SUPPLEMENTARY MATERIAL

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

Guillem Hurault, Elisa Domínguez-Hüttinger, Sinéad M. Langan, Hywel C. Williams and Reiko J. Tanaka

A. Description of the extended model

We extended our mechanistic Bayesian model to take into account information present in SWET dataset that were not available in Flares dataset, notably to investigate the effects of potential risk factors on severity scores and heterogeneity in treatment responsiveness.

The model is described by an exponentially modified Gaussian distribution,

$$\boldsymbol{S}_k(t+1) \sim N\left(\hat{S}_k(t+1), \sigma_S^2\right), \text{ with}$$

$$\hat{S}_k(t+1) = w_S^{(k)} \boldsymbol{S}_k(t) + \hat{T}_k(t) + w_{Home} \boldsymbol{Home}^{(k)}(t) + \widehat{Dem_k} + R_k(t) + b_S,$$

where $\hat{T}_k(t)$, $\boldsymbol{Home}^{(k)}(t)$ and $\widehat{Dem_k}$ represent the contribution of treatment, whether the patient "slept at home" and the contribution of demographics factors, respectively.

Demographics factors include the presence of filaggrin mutation ($\boldsymbol{FLG}^{(k)}$), sex ($\boldsymbol{Sex}^{(k)}$), age ($\boldsymbol{Age}^{(k)}$), and "white" ethnicity ($\boldsymbol{White}^{(k)}$):

$$\widehat{Dem_k} = w_{FLG} \boldsymbol{FLG}^{(k)} + w_{Sex} \boldsymbol{Sex}^{(k)} + w_{Age} \boldsymbol{Age}^{(k)} + w_{White} \boldsymbol{White}^{(k)}.$$

The contribution of treatment, $\hat{T}_k(t)$, is summarised in Figure S5 and is modelled by a linear combination of the treatment usage,

$$\hat{T}_k(t) = w_{SU}^{(k)} \boldsymbol{SU}_k(t) + w_{CS}^{(k)} \boldsymbol{CS}(t) + w_{CI}^{(k)} \boldsymbol{CI}_k(t),$$

for step-up ($\boldsymbol{SU}_k(t)$), topical steroid ($\boldsymbol{CS}_k(t)$) and calcineurin inhibitors ($\boldsymbol{CI}_k(t)$). We assumed a hierarchical prior for $w_{SU}^{(k)} \sim N(\mu_{SU}, \sigma_{SU}^2)$, and expressed $w_{CS}^{(k)}$ and $w_{CI}^{(k)}$ as a function of the daily quantity of treatment of different potencies used:

$$w_T^{(k)} = \sum_P w_{T,P}\, \hat{q}_{T,P}^{(k)} + b_T^{(k)}, \quad T \in \{CS, CI\} \text{ and } P \in \{\text{Mild, Moderate, Potent, Very Potent}\},$$

where

- $\hat{q}_{T,P}^{(k)}$ is the estimated daily quantity of treatment T of potency P used,
- $w_{T,P}$ is the relative contribution of treatment T of potency P on the severity score, and
- $b_T^{(k)}$ is the intrinsic responsiveness of the k-th patient to treatment T. We assumed a hierarchical prior for $b_T^{(k)} \sim N(\mu_T, \sigma_T^2)$.

The daily quantity of treatment used by the k-th patient, $\hat{q}_{T,P}^{(k)}$, is estimated from the reported total quantity of treatment used $\boldsymbol{Q}_{T,P}^{(k)}$. If $\boldsymbol{Q}_{T,P}^{(k)} = 0$, we assume $\hat{q}_{T,P}^{(k)} = \boldsymbol{Q}_{T,P}^{(k)} = 0$. Otherwise, we estimate $\hat{q}_{T,P}^{(k)}$ by $\hat{q}_{T,P}^{(k)} = \frac{\hat{Q}_{T,P}^{(k)}}{N_{T,P}^{(k)}}$, where $\hat{Q}_{T,P}^{(k)}$ is the total quantity of treatment used and $\boldsymbol{N}_{T,P}^{(k)}$ is the number of treatment applications estimated by a multiplicative error model, $\log(\hat{Q}_{T,P}^{(k)}) \sim N\left(\log(\boldsymbol{Q}_{T,P}^{(k)}), \frac{\sigma_Q^2}{\boldsymbol{Conf}^{(k)}}\right)$, where $\boldsymbol{Q}_{T,P}^{(k)}$ is the quantity reported by the k-th patient with confidence $\boldsymbol{Conf}^{(k)} \in \{1 = \text{not all sure}, 2 = \text{not sure}, 3 = \text{sure}, 4 = \text{very sure}\}$.

B. Missing value imputation

As is often the case with real-world data, especially clinical data, Flares and SWET datasets contains missing values in the bother score (38.8% and 1.9%, respectively). Ignoring missing values, such as removing them entirely, could result in a dramatic reduction of the available data and information, especially when dealing with time-series where the observations are related. Appropriate imputation of missing values is important to avoid making biased prediction. Simple imputation methods, such as imputation by a default value or the mean of the observations, are often implemented. However, they do not fit well in a Bayesian framework where uncertainties in any unknowns are modelled.

In this study, we treated missing values as unknown (parameters) to be imputed by the Bayesian model in a semi-supervised setting. We constructed a vector S_k representing the time-series of the *k*-th patient by concatenating both data (observed score) and parameters (missing score). Therefore, priors for missing values follow the same distribution as the likelihood when values are observed. Here we assumed the missing scores were missing completely at random.

We did not let the model impute the missing values for the other covariates (treatment and risk factors for the extended model) to avoid reverse causality. For instance, if we let the model impute the missing value for treatment data, it could determine the value *a posteriori* based on the knowledge that the severity is decreasing. Instead, we made a conservative assumption to replace missing values for $\boldsymbol{T}_k(t)$ or $\boldsymbol{CS}_k(t)$, $\boldsymbol{CI}_k(t)$, $\boldsymbol{SU}_k(t)$ with 0 (no use of treatment). As a result, the effects of treatment on future severity scores are more likely to be underestimated than overestimated. Similarly, missing $\boldsymbol{Conf}^{(k)}$ (2/327 patients) were imputed by "not at all sure", missing $\boldsymbol{FLG}^{(k)}$ (22/327 patients) were imputed by 0 (absence of mutation), missing $\boldsymbol{White}^{(k)}$ (2/327 patients) were imputed by 0 (non-white or do not wish to declare) and missing $\boldsymbol{Home}^{(k)}(t)$ were imputed by 1 ("slept at home", the most common answer).

We assumed $\boldsymbol{T}_k(t) = 0$ during testing (where t represents the future and is therefore missing) to avoid making biased predictions based on the knowledge about the future use of treatment.

C. Choice of priors

Priors are inherent to a Bayesian model, as they correspond to the initial distributions that are updated upon observations of data to form the posterior distribution.

We chose our priors to be weakly informative. Weakly informative priors are priors designed to rule out unreasonable parameter values (e.g. noise parameters outside the range of the scores) without excluding any value that could make sense. The influence of weakly informative priors is expected to disappear with enough data. We confirmed that our priors were reasonable by conducting prior predictive checks and that our results were not sensitive to the choice of realistic priors.

We set hierarchical (population) priors for patient-dependent treatment parameters by $w_T^{(k)} \sim N(\mu_T, \sigma_T^2)$, $w_{SU}^{(k)} \sim N(\mu_{SU}, \sigma_{SU}^2)$, $b_{CS}^{(k)} \sim N(\mu_{CS}, \sigma_{CS}^2)$ and $b_{CI}^{(k)} \sim N(\mu_{CI}, \sigma_{CI}^2)$, where μ and σ^2 are the mean and variance of the Gaussian distribution, respectively. We assumed a uniform prior for $w_S^{(k)} \sim U(0,1)$. We did not define a hierarchical prior for $w_S^{(k)}$ for computational reasons as several parametrisations (e.g. Beta distribution, non-centred normal distribution with a logit link) did not converge after two days of sampling with a high-performance computing cluster. Using a hierarchical or a uniform prior should not result in a noticeable difference in the posterior for $w_S^{(k)}$ since we can

expect the influence of the prior to be minimal for long enough time-series (more than 100 days long) used in this study.

We chose $b_S \sim T_2(0,1)$, $\mu_T \sim T_2(0,0.5)$, $\sigma_T \sim T_2^+(0,0.25)$ and $\sigma_P \sim T_2^+(0,\ 0.2)$, where $T_\nu(\mu,\sigma)$ represents the Student's t-distribution with ν degrees of freedom, centered around μ with scale σ, and the superscript $^+$ denotes half (truncated) distributions defined on positive reals.

We also chose $\sigma_S \sim Gamma(2.6, 3.1)$ to ensure that 98% of the mass is between 0.2 and 2.5. Indeed, $\sigma_S < 0.2$ and $\sigma_S > 2.5$ are unlikely due to discretisation (e.g. if 3.4 is discretised as 3, the error is 0.4) and the score range (0-10), respectively.

For the extended model, we chose $w_{FLG}, w_{Sex}, w_{White}, w_{Home} \sim T_2(0,0.5)$, $w_{Age} \sim T_2(0,0.1)$, $\mu_{SU}, \mu_{CS}, \mu_{CI} \sim T_2(0,0.5)$, $\sigma_{SU}, \sigma_{CS}, \sigma_{CI} \sim T_2^+(0,0.25)$, $\sigma_Q \sim N^+(0,0.25^2)$, $w_{CS,\text{ Mild}}, w_{CS,\text{ Moderate}}, w_{CS,\text{ Potent}}, w_{CS,\text{ Very Potent}} \sim N(0,0.5^2)$ and $w_{CI,\text{ Mild}}, w_{CI,\text{ Moderate}} \sim N(0,0.5^2)$.

D. Inference method

This section provides some details of the inference method used in this paper.

A Bayesian model is fully determined by its posterior distribution $p(\theta|x)$, where θ represents the model parameters (including missing values) and x the observed variables ($\boldsymbol{S}_k(t)$ and $\boldsymbol{T}_k(t)$).

In most cases, including this study, it is not possible to compute $p(\theta|x)$ analytically using Bayes' theorem. Instead, Markov chain Monte-Carlo (MCMC) methods, like the Hamiltonian Monte-Carlo algorithm used in this paper, aim to sample the posterior distribution using Markov chains whose stationary distribution is $p(\theta|x)$. The first samples from the chains are discarded (burn-in or warm-up) to limit the influence of initial conditions.

When we use MCMC methods, it is not possible to know whether a Markov chain has converged. However, we can look for signs of a lack of convergence. Not observing any issues (such as numerical errors, chains sampling different areas in the parameter space or not sampling the parameter space at all) while running Markov chains for a long time is a good indicator for potential convergence. Inspection of trace plots (time series plots of MCMC draws) can assess the mixing of the chains and whether the chains explore the posterior distribution well. It is also possible to run multiple chains in parallel from different initial values and check whether these chains sample the same distribution with a Gelman-Rubin convergence diagnostic $\hat{R}$ smaller than 1.1 (Tables S1 and S2).

The precision of parameter estimates can be assessed with the effective sample size n_{eff} (also shown in Tables S1 and S2). Given a parameter θ whose distributions is estimated by N independent and identically distributed samples, the sample mean $\bar{\theta}$ is given by $\bar{\theta} \sim N\left(\mu_\theta, \left(\frac{\sigma_\theta}{\sqrt{n}}\right)^2\right)$, according to the central limit theorem, where μ_θ and σ_θ are the true mean and the standard deviation of θ, respectively. Therefore, the resolution of $\bar{\theta}$ is proportional to $\frac{1}{\sqrt{n}}$, suggesting that 100 times more samples are required if one more digit of precision is needed for μ_θ. In MCMC, the samples are autocorrelated, and thus the error is proportional to $\frac{1}{\sqrt{n_{\text{eff}}}}$, rather than $\frac{1}{\sqrt{n}}$, where n_{eff} is the effective sample size.

E. Multi-category calibration

We applied calibration during the validation process to correct for potential mismatch between predicted probabilities and observed frequency (e.g. when a 50% forecast does not happen 50% of the time). At each week *W*,

- predictions from the previous week (*W*-1) were compared to the outcomes of week *W* to update the calibration model, and
- the model was trained with the data up to week *W*, and uncalibrated predictions for week *W*+1 were generated and fed to the calibration model to return the calibrated predictions.

Designing a calibration model is relatively straightforward for a binary classification task (with a positive and negative class), by adding more probabilities to the positive outcome while removing the same amount for the negative outcome. However, this study deals with a multi-category forecast with 11 categories (corresponding to the severity score of 0 to 10), and no gold standard for calibrating a multi-category forecast has been proposed.

We therefore proposed a multi-category calibration method that applies binary classification followed by the coupling of the adjusted probabilities, using pairwise "one-against-all" isotonic regressions. Firstly, we decomposed the multi-category problem in "one-against-all" binary problems: "bother $=$ 0 vs bother $\neq$ 0", "bother $=$ 1 vs bother $\neq$ 1", etc. Then we calibrated each of these binary classification problems with isotonic regressions, a non-parametric approach to fit monotonically increasing curves. Finally, we used pairwise coupling to combine the probabilities of the binary forecasts to derive the calibrated multi-category forecast, $\hat{f}_i$, for the *i*-th category (*i*=0, …, 10) by

$$\hat{f}_i = f_i + \alpha\, \Delta_i \cdot [|\Delta_i| > 0.01],$$

where f_i is the non-calibrated multi-category forecast and $\Delta_i = p_i - f_i$ is the adjustment term, which corresponds to the difference (to be adjusted) between the calibrated forecast p_i and f_i in the binary problem "bother $= i$ vs bother $\neq i$". We chose $\alpha = 0.8$ to scale down the adjustment to prevent overfitting and included the Iverson bracket $[|\Delta_i| > 0.01] = \begin{cases} 0 \text{ if } |\Delta_i| \leq 0.01 \\ 1 \text{ otherwise} \end{cases}$ to keep the original forecast if the adjustment was small ($|\Delta_i| \leq 0.01$). The $\hat{f}_i$'s were normalised to sum to 1. The calibration procedure could be applied iteratively, but we found that one pass was enough to calibrate the forecasts effectively.

F. Performance metrics

The accuracy of probabilistic models is assessed by calibration and discrimination. Calibration evaluates whether the probability estimates are accurate (e.g. whether an event with a probability of 50% occurs 50% of the time), while discrimination evaluates the ability to differentiate observations with different outcomes (e.g. whether someone with a disease is assigned a higher probability of having a disease than someone without a disease). In this study, we assessed model calibration (whether forecast probabilities are accurate) by a normalised quadratic scoring rule (ranked probability skill score, RPSS) and discrimination (whether the ranking of probabilities is accurate) by the area under the receiving operating characteristic (AUROC). These metrics are appropriate for evaluating forecasts of categorical ordinal data.

F.1 RPSS for calibration

We qualitatively assessed calibration of our model with calibration curves that plot forecast probabilities against observed frequencies (Fig 3C-D), and also quantified calibration with a scoring rule using RPSS.

The most common quadratic scoring rule is the mean square error, aka the Brier score (*BS*), which is defined as the mean square difference between the predicted probabilities and the actual outcome. A perfectly calibrated classifier has a *BS* of 0 by definition. For a multi-category outcome, the Brier score is an average of $BS(t) = \sum_{i=1}^{R}\left(f_{t,i} - o_{t,i}\right)^2$ at each prediction t over N predictions, and is described by

$$BS = \frac{1}{N}\sum_{t=1}^{N}\sum_{i=1}^{R}\left(f_{t,i} - o_{t,i}\right)^2,$$

where N is the number of predictions, R is the number of categories, $f_{t,i}$ is the forecast probability to predict the i-th category by the t-th prediction, and $o_{t,i} \in \{0,1\}$ describes whether the t-th prediction is in the i-th category.

The Brier score is defined for multi-category forecasts but is not suited well for ordinal outcomes, such as the severity score from 0 to 10 in this study. The ranked probability score (RPS) extends the Brier score to ordinal outcomes by computing the mean square difference between the cumulative forecast distribution $F_t(i) = \sum_{j=1}^{i} f_{t,j}$ and the cumulative outcome distribution $O_t(i) = \sum_{j=1}^{i} o_{t,j}$, and is defined by

$$RPS = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{R-1}\sum_{i=1}^{R}\left(F_t(i) - O_t(i)\right)^2.$$

Given that we can rewrite $RPS = \frac{1}{R-1}\sum_{i=1}^{R} BS(\text{outcome} \le i)$ with $BS(\text{outcome} \le i) = \frac{1}{N}\sum_{t=1}^{N}\left(F_t(i) - O_t(i)\right)^2$, RPS can be seen as the averaged Brier score of the binary classification problems, "$outcome \le i$" versus "$outcome > i$".

To provide a more interpretable metric than the RPS, we used the ranked probability skill score, $RPSS = 1 - \frac{RPS}{RPS_0}$, which quantifies the improvement of accuracy of the probabilistic forecast by the proposed model from that of the baseline forecast, RPS_0. $RPSS = 0$ represents a prediction not better than the baseline forecast, and $RPSS = 1$ is a perfect prediction. We chose the baseline RPS to be the expected RPS for a chance-level forecast, i.e. $RPS_0 \approx 0.182$. A chance-level forecast has the advantage of being constant over time and reflects the situation when no data is available to train a model (e.g. if we are to make predictions the first day a patient enters a study). While naïve, we consider a chance-level forecast to be an appropriate upper bound for the RPS.

F.2 AUROC for discrimination

We assessed discrimination by computing the area under the receiving operating characteristic (AUROC). The AUROC in a binary classification task (negative and positive classes) can be interpreted as the probability that a random positive instance is assigned to a higher probability than a random negative instance. A multi-category AUROC can be derived by averaging the AUROC from a set of "one-vs-all" binary classification problems. We derived an ordinal AUROC as the average AUROC of the binary classification problems "Bother $\le i$" versus "Bother $> i$" for $i \in 0 \ldots 9$.

F.3 Learning curves

To investigate whether the model learns/improves its performance as more data comes in, we plotted the evolution of the RPSS as a function of the training iterations of the forward chaining (Fig 3a). However, the RPSS was not computed on the same population at each iteration due to missing observations. These subpopulations were not representative of the entire population as patients with controlled AD tended to drop out earlier in clinical trials than patients with uncontrolled AD. The fact that the patients with controlled AD are easier to predict (low noise, high RPSS) than patients with uncontrolled AD (high noise, low RPSS) resulted in Simpson's paradox, where the RPSS averaged across all the available patients hit a maximum then decrease, although each individual RPSS may increase (Fig S10).

We therefore controlled for the patient-dependence as well as other factors by modelling the RPSS at the observation-level and used the mean fit as an unbiased estimate of the RPSS averaged across observations. We used a Generative Additive Model (GAM) with cubic splines to achieve a flexible fit to the evolution of the RPSS while avoid overfitting. The performance for the first training iteration ($i = 0$) was estimated separately as it corresponds to a uniform forecast. We also controlled for the prediction horizon, since predictions at each iteration were made for the entire week but the performance was expected to decrease as the prediction horizon increases. In addition, we introduced a mixed effect on the intercept to control for patient-dependence. The model was fitted using the gamm4 package in R to $RPSS \sim [i = 0] + [i > 0]{:}\, t + [i > 0]{:}\, s(i) + (1|Patient)$, where

- [.] is the Iverson bracket: $[A] = 1$ if A is true and 0 otherwise.
- the coefficient for $[i = 0]$ corresponds to the RPSS estimate for $i = 0$,
- $[i > 0]{:}\, t$ represents the interaction between $[i > 0]$ (1 if $i > 0$, 0 otherwise) and the prediction horizon t (the corresponding coefficients measures how much RPSS is lost as t increases),
- $s(i)$ represents a cubic spline on i, which can be written as a linear combination of piecewise cubic polynomial basis function $b_j(i)$ and coefficients β_j, $\beta_1\, b_1(i) + \beta_2\, b_2(i) + \cdots + \beta_l\, b_l(i)$. This term models the evolution of RPSS as more data comes in, and
- $(1|Patient)$ represents a random effect on the intercept for different patients.

The RPSS fit for one-step-ahead prediction is shown in Fig 3a. We estimated the calibration loss to be 0.9% in Flares and 3.5% in SWET when the prediction horizon (t) is increased by one day (e.g. one day forecast versus two days forecasts).

Unlike the RPSS, it is not possible to control for the fact that the AUROC at different iterations may be computed from different subpopulations of patients, since the AUROC cannot be computed at the observation-level. We can nonetheless mitigate Simpson's paradox in the evolution of AUROC by weighting observations with the population size used to derive each AUROC. We implemented a Beta regression with a logit link for the mean and a log link for the sample size, since the AUROC is a probability. We used B-splines with 5 degree of freedoms to model the evolution of the AUROC and controlled for the prediction horizon. The model was fitted for $i > 0$ (since for a uniform forecast, $AUROC = 0.5$) using the betareg and splines packages in R, with $AUROC \sim t + s(i)$.

The AUROC was fitted for prediction one-step-ahead (Fig 3b). We also estimated the odd ratio for the discrimination loss to be 0.96 in Flares dataset and 0.85 in SWET dataset when the prediction horizon (t) is increased by one day (e.g. one day forecast versus two days forecasts, respectively).

G. Comparison of our model with the null model

We compared the predictive performance metrics for model calibration and discrimination between our model and a null model. As the null model, we used a Gaussian random walk model $\boldsymbol{S}_k(t+1) \sim N(\boldsymbol{S}_k(t), \sigma_S^2)$, where σ_S^2 is the variance of the random walk. Note that σ_S^2 here is different from the variance of the Gaussian component of the exponentially modified Gaussian distribution of our model.

The null model had a lower performance overall, although the difference is less striking for the model fitted to SWET dataset than that to Flares dataset. The similar performance between the null model and our model for SWET dataset could be explained by the fact that 58% of the patients in SWET dataset follows a "quasi" random-walk behaviour (characterised by a strong autocorrelation, e.g. $w_S^{(k)} > 0.75$ and mild skewness, e.g. $P(t) < 1$), compared to only 17% of the patients in Flares dataset. It could also be explained by the fact that the relative benefit of using our model, compared to a random walk model, for one-day-ahead forecasts is less pronounced when working with a complete dataset (e.g. SWET dataset) than when working with a dataset with many missing values (e.g. Flares dataset).

However, the null model did not "learn" the dynamic patterns of the severity scores from the data and merely reacted to distributional changes in the data, as indicated by heteroscedasticity of the model (the variance that changes with time) (Fig S12). Therefore, the random walk model is not generalisable to unseen data even though it has only one patient-nonspecific parameter. Our model, in comparison, is statistically valid, interpretable and flexible enough to capture different trajectory patterns, including random walks.

For all these reasons, we believe our model is superior to the Gaussian random walk model.

Table S1: Posterior summary statistics for the population-level parameters of the models trained on Flares and SWET dataset. The potential scale reduction factor $\hat{R}$ indicates whether different MCMC chains are sampling the same distribution, an indicator of convergence. $\hat{R} < 1.1$ suggests no evidence for an absence of convergence. The effective sample size n_{eff} is an estimate of the number of independent draws from the chains. Higher n_{eff} corresponds to more precise estimates (smaller standard error of the mean).

Parameter	Interpretation	Dataset	Mean	95% CI	$\hat{R}$	n_{eff}	SE
σ_S	Standard deviation of the evolution of *S* (severity)	Flares	0.651	[0.616, 0.688]	1.033	199	0.001
		SWET	0.640	[0.631, 0.649]	1.011	1029	0.000
σ_P	Standard deviation of the relative evolution of *P* (flare triggers)	Flares	0.051	[0.042, 0.061]	1.004	999	0.000
		SWET	0.076	[0.068, 0.084]	1.004	1389	0.000
μ_T	Population mean of the responsiveness to treatment	Flares	-0.200	[-0.330, -0.070]	1.001	2904	0.001
		SWET	-0.196	[-0.245, -0.146]	1.001	3510	0.000
σ_T	Population standard deviation of the responsiveness to treatment	Flares	0.373	[0.263, 0.507]	1.002	2077	0.001
		SWET	0.369	[0.321, 0.423]	1.002	2594	0.001
b_S	Intercept of the evolution of *S*	Flares	0.100	[0.049, 0.151]	1.015	468	0.001
		SWET	0.294	[0.267, 0.321]	1.004	2348	0.000

Table S2: Posterior summary statistics for the population-level parameters of the extended model.

Parameter	Interpretation	$\hat{R}$	n_{eff}	Mean	SD	SE	2.5%	50%	97.5%
b_S	Intercept of the evolution of S	1.006	1520	0.396	0.042	0.001	0.312	0.397	0.479
σ_S	Standard deviation of the evolution of S	1.007	599	0.634	0.005	0.000	0.626	0.634	0.643
σ_P	Standard deviation of the relative evolution of P	1.006	411	0.076	0.004	0.000	0.069	0.076	0.084
σ_Q	Standard deviation of the true total quantity of treatment	1.013	708	0.561	0.177	0.007	0.161	0.565	0.896
w_{FLG}	Expected change in the presence of filaggrin mutation	1.002	1202	0.156	0.031	0.001	0.096	0.156	0.218
w_{Sex}	Expected change if the patient is male (compared to female)	1.001	1693	0.020	0.027	0.001	-0.033	0.021	0.072
w_{Age}	Expected change for one extra year of age	1.006	1424	-0.008	0.003	0.000	-0.014	-0.008	-0.002
w_{White}	Expected change when the patient is of white ethnicity versus other/unknown ethnicity	1.004	1451	-0.048	0.034	0.001	-0.113	-0.048	0.020
w_{Home}	Expected change when the patient is sleeping at home	1.003	3106	-0.069	0.018	0.000	-0.105	-0.069	-0.033
μ_{SU}	Population mean of the responsiveness to step-up	1.002	1381	-0.105	0.028	0.001	-0.160	-0.105	-0.050
σ_{SU}	Population standard deviation of the responsiveness to step-up	1.009	871	0.293	0.033	0.001	0.230	0.293	0.360
μ_{CS}	Population mean of the responsiveness to corticosteroids	1.004	909	-0.181	0.034	0.001	-0.247	-0.181	-0.113
σ_{CS}	Population standard deviation of the responsiveness to corticosteroids	1.011	673	0.321	0.029	0.001	0.266	0.321	0.379
$w_{CS,Mild}$	Change in corticosteroids responsiveness due to one daily additional gram of mild corticosteroids	1.005	1598	0.013	0.030	0.001	-0.047	0.013	0.072
$w_{CS,Moderate}$	Change in corticosteroids responsiveness due to one daily additional gram of moderate corticosteroids	1.008	513	0.022	0.023	0.001	-0.022	0.021	0.068
$w_{CS,Potent}$	Change in corticosteroids responsiveness due to one daily additional gram of potent corticosteroids	1.003	943	0.036	0.027	0.001	-0.016	0.035	0.091
$w_{CS,VeryPotent}$	Change in corticosteroids responsiveness due to one daily additional gram of very potent corticosteroids	1.003	1831	-0.041	0.100	0.002	-0.256	-0.036	0.147
μ_{CI}	Population mean of the responsiveness to calcineurin inhibitors	1.005	1733	0.093	0.052	0.001	-0.013	0.093	0.194
σ_{CI}	Population standard deviation of the responsiveness to calcineurin inhibitors	1.027	266	0.147	0.069	0.004	0.026	0.143	0.291
$w_{CI,Mild}$	Change in calcineurin inhibitors responsiveness due to one daily additional gram of mild calcineurin inhibitors	1.004	1508	0.030	0.064	0.002	-0.091	0.027	0.165
$w_{CI,Moderate}$	Change in calcineurin inhibitors responsiveness due to one daily additional gram of mild calcineurin inhibitors	1.010	693	-0.461	0.128	0.005	-0.727	-0.457	-0.223

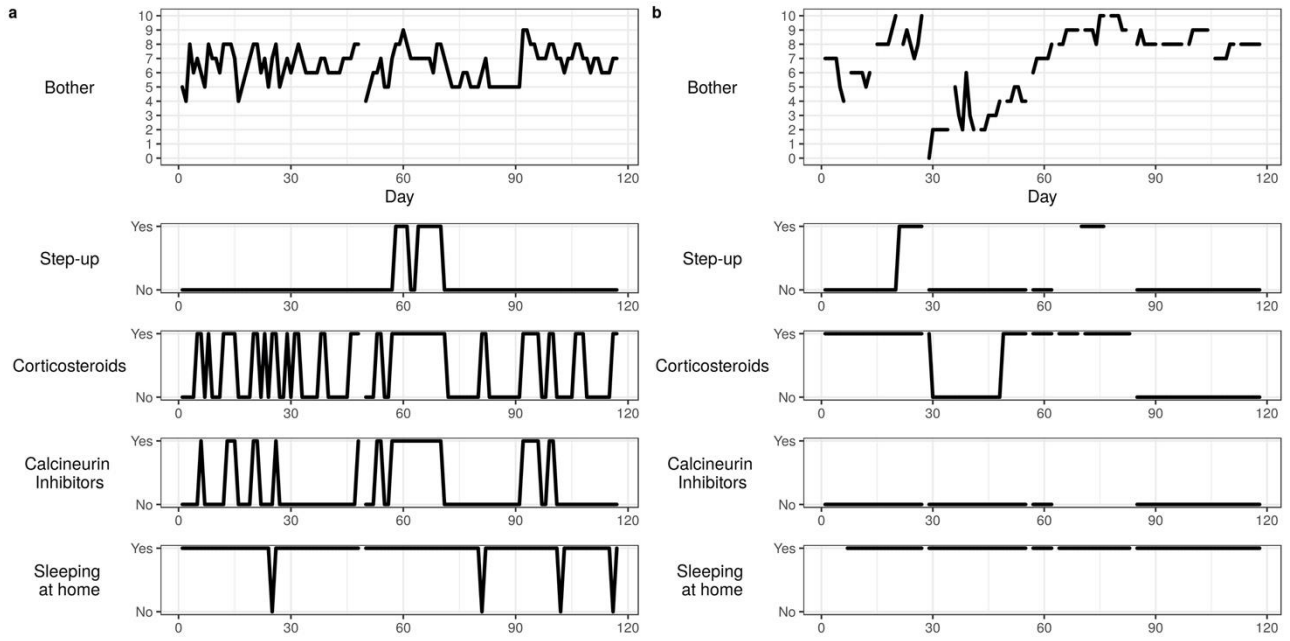


Figure S1: Example data from SWET dataset. Discontinuities represent missing values. Data from Flares dataset is similar but with only bother score and step-up, and with more missing values.

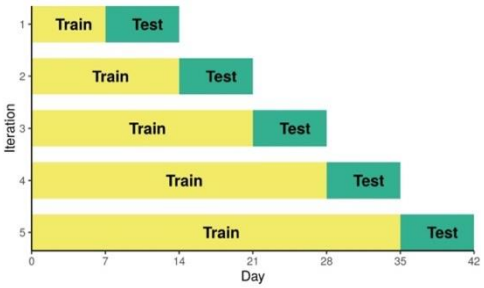


Figure S2: Illustration of the forward chaining validation procedure. The model was trained with the first week's data and tested on the second week's data, then re-trained on the first two weeks' data and tested on the third week's data, and so on, up to week 39 and 16 for Flares and SWET datasets, respectively.

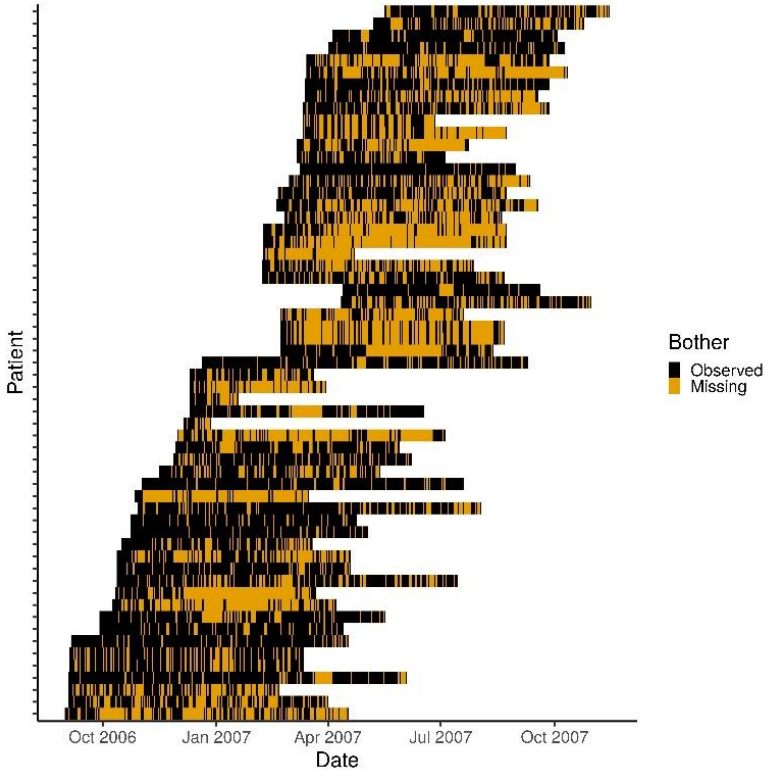


Figure S3: Missing bother scores in Flares dataset. Black and orange indicate observed and missing scores, respectively. The x-axis indicates the date of the measurement.

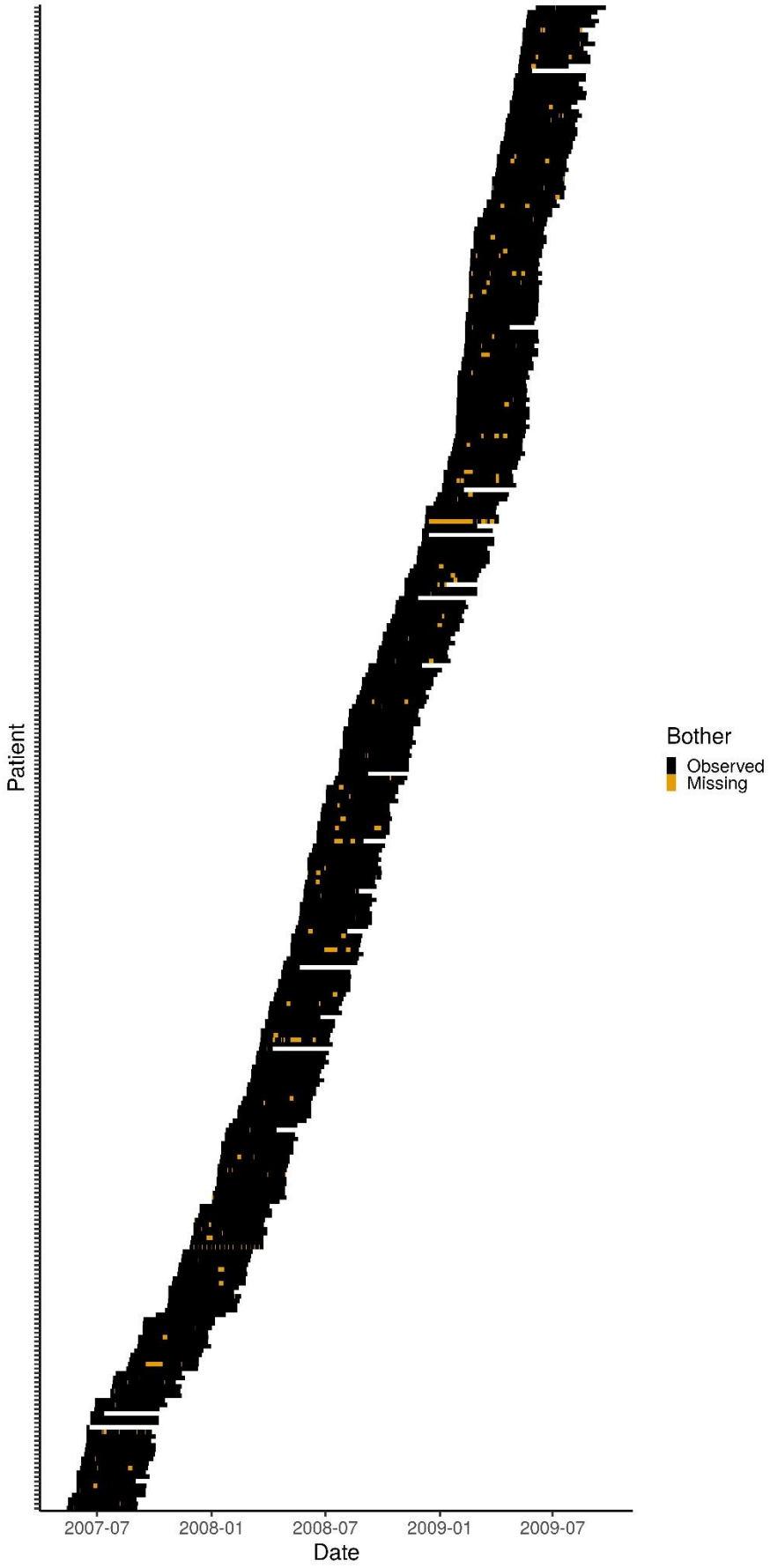


Figure S4: Missing bother scores in SWET dataset. Black and orange indicate observed and missing scores, respectively. The x-axis indicates the date of the measurement.

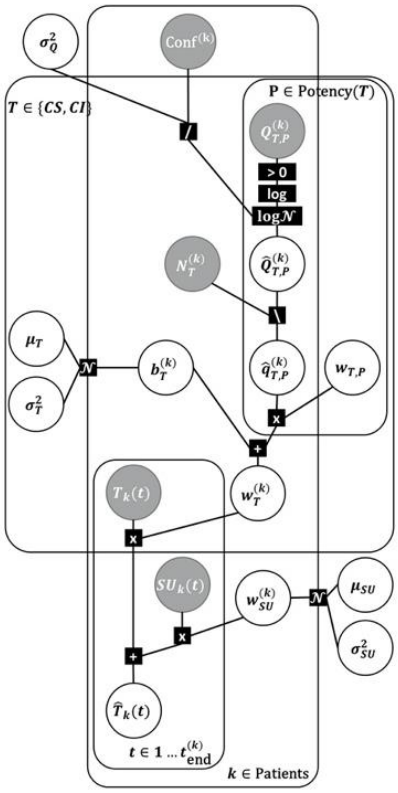


Figure S5: A factor graph that represents the structure of the term corresponding to contribution of treatment in the extended model, $\hat{T}_k(t) = w_{SU}^{(k)} \boldsymbol{SU}_k(t) + w_{CS}^{(k)} \boldsymbol{CS}(t) + w_{CI}^{(k)} \boldsymbol{CI}_k(t)$ (Supplementary A). The grey and white circles represent the observed and latent variables, respectively. The variables are connected to factors (square nodes) that represent the operations or conditional probability distributions. For instance, $b_T^{(k)}$ is normally distributed with mean μ_T and variance σ_T^2, and $\hat{q}_{T,P}^{(k)}$ is defined by $\frac{\hat{Q}_{T,P}^{(k)}}{N_T^{(k)}}$. Plates (squared ovals) represent the variables that are repeated in the model. For example, all variables in the $T \in \{CS, CI\}$ plate are duplicated for corticosteroids (CS) and calcineurin inhibitors (CI).

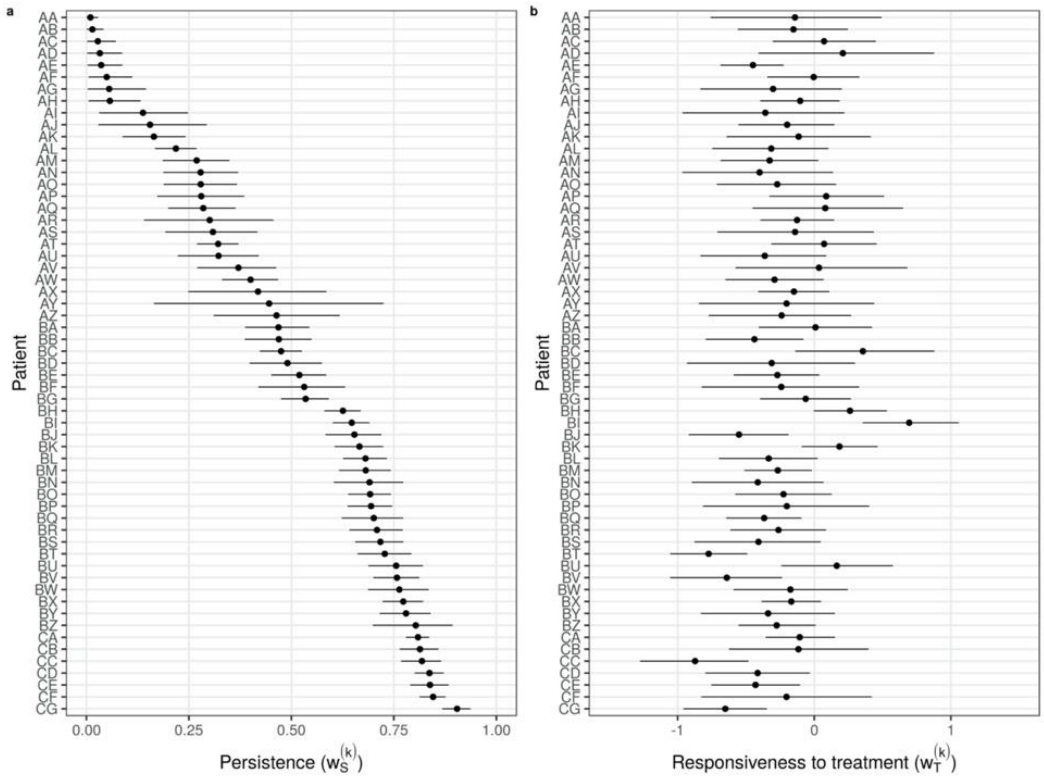


Figure S6: Estimates of the patient-dependent model parameters ($w_S^{(k)}$ and $w_T^{(k)}$) fitted to Flares dataset. Black circles and the line segments represent the mean posterior and the 90% credible interval, respectively. Estimates greatly vary from one patient to another, confirming their patient-dependence. A: $w_S^{(k)}$ (persistence of the severity score). The closer $w_S^{(k)}$ is to 1, the more persistent the severity score is. B: $w_T^{(k)}$ (responsiveness to treatment). The value of $w_T^{(k)}$ quantifies the expected change in the severity score by the treatment.

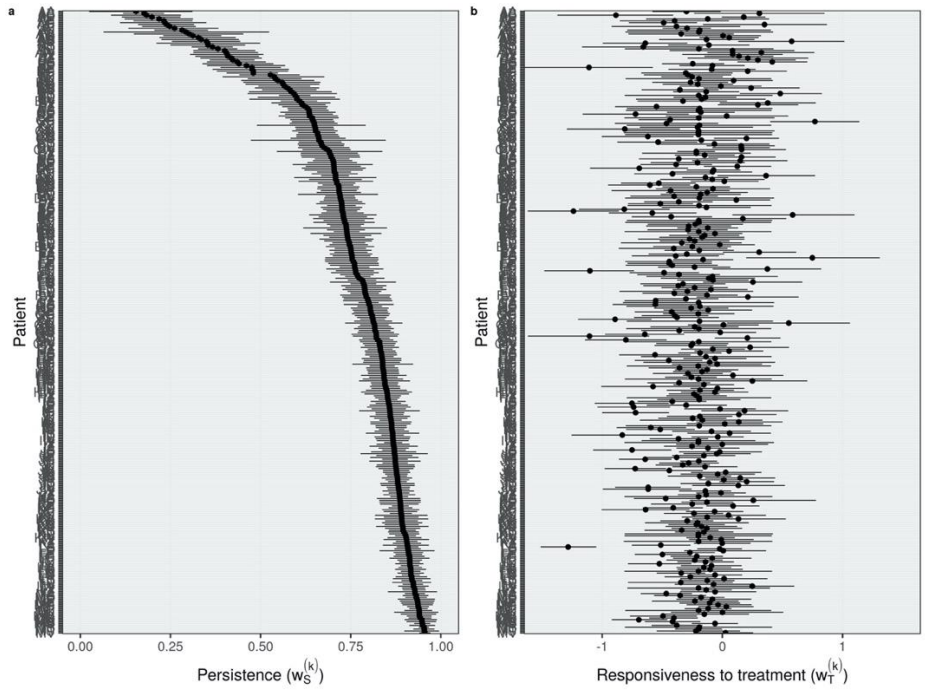


Figure S7: Estimates of the patient-dependent model parameters ($w_S^{(k)}$ and $w_T^{(k)}$) fitted to SWET dataset. Black circles and the line segments represent the mean posterior and the 90% credible interval, respectively. Estimates greatly vary from one patient to another, confirming their patient-dependence. A: $w_S^{(k)}$ (persistence of the severity score). The closer $w_S^{(k)}$ is to 1, the more persistent the severity score is. B: $w_T^{(k)}$ (responsiveness to treatment). The value of $w_T^{(k)}$ quantifies the expected change in the severity score by the treatment.

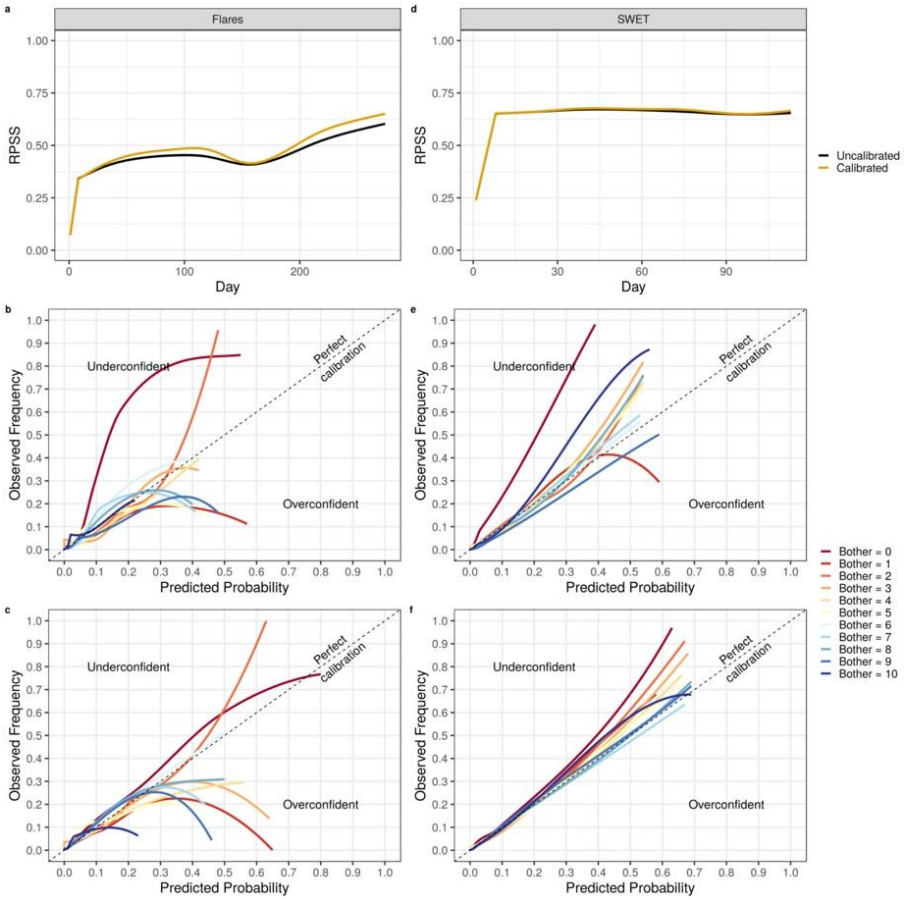
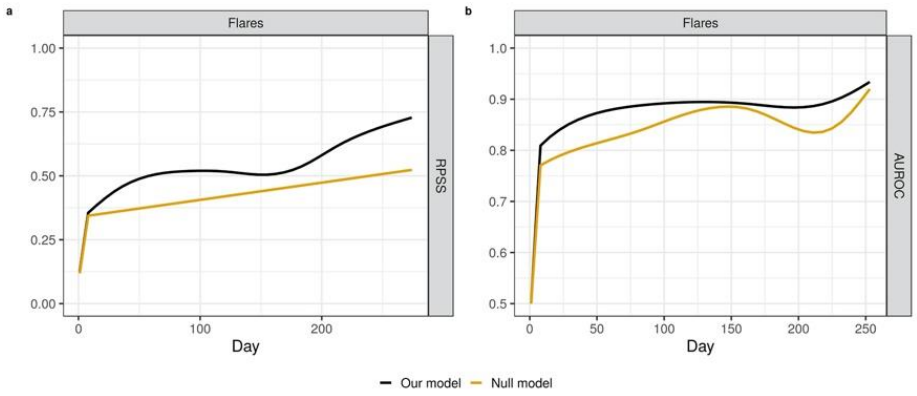
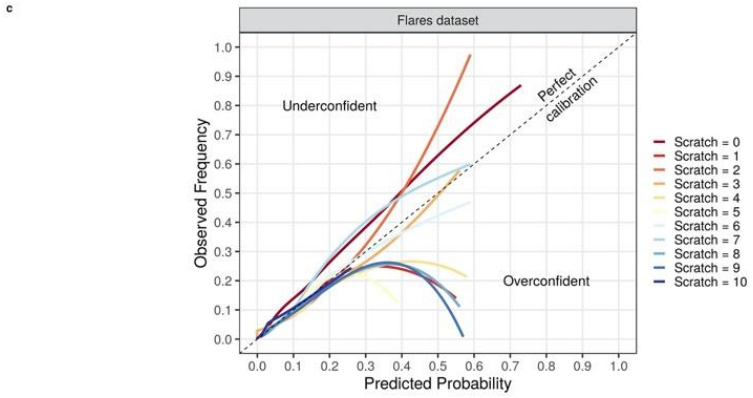


Figure S8: Effect of multi-category calibration on performance for the model trained with Flares (a- c) or SWET (d-f). a, d: RPSS learning curve. b, e: Calibration curves before calibration. c, f: Calibration curves after calibration.





d

Parameter	$\hat{R}$	n_{eff}	Mean	SD	SE	2.5%	50%	97.5%
b_S	1.005	1116	0.260	0.028	0.001	0.204	0.259	0.314
σ_S	1.013	410	0.745	0.016	0.001	0.713	0.746	0.776
σ_P	1.001	1968	0.051	0.005	0.000	0.043	0.051	0.061
μ_T	1.001	4753	-0.220	0.062	0.001	-0.343	-0.219	-0.101
σ_T	1.001	3110	0.324	0.056	0.001	0.224	0.320	0.441

Figure S9: Performance (A-C) and fit (D) of the model predicting the "scratch" severity score that was only available in Flares dataset. A-B: Learning curves for RPSS (A) and AUROC (B) for our model (black) compared to the null model (orange). C: Calibration curves. D: Posterior summary statistics of the main parameters.

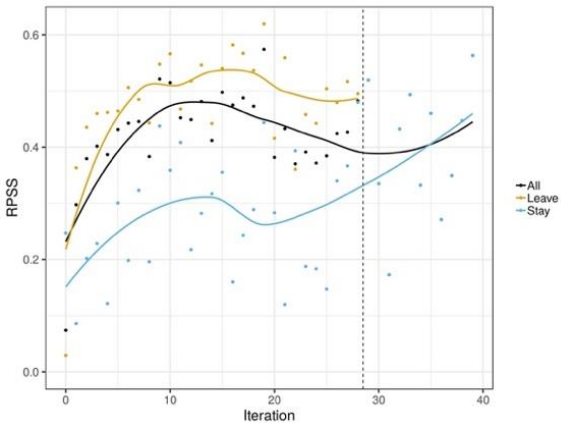


Figure S10: Learning curves of RPSS for the model trained on Flares dataset. The orange and blue circles correspond to the patients who dropped out of the study before and after the 28th iteration, respectively. The total RPSS (black) is the average of the two curves weighted by the proportion of the patients in each group at a given time. The orange and blue curves can both increase, while the average decreases (Simpson's paradox).

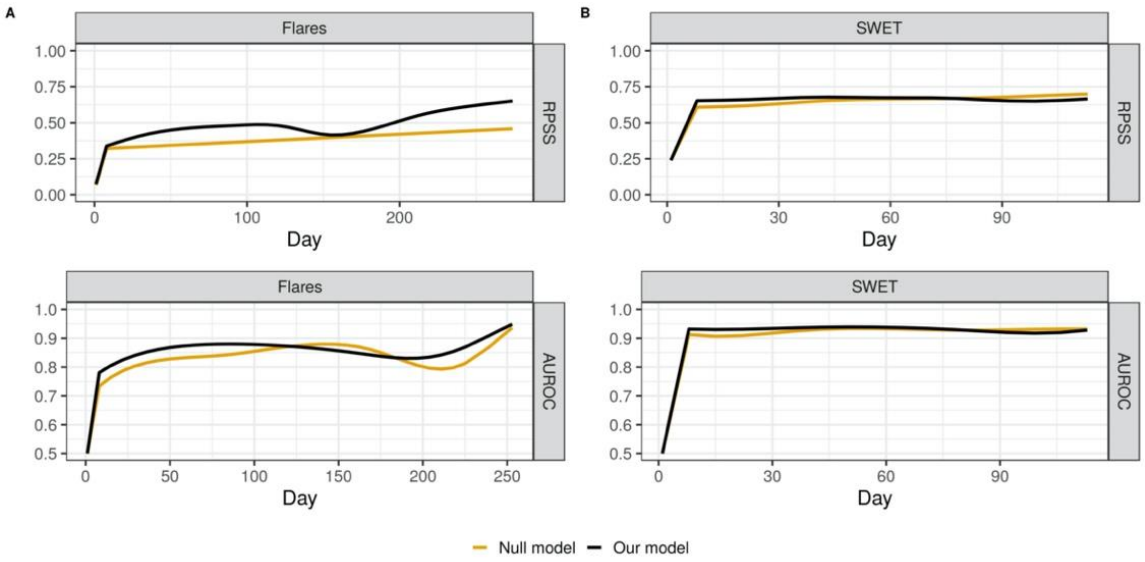


Figure S11: Learning curves of the calibration (RPSS) and discrimination (AUROC) for one-day ahead predictions of our model (black) compared to the null model (orange) trained with Flares (A) and SWET (B) dataset. The null model had a lower performance overall, although the difference is less striking for the model fitted to SWET dataset than that to Flares dataset.

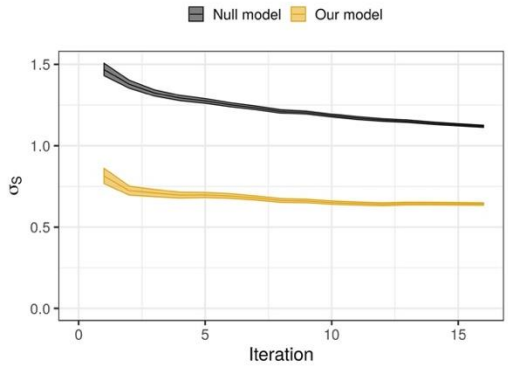


Figure S12: Evolution of σ_S (standard deviation of the predictive distribution and standard deviation of the Gaussian component of the predictive distribution for the null model and our model, respectively) along with the forward chaining iteration for the model trained with SWET. Band corresponds to the 95% credible interval. σ_S does not converge for the null random-walk model, although σ_S is its only model parameter, indicating that the null model is not stationary and not generalisable to unseen data, unlike our proposed model.