individual patients, using reinforcement learning for example, in addition to conventional computational methods using optimal control theory and bifurcation analysis.<sup>24</sup> Our method could be tested further as part of an intervention using a personalized approach in a future pragmatic randomized controlled trial and compared with conventional standard approaches.

### ACKNOWLEDGEMENTS

We thank Professor Kim S Thomas for sharing SWET dataset and constructive comments on the manuscript. The SWET trial was funded by the NIHR Health Technology Assessment Programme.

### DATA AVAILABILITY STATEMENT

All the codes are available at <https://github.com/ghurault/mbml-eczema>.

### ORCID

Guillem Hurault  <https://orcid.org/0000-0002-1052-3564>

Elisa Domínguez-Hüttinger  <https://orcid.org/0000-0002-9086-099X>

Sinéad M. Langan  <https://orcid.org/0000-0002-7022-7441>

Hywel C. Williams  <https://orcid.org/0000-0002-5646-3093>

Reiko J. Tanaka  <https://orcid.org/0000-0002-0769-9382>

### REFERENCES

1. Johansson SGO, Bieber T, Dahl R, et al. Revised nomenclature for allergy for global use: Report of the Nomenclature Review Committee of the World Allergy Organization, October 2003. *J Allergy Clin Immunol*. 2004;113(5):832-836.
2. Weidinger S, Novak N. Atopic dermatitis. *Lancet*. 2016;387(10023):1109-1122.
3. Drucker AM, Wang AR, Li WQ, Sevetson E, Block JK, Qureshi AA. The burden of atopic dermatitis: summary of a report for the national eczema association. *J Invest Dermatol*. 2017;137(1):26-30.
4. Bieber T, D'Erme AM, Akdis CA, et al. Clinical phenotypes and endophenotypes of atopic dermatitis: Where are we, and where should we go? *J Allergy Clin Immunol*. 2017;139(4):S58-S64.
5. Galli SJ. Toward precision medicine and health: Opportunities and challenges in allergic diseases. *J Allergy Clin Immunol*. 2016;137(5):1289-1300.
6. Senn S. Statistical pitfalls of personalized medicine. *Nature*. 2018;563(7733):619-621.
7. Lipton ZC. The Mythos of Model Interpretability. *ACM Queue*. 2018;16(3):30-57.
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215.
9. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Mag*. 2017;38(3):50.
10. Bishop CM. Model-based machine learning. *Philos Trans R Soc A Math Phys Eng Sci*. 2013;371(1984):20120222.
11. Simpson A, Tan VYF, Winn J, et al. Beyond atopy. *Am J Respir Crit Care Med*. 2010;181(11):1200-1206.
12. Zhang Y, Berhane K. Bayesian mixed hidden markov models: a multi-level approach to modeling categorical outcomes with differential misclassification. *Stat Med*. 2014;33(8):1395-1408.



13. Domínguez-Hüttinger E, Christodoulides P, Miyauchi K, et al. Mathematical modeling of atopic dermatitis reveals “double-switch” mechanisms underlying 4 common disease phenotypes. *J Allergy Clin Immunol*. 2017;139(6):1861-1872.e7.
14. Christodoulides P, Hirata Y, Domínguez-Hüttinger E, et al. Computational design of treatment strategies for proactive therapy on atopic dermatitis using optimal control theory. *Philos Trans A Math Phys Eng Sci*. 2017;375(2096):20160285.
15. Langan SM, Silcocks P, Williams HC. What causes flares of eczema in children? *Br J Dermatol*. 2009;161(3):640-646.
16. Thomas KS, Dean T, O'Leary C, et al. A randomised controlled trial of ion-exchange water softeners for the treatment of eczema in children. *PLoS Med*. 2011;8(2):e1000395.
17. Carpenter B, Gelman A, Hoffman MD, et al. Stan : a probabilistic programming language. *J Stat Softw*. 2017;76(1):1-32.
18. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-472.
19. Steen N, Hutchinson A, Mccoll E, et al. Development of a symptom based outcome measure for asthma. *BMJ*. 1994;309(6961):1065.
20. Charman CR, Venn AJ, Williams HC. The patient-oriented eczema measure development and initial validation of a new tool for measuring atopic eczema severity from the patients' perspective. *Arch Dermatol*. 2014;140:1513-1519.
21. Tofte S, Graeber M, Cherill R, Omoto M, Thurston M, Hanifin JM. Eczema area and severity index (EASI): A new tool to evaluate atopic dermatitis. *J Eur Acad Dermatology Venereol*. 1998;11:S197.
22. Stalder JF, Taïeb A, Atherton DJ, et al. Severity scoring of atopic dermatitis: The SCORAD index: Consensus report of the european task force on atopic dermatitis. *Dermatology*. 1993;186(1):23-31.
23. Eyerich K, Brown SJ, Perez White BE, et al. Human and computational models of atopic dermatitis: A review and perspectives by an expert panel of the International Eczema Council. *J Allergy Clin Immunol*. 2019;143(1):36-45.
24. Tanaka G, Domínguez-Hüttinger E, Christodoulides P, Aihara K, Tanaka RJ. Bifurcation analysis of a mathematical model of atopic dermatitis to determine patient-specific effects of treatments on dynamic phenotypes. *J Theor Biol*. 2018;448:66-79.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Hurault G, Domínguez-Hüttinger E, Langan SM, Williams HC, Tanaka RJ. Personalized prediction of daily eczema severity scores using a mechanistic machine learning model. *Clin Exp Allergy*. 2020;00:1-9. <https://doi.org/10.1111/cea.13717>