

**Towards a data-driven personalised management
of Atopic Dermatitis severity**

Guillem HURALT

17 May 2022

*Submitted in partial fulfilment of the requirements for
the degree of Doctor of Philosophy and the Diploma of Imperial College London.*

Abstract

Atopic Dermatitis (AD, eczema) is a common inflammatory skin disease, characterised by dry and itchy skin. AD cannot be cured, but its long-term outcomes can be managed with treatments. Given the heterogeneity in patients' responses to treatment, designing personalised rather than "one-size-fits-all" treatment strategies is of high clinical relevance. In this thesis, we aim to pave the way towards a data-driven personalised management of AD severity, whereby severity data would be collected automatically from photographs without the need for patients to visit a clinic, be used to predict the evolution of AD severity, and generate personalised treatment recommendations.

First, we developed EczemaNet, a computer vision pipeline using convolution neural networks that detects areas of AD from photographs and then makes probabilistic assessments of AD severity. EczemaNet was internally validated with a medium-size dataset of images collected in a published clinical trial and demonstrated fair performance.

Then, we developed models predicting the daily to weekly evolution of AD severity. We highlighted the challenges of extracting signals from noisy severity data, with small and practically not significant effects of environmental factors and biomarkers on prediction. We showed the importance of using high-quality measurements of validated and objective (vs subjective) severity scores. We also stressed the importance of modelling individual severity items rather than aggregate scores, and introduced EczemaPred, a principled approach to predict AD severity using Bayesian state-space models. Our models are flexible by design, interpretable and can quantify uncertainty in measurements, parameters and predictions. The models demonstrated good performance to predict the Patient-Oriented SCORing AD (PO-SCORAD).

Finally, we generated personalised treatment recommendations using Bayesian decision analysis. We observed that treatment effects and recommendations could be confounded by the clinical phenotype of patients. We also pretrained our model using historical data and combined clinical and self-assessments.

In conclusion, we have demonstrated the feasibility and the challenges of a data-driven personalised management of AD severity.

Statement of originality

I hereby declare that this thesis and the work reported herein is the product of my own research, except where explicitly stated, and that I have acknowledged and referenced the work of others.

Guillem Hurault (2022)

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.



Acknowledgements

I would like to express my sincere gratitude to many people who have supported me during this PhD journey:

- First and foremost, my supervisor, Dr. Reiko Tanaka, for convincing me and offering me the opportunity to do a PhD in her lab. I am very grateful for her continuous support and guidance, for trusting me and giving me the autonomy to conduct my research, and for fostering an amazing lab environment.
- My colleagues and friends, particularly Tara Hameed and Harley Day, for the fun and insightful discussions. I am also thankful to Kailash Arulkumaran and Kevin Pan who contributed to the work presented in this thesis, and to my other colleagues in the Tanaka group: Gemma Pitotti, Natasha Motsi, Jamie Lee, Ariane Duverdier, Victor Ubels, Prakrati Dangarh, Seokjun Lee, Dr. Yuxin Sun, Dr. Mansoor Sheikh, Dr. Rahman Attar, Dr. Takuya Miyano. I reiterate my thanks to Tara Hameed and Ariane Duverdier who provided me feedback on this thesis.
- The many students I helped supervise, and with whom I have learned a lot: Zihao Wang, Antonia Bideharn, Yidi Wu, Umar Shehzad, Ricardo Mokhtari, Aditya Jalin, Pietro Vitiello, Lapo Rastrelli, Valentin Delorieux, Marianne Defresne, Alexis Claudon, Ryan Healey, Adrien Robert De Saint-Victor, Mikhail Iakovlev, Jack McKeon, Corinne Yau, Vick Lau, Sean Chan, Miroslav Gasperek, Lin Rong, Marianna Manou-Kaklamani, Jonathan Ish-Horowicz, Lucas Ducrot, Artur Isufaj, Runze Xu, Anna McGovern.
- My research collaborators: Prof. Hywel C. Williams, Prof. Kim S. Thomas, Dr. Elisa Domínguez-Hüttinger, Prof. Sinéad M. Langan, Dr. Kangmo Ahn, Dr. Young-Min Kim, Dr. Evelien Roekevisch, Dr. Mandy E. Schram, Dr. Krisztina Szegedi, Dr. Sanja Kezic, Prof. Maritza A. Middelkamp-Hup, Prof. Phyllis I. Spuls, Prof. Jean-François Stalder, Sophie Mery, Dr. Alain Delarue, Dr. Markéta Saint Aroman, Dr. Gwendal Josse, Dr Sébastien Barbarot, Dr. Thérèse Nocera, Yann Kling.
- The British Skin Foundation and Pierre Fabre Laboratories for supporting financially the research projects presented in this thesis.
- The Department of Bioengineering for creating a welcoming environment. A special thanks to Dr Angela Kedgley, the Director of Postgraduate studies, who has always been very helpful, and my mentor Prof. Mengxing Tang. I would also like to thank my fellow Bioengineering PhD representatives with whom I had the pleasure to work with: Caryn Urbanczyk, Julia Agramunt, Krysia Broda, Konstantinos Kalyviotis, Leah Xu.

-
- My friends, especially those in the department and particularly Mikolaj Kegler and Alison Pouplin.
 - The R and Stan communities.
 - My family, especially my parents Elisabeth and Frédéric, and my partner Marta, for their support, love and encouragement.

Table of contents

Abstract	3
Declarations	4
Acknowledgements	5
Table of Contents	7
List of Figures	13
List of Tables	16
Nomenclature	17
1 Introduction	20
1.1 Motivation	20
1.2 Aims and objectives	22
1.3 Modelling strategy	22
1.3.1 Model-based vs model-free approaches to decision-making	23
1.3.2 Causality	24
1.3.3 Interpretability	25
1.3.4 Uncertainty quantification	26
1.3.5 Comparing modelling approaches	27
1.3.6 Bayesian modelling	28
1.4 Structure of this thesis	29
2 Background	30
2.1 Atopic Dermatitis	30
2.1.1 Pathogenic mechanisms	31
2.1.2 Treatments	31
2.1.3 Severity scores	32

2.1.4	Properties of severity scores	36
2.2	Bayesian modelling	39
2.2.1	Bayesian statistics	39
2.2.2	Priors	40
2.2.3	Inference algorithms	41
2.2.4	Bayesian workflow	44
2.2.5	Evaluating probabilistic forecasts	44
2.3	Time-series forecasting	47
2.3.1	Forward chaining	48
2.3.2	Learning curves	48
2.3.3	Reference models for time-series forecasting	50
2.3.4	State-space models	52
2.4	Ordered logistic distribution	53
3	Automating the assessment of AD severity	56
3.1	Introduction	57
3.2	Data	58
3.3	Methods	59
3.3.1	Region of Interest detection	59
3.3.2	Severity prediction	59
3.4	Experiments and evaluation	61
3.4.1	Region of Interest detection	61
3.4.2	Severity prediction	62
3.5	Discussion	64
3.6	Afterword	65
4	A statistical model to predict AD severity	67
4.1	Introduction	67
4.2	Methods	68
4.2.1	General approach	68
4.2.2	Data	69
4.2.3	Bayesian models	70
4.2.4	Model fitting	72
4.2.5	Model validation	72
4.3	Results	73
4.3.1	Model fitting	73

4.3.2	Model validation	73
4.3.3	Effects of treatment modalities and risk factors on predictions	75
4.4	Discussion	77
4.4.1	Main findings	77
4.4.2	Strengths of our approach	78
4.4.3	Limitations of the study and future directions	79
5	The role of environmental factors in AD severity prediction	81
5.1	Introduction	82
5.2	Methods	83
5.2.1	Data	83
5.2.2	Mixed effect autoregressive ordinal logistic regression model	84
5.2.3	Model validation	84
5.3	Results	85
5.3.1	Model validation	85
5.3.2	Effect of environmental factors on the model's predictions	86
5.4	Discussion	87
6	The role of biomarkers in AD severity prediction	90
6.1	Introduction	90
6.2	Methods	92
6.2.1	Data	92
6.2.2	Model overview	93
6.2.3	Model validation	94
6.3	Results	95
6.3.1	Model fit and validation	95
6.3.2	Effects of biomarkers on the model's predictions	98
6.4	Discussion	99
7	The role of measurements in AD severity prediction	101
7.1	Introduction	102
7.2	Methods	103
7.2.1	Datasets	103
7.2.2	EczemaPred	106
7.2.3	Model validation	109
7.3	Results of PO-SCORAD models	110
7.3.1	Predictions of severity items	110

7.3.2	Predictions of PO-(o)SCORAD	110
7.3.3	Decomposition of prediction uncertainty in EczemaPred	115
7.4	Results of POEM models	115
7.5	Discussion	119
7.5.1	Main findings	119
7.5.2	Choosing the right score for severity prediction	119
7.5.3	Strengths of our approach	120
7.5.4	Limitations and future directions	121
8	Towards generating treatment recommendations	122
8.1	Introduction	122
8.2	Methods	123
8.2.1	Data	124
8.2.2	Model	125
8.2.3	Priors	126
8.2.4	Treatment recommendation	127
8.2.5	Inference and validation	128
8.3	Results	128
8.3.1	Multivariate dynamic	128
8.3.2	Calibration of PO-SCORAD with SCORAD	130
8.3.3	Treatment effects and recommendations	130
8.3.4	Model validation	132
8.4	Discussion	133
9	Conclusion	136
9.1	Summary	136
9.1.1	Collecting AD severity data	136
9.1.2	Predicting AD severity	136
9.1.3	Generating treatment recommendations	138
9.2	Future directions	138
9.2.1	Research	139
9.2.2	Engineering	140
	References	142
	A Supplementary figures to Chapter 3	161
	B Appendix to Chapter 4	164

B.1	Clinical data	164
B.1.1	Flares dataset	164
B.1.2	SWET dataset	164
B.2	Description of the extended model	164
B.3	Missing value imputation	166
B.4	Choice of priors	166
B.5	Learning curves	168
B.6	Supplementary tables	170
B.7	Supplementary figures	172
C	Supplementary figure to Chapter 5	180
D	Appendix to Chapter 6	181
D.1	Choice of priors	181
D.1.1	Priors for the baseline model	181
D.1.2	Regularised horseshoe prior	183
D.1.3	Reference model priors	183
D.2	Supplementary tables	184
D.3	Supplementary figures	185
E	Appendix to Chapter 7	187
E.1	EczemaPred models	187
E.1.1	Binomial Markov chain	187
E.1.2	Binomial random walk	190
E.1.3	Ordered logistic random walk (v1)	191
E.1.4	Ordered logistic random walk (v2)	192
E.1.5	Multivariate dynamics	193
E.2	Performance metrics	194
E.2.1	Accuracy	194
E.2.2	Learning curves	195
E.3	Reference models	196
E.3.1	Markov chain model	196
E.3.2	Priors for the other reference models	196
E.4	Supplementary PO-SCORAD figures	198
E.5	Supplementary POEM Figures	205
F	Appendix to Chapter 8	206

F.1	Model	206
	F.1.1 Latent dynamics	207
	F.1.2 Measurements	209
F.2	Priors	210
	F.2.1 Power prior	210
	F.2.2 Correlations between severity items	212
	F.2.3 Trend	212
	F.2.4 Inference of daily treatment usage	212
	F.2.5 Treatment effects	213
	F.2.6 Calibration	213
F.3	Treatment recommendations	213
	F.3.1 Utility function	213
	F.3.2 Objective function	214
	F.3.3 Decision profiles	214
F.4	Supplementary figures	216

List of Figures

1.1	Proposed pipeline for a data-driven personalised management of Atopic Dermatitis severity.	22
2.1	Itchy skin with signs of redness and scratch	30
2.2	SCORAD intensity scores	35
2.3	Overview of the Bayesian workflow	45
2.4	Forward chaining with a horizon of 4 days	48
2.5	Schematic of a state-space model	53
2.6	Illustration of an ordered logistic distribution	54
3.1	Disease signs and their relationship to severity scores	57
3.2	EczemaNet overview	60
3.3	EczemaNet predictive performance	64
4.1	Bayesian model of AD severity dynamics	69
4.2	Posterior predictive distribution of AD severity scores for four representative patients from Flares dataset	74
4.3	Model comparison	75
4.4	Fitting of the extended model	76
4.5	Estimated effects of potential risk factors and responsiveness to treatments on the severity score	77
5.1	Example trajectories of the six AD sign scores and the derived AD symptom state for a representative patient.	83
5.2	Model comparison	85
5.3	Comparison of the predictive performance for the models predicting the AD symptom state	86
5.4	Effects of environmental factors	87
6.1	An overview of the Bayesian state-space model for probabilistic predictions of AD severity scores	93
6.2	The posterior predictive distribution of four representative patients by our model predicting EASI	96

6.3	Predictive performance for EASI by our Bayesian state-space model and the reference models, measured by the lpd	97
6.4	Effects of covariates in our model’s predictions of EASI	98
7.1	Example trajectories of PO-SCORAD and its severity items for representative patients from datasets 1 and 2	105
7.2	Model overview	108
7.3	Predictive performance for 4-days-ahead forecasts by EczemaPred models and reference models measured by lpd.	111
7.4	PO-SCORAD prediction by EczemaPred for four representative patients from dataset 1 and dataset 2	112
7.5	Learning curves for 4-days-ahead forecasts of PO-SCORAD evaluated by lpd and accuracy, as a function of the number of training observations, for datasets 1 and 2.	114
7.6	Results of the model with an ordered logistic distribution and multivariate latent random walk dynamic	116
7.7	Predictive performance estimates for one-week-ahead of the different symptoms and POEM by several models	118
8.1	Model and method overview	124
8.2	Data from a representative patient	125
8.3	Power prior contribution and correlogram	129
8.4	Calibration of PO-SCORAD measurements using SCORAD	131
8.5	Treatment effects and recommendations	132
8.6	Analysis of treatment recommendations for a risk neutral patient and a “normal” perceived cost of treatments	133
A.1	Data inclusion, exclusion and validation splits	161
A.2	Architectures of EczemaNet, baselines and ablations	162
A.3	Base architecture comparisons	163
A.4	SASSAD prediction	163
A.5	TISS prediction	163
B.1	Learning curves of RPS for the model trained on Flares dataset	169
B.2	Example data from SWET dataset	172
B.3	Missing bother scores in Flares dataset	173
B.4	Missing bother scores in SWET dataset	174
B.5	Factor graph of the treatment term from the extended model	175
B.6	Estimates of the patient-dependent model parameters fitted to Flares dataset .	176
B.7	Estimates of the patient-dependent model parameters fitted to SWET dataset .	177

B.8	Calibration curves	178
B.9	Results of the model predicting the “scratch” severity score with the Flares dataset	179
C.1	Distribution of the AD signs scores across time and patients.	180
D.1	K-fold cross-validation in a forward chaining setting	185
D.2	Performance of our model and reference models to predict EASI	185
D.3	Predictive performance of our model and reference models for oSCORAD, SCORAD and POEM	186
E.1	Two state Markov Chain	188
E.2	Distribution of the nine severity items and PO-(o)SCORAD in dataset 1.	198
E.3	Distribution of the nine severity items and PO-(o)SCORAD in dataset 2.	198
E.4	Predictive performance of the Extent model with datasets 1 and 2	199
E.5	Predictive performance of the Dryness model with datasets 1 and 2	199
E.6	Predictive performance of the Redness model with datasets 1 and 2	200
E.7	Predictive performance of the Swelling model with datasets 1 and 2	200
E.8	Predictive performance of the Oozing model with datasets 1 and 2	201
E.9	Predictive performance of the Scratching model with datasets 1 and 2	201
E.10	Predictive performance of the Thickening model with datasets 1 and 2	202
E.11	Predictive performance of the Itching model with datasets 1 and 2	202
E.12	Predictive performance of the Sleep loss model with datasets 1 and 2	203
E.13	Learning curves of models predicting PO-oSCORAD, measured by lpd and accuracy, as a function of the number of training observations, for datasets 1 and 2	203
E.14	PO-SCORAD predictive performance changes as the prediction horizon is in- creased by one day, measured by Accuracy and lpd, for datasets 1 and 2	204
E.15	PO-oSCORAD predictive performance changes as the prediction horizon is increased by one day, measured by Accuracy and lpd, for datasets 1 and 2	204
E.16	lpd learning curves for one-week-ahead forecasts as a function the number training observations or equivalently training week	205
F.1	Estimates of the measurement and latent dynamic standard deviations for all severity items	216
F.2	Minimum and maximum of the expected trend component, for each patient and each severity item	216
F.3	Mean and 90% credible interval of the characteristic learning time τ of the calibration process	217
F.4	Predictive performance estimates for four-days-ahead predictions after training the model with 65 days of data	217

List of Tables

3.1	Results of experiments in terms of F_1 score and RPS for all 7 disease signs . . .	64
7.1	Characteristics of PO-SCORAD datasets	104
B.1	Posterior summary statistics for the population-level parameters of the model trained on the Flares dataset	170
B.2	Posterior summary statistics for the population-level parameters of the model trained on the SWET dataset	170
B.3	Posterior summary statistics for the population-level parameters of the extended model	171
D.1	Posterior summary statistics of the population-level parameters for the model predicting EASI without covariates.	184
D.2	MCID and MDC comparison	184
F.1	Decision profiles for treatment recommendations	215

Nomenclature

AD	Atopic Dermatitis
CNN	Convolutional Neural Network
DL	Deep Learning
FLG	Filaggrin
HMC	Hamiltonian Monte-Carlo
HOME	Harmonizing Outcome Measures for Eczema
LOWESS	LOcally Weighted Scatterplot Smoothing
MCMC	Markov Chain Monte-Carlo
MDC	Minimal Detectable Change
MID	Minimal Important Difference
ML	Machine Learning
RoI	Region of Interest
SNR	Signal-to-Noise Ratio
SSM	State-Space Model
SWET	Softened Water Eczema Trial

Severity scores

EASI	Eczema Area and Severity Index (severity score)
oSCORAD	objective SCORAD (severity score)
PO-SCORAD	Patient-Oriented SCORAD (severity score)
POEM	Patient-Oriented Eczema Measure (severity score)
SASSAD	Six Area Six Sign Atopic Dermatitis (severity score)
SCORAD	SCORing Atopic Dermatitis (severity score)
TISS	Three Item Severity Score

Performance metrics

BS	Brier Score
CRPS	Continuous Ranked Probability Score

lpd	log predictive density
RMSE	Root Mean Squared Error
RPS	Ranked Probability Score

Distributions

$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}^+(0, \sigma^2)$	Half-normal distribution with variance σ^2
$\mathcal{N}_{[a,b]}(\mu, \sigma^2)$	Truncated normal distribution in $[a, b]$, with mean μ and variance σ^2
$\log \mathcal{N}(\mu, \sigma^2)$	Log-Normal distribution with log mean μ and log variance σ^2 : if $y \sim \log \mathcal{N}(\mu, \sigma^2)$, then $\log(y) \sim \mathcal{N}(\mu, \sigma^2)$
logit \mathcal{N}	Logit-Normal distribution with logit mean μ and logit variance σ^2 : if $y \sim \text{logit } \mathcal{N}(\mu, \sigma^2)$, then $\text{logit}(y) \sim \mathcal{N}(\mu, \sigma^2)$
$\mathcal{C}(\mu, \sigma)$	Cauchy distribution with location μ and scale σ
$\mathcal{B}(N, p)$	Binomial distribution for N trials with a probability of success p
Beta(α, β)	Beta distribution with shape parameters α and β
$\mathcal{U}(a, b)$	Uniform distribution (discrete or continuous) with bounds a and b , $a < b$

Conventions

Unless stated otherwise, we will use the following mathematical conventions throughout this thesis:

- Greek letters are used to denote parameters and Latin letters are used to denote data.
- Superscripts in parentheses are used to index patients.
- Vectors are denoted in bold.
- Matrices are denoted with a capital letter.
- Subscripts are used to subset vectors or matrices.

For data and parameters, we will use the following symbols:

- y to refer to outcomes. For example, $\mathbf{y}^{(k)}(t)$ represents the vector of outcomes for the k -th patient at time t . $y_i^{(k)}(t)$ denotes the i -th element of $\mathbf{y}^{(k)}(t)$.
- M to refer to the upper bound of y . Usually y is positive, so $y \in [0, M]$.
- \hat{y} to refer to the prediction of y (linear predictor, latent score, etc.).
- k to index patients.
- t to index time.
- x to refer to fixed predictors (time-independent, such as demographics). The corresponding parameters are noted as β . In particular, β_0 corresponds to the intercept.

- u to refer to control inputs (time-dependent, such as treatment usage). The corresponding parameters are noted with θ .
- μ to refer to the mean or location of a distribution.
- σ to refer to the standard deviation or scale of a distribution.
- α to refer to the slope or autocorrelation parameter.

Finally,

- $p(\cdot)$ to denote a probability density function. For example, if $y \sim \mathcal{N}(\mu, \sigma^2)$, then $p(y | \mu, \sigma^2) = f_{\mathcal{N}}(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- $P(\cdot)$ to denote the probability of an event.

Chapter 1

Introduction

1.1 Motivation

Atopic Dermatitis is a chronic inflammatory skin disease characterised by a dry and itchy skin [1]. AD is a complex disease, and despite affecting up to 20% of the paediatric population worldwide [2] and having a high socio-economic impact [3], its pathogenic mechanisms are not fully understood yet. Although AD is often considered as a specific form of eczema presenting an atopic phenotype, “eczema” is commonly used as a synonym to AD [4] [5]. Because of this ambiguity, the terms “Atopic Dermatitis” and “eczema” sometimes refer to different clinical phenotypes in the literature. Usage of the terms “Atopic Dermatitis” and “eczema” also differs between scientific fields and languages [6]. In this thesis, we will follow the World Allergy Organization nomenclature and use “eczema” as a synonym to Atopic Dermatitis [7].

AD cannot be cured, but its long-term outcomes can be managed/controlled with the use of treatments. Managing AD is nonetheless challenging for patients [8] and often results in low adherence to treatment [9]. In addition, treatment responses vary considerably from patient to patient [10]. As a result, personalised (precision) medicine, i.e. tailoring medical treatments to each patient¹ as opposed to a “one-size-fits-all” approach to treatment, is of high clinical relevance for AD [10] [11]. Together with tools to support eczema self-management such as smartphone apps [12], personalised medicine could help patients become more involved in the management of their condition and improve medical outcomes.

Personalised medicine is made possible thanks to many technological advances, such as the ability to generate large volume of data (e.g. health monitoring using IoT devices, gene

¹Even though personalised medicine is often associated with genomics, to identify genes that predispose specific outcomes or treatment responses, we do not restrict personalised medicine to the study of genes in this thesis.

sequencing, etc.) as well as the increase in computational power and the development of new techniques to analyse data. Machine Learning (ML), in particular, has been very successful in fields such as computer vision or natural language processing [13], and holds promises to analyse medical data and uncover complex patterns that are hardly detectable by the human eye [14] [15]. However, there has been little AD research using ML, or more generally advanced data analytics methods, so far [16].

Data-driven computational methods could offer valuable support tools for medical decision-making for AD, notably to generate personalised² recommendations³. The idea of recommending treatment algorithmically to standardise and improve disease management and control has already been demonstrated in a randomised clinical trial, where a rule-based treatment algorithm was derived from a literature review [17]. A computational and data-driven treatment recommendation algorithm could potentially do even better, as the outcomes of different treatment regimens could be simulated beforehand, to assess which treatment is most likely to be effective for a particular patient. These simulations could be personalised to the characteristics of patients and included in a cost-benefit analysis to guide treatment recommendations⁴.

Generating treatment recommendations requires the ability to accurately predict the future evolution of eczema severity. However, past studies have mostly focused on quantifying associations with disease outcomes [16], or remain preliminary with significant limitations [18]. Instead, predictions need to be generated and evaluated on out-of-sample data, as associations often do not generalise to unseen data [19]. Beyond predictions, modelling the evolution of AD severity could help our understanding of the disease dynamics, using predictive performance as a way to compare competing theories [20]. For example, predictive models could be used to investigate the influence of past severity, treatments, environmental factors and even biomarkers on future severity [21] [10]. In a clinical setting, such models could also help track and generate insights into the evolution of the disease, or could be used as a tool to initiate a discussion with patients (e.g. showing potential outcomes under different treatment conditions).

To make AD severity prediction and treatment recommendation tools available to a wide audience, it is crucial to develop the means to collect the appropriate severity data [22]. AD severity is measured by trained clinical staff (often nurses) when a patient visits a clinic, i.e. infrequently, especially if patients live in a remote area. Automatic assessments of AD severity using images taken from a smartphone could help patients track the evolution of their symptoms

²In a decision analysis, one must consider subjective criteria such as patients' preferences. Therefore we prefer to talk about "personalised" rather than "optimal" treatment recommendations.

³We prefer the term "recommendation" as opposed to "decision" in this thesis, because algorithmic recommendations/decisions should not be viewed as prescriptive, but as a tool to support decision-making. For example, the final decision to initiate treatment may rely on external factors, not available to the machine.

⁴Our main focus in this thesis is on treatment recommendations, but recommendations are not necessarily limited to treatments and could relate to other possible interventions, such as environmental factors exposures (e.g. staying indoor during a pollution peak).

daily, and contribute data for the development and large scale deployment of decision support tools.

1.2 Aims and objectives

In this thesis, we aim to pave the way towards a closed-loop system, whereby AD severity data would be collected automatically from camera images, this data would be used to predict the evolution of AD severity, and the resultant predictions would help generate personalised treatment recommendations [21] (Fig. 1.1). We hope such data-driven personalised pipeline for managing AD severity would improve the care of AD patients.

The objective of this thesis is threefold:

1. Automating the assessment of AD severity using camera images (Chapter 3).
2. Developing and exploring the requirements for personalised predictive models of AD severity (Chapters 4, 5, 6, 7 and 8). This will be the main focus of this thesis.
3. Generating personalised treatment recommendations (Chapter 8).

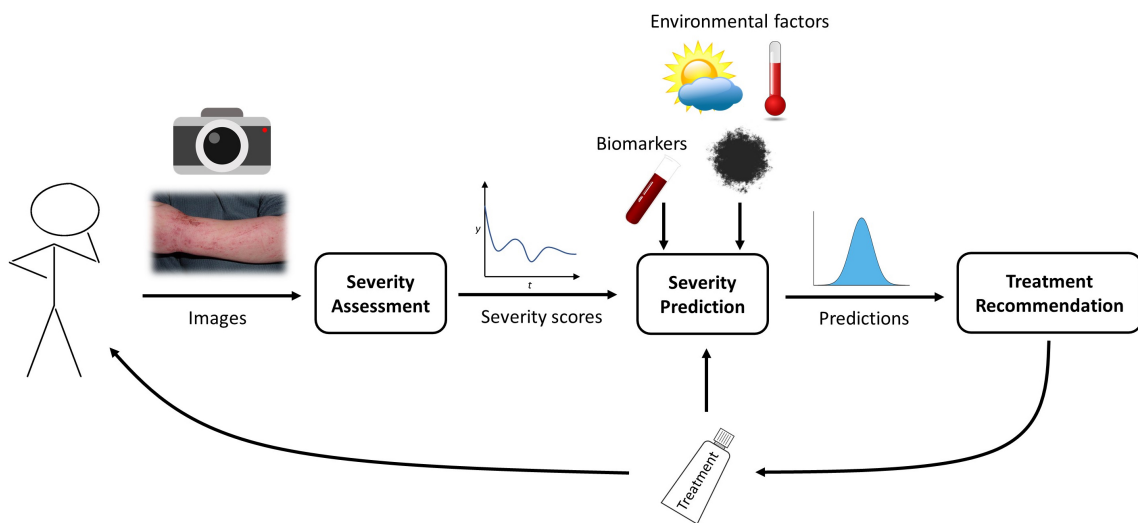


Figure 1.1: Proposed pipeline for a data-driven personalised management of Atopic Dermatitis severity.

1.3 Modelling strategy

Automating the assessment of AD severity using camera images is a classic computer vision task, and there is little debate about the superiority of Deep Learning (DL) techniques in recent

years, compared to more traditional feature-engineering approaches [13] [23].

However, the choice of a model, or more generally of a modelling strategy, is less straightforward for the tasks of predicting the evolution of eczema severity and subsequently generating treatment recommendations.

For these tasks, we will use data collected in different clinical trials or observational studies⁵, although the ultimate goal is to collect data directly from photographic images taken by patients. The datasets used in this thesis are similar, in that they consist of longitudinal data (time-series from different patients) of AD severity measurements. The measurements are imperfect with the presence of measurement errors (cf. Section 2.1.3), and often contain a non-negligible fraction of missing values (e.g. when patients forget to record their severity). Finally, all the datasets are relatively small, in the order of 100 patients and 100 timepoints per time-series or less.

With that in mind, in this section, we will discuss considerations, key requirements and challenges for our models. Based on these, we will argue in favour of statistical modelling as our preferred modelling strategy, and particularly Bayesian modelling.

1.3.1 Model-based vs model-free approaches to decision-making

We will adopt a model-based approach for generating treatment recommendations. This approach consists of first developing a model of the evolution eczema severity, and then generating treatment recommendations by comparing the predictions of the model under different treatment conditions.

Model-free approaches are different from model-based ones in that they can learn how to make optimal decisions without requiring a model of the environment. Model-free approaches are popular in the field of reinforcement learning (RL) [24], whereas statistical decision analysis and control engineering rely on models. Compared to model-free approaches, model-based approaches are, by definition, prone to model misspecifications. However, model-based approaches require less data for training, as the exploration of the action space is guided by the model [25], making them more suitable in our context of working with small data. Having a model of the environment may also be desirable when causality, interpretability, and uncertainty quantification aspects (that we will review below) are important.

In addition to being an intermediate step towards generating treatment recommendations, and in the absence of prior work, modelling the evolution of eczema severity can be useful

⁵We use data from multiple studies as no dataset contains all the necessary information to explore the different aspects of AD severity prediction.

on its own. According to the tripartite division of statistical problems by Bernardo and Smith [26], our problem is \mathcal{M} -complete in the sense that we believe a true model of the human body suffering from AD exists, but we do not presume it can be written down. This is in contrast to \mathcal{M} -closed problems where a true model exists, can be written down and is among a list of models that a researcher can choose from; or \mathcal{M} -open problems where a true model exists but cannot be specified (or that no true model can even be conceptualised [27]). If our problem were \mathcal{M} -open, we could only aim at making predictions, as it would not be possible to say anything about the underlying mechanisms. However, in \mathcal{M} -complete problems, the role of modelling is to produce surrogate (incorrect) models that are tractable and can be used for multiple purposes, including descriptive modelling (summarising data), making inferences or good predictions [28].

1.3.2 Causality

Designing treatment recommendations is a decision-making problem, with the implicit understanding that decisions have a causal impact on the outcomes we consider. Since decisions correspond to interventions (“what if” questions), in order to make the best decision, one must understand the causal structure of the problem [29] [30]. In other words, it is not sufficient to select the action that is associated with the best outcome, if the latter is not caused by the former. It is thus desirable to integrate causal considerations in our models of the evolution of AD severity.

Causal mechanisms can be learnt by conducting interventions on the environment. For example, autonomous systems, such as robots learning how to move, usually have the ability to act on their environment, observe the consequences of their actions and conduct hypotheses (formally or informally) to learn causal relationships between actions and their outcomes. In the medical field, hypotheses can be tested in randomised controlled trials, but it is not possible to let an algorithm experiment treatments on patients for obvious ethical reasons. When online data collection is not possible, one must rely on previously collected data, as in the case of this thesis. In this “offline” data setting, learning how to make decisions is more difficult, as it requires counterfactual predictions (e.g. what would have happened if the patient had taken a different treatment) [31] [32]. For example, ML approaches focus almost exclusively on association rules [30], and can mistake associations for causation. A “naive” ML system could thus recommend patients to avoid going to the doctor to reduce their probability of illness.

Answering causal questions (causal inference) from observational data is challenging. In particular, causal inference requires assumptions and cannot be entirely data-driven: one must specify the causal structure problem and assumes the absence of unobserved confounders

(ignorability or back-door criterion) [32]. However, we do not think we can reasonably assume that the data available to us include measurements of all potential confounding factors to estimate causal effects of treatments on AD severity⁶.

Non-causal predictive models can nonetheless be useful for causal decision-making, as illustrated by their ubiquitous use in industry [33]. For example, the ranking of treatment effects (especially knowing which treatment is the best) is more important for causal decision-making than the precise estimation of causal treatment effects. As a result, learning from confounded data introduces biases in causal treatment effects that do not necessarily influence causal decision-making, while potentially reducing errors due to variance. Non-causal predictive models can thus serve as proxies for causal decision-making, and sometimes even outperform causal models [33]. To illustrate this proxy view, we can consider the suspected role of biomarkers for patient stratification, i.e. biomarkers that would identify the most responsive patients to a given treatment [10]. Unlike treatments or environmental factors, we would not assume that biomarkers directly cause changes in severity, but that biomarkers and severity changes have common causes and are thus associated. Without being a causal driver, biomarkers can therefore be useful as surrogates for multiple causal mechanisms in a predictive model of the evolution of AD severity.

As our work is preliminary, we chose to develop non-causal models, even if the ultimate goal is to generate treatment recommendations⁷. We may nonetheless want to integrate causal considerations in our models, as understanding the causal structure of the problem can guide model development, lead to improvements in predictive performance, result in useful (albeit biased) inference, and help build trust in the model's outputs (see Section 1.3.3).

1.3.3 Interpretability

Ideally, we would like our models to be somewhat interpretable, to engender trust in their outputs. Existing regulations such as the European Union's General Data Protection Regulation (GDPR) highlights patients' "right to an explanation" for automated decision-making [34]. In particular, "black-box" models or algorithms may fail to be accepted by the medical community or patients, as interpretability is an important desideratum for high stake decision problems [22] [35].

⁶For example, treatment effects could be confounded by environmental factor exposures through the patient's "behaviour". That is, the patient's behaviour could influence both their environmental exposures and treatment usage. Alternatively, we can imagine a situation where treatment is used as a preventive measure before being exposed to environmental factors (if a forecast is available), thus confounding the causal effect of treatments.

⁷We will primarily focus on modelling the evolution of AD severity in this thesis, only addressing the treatment recommendation question in Chapter 8.

Interpretability is an ill-defined concept, covering many different aspects, but is often associated with the notion of transparency, in terms of simulatability (whether we can contemplate and understand the model at once), decomposability (whether individual parts such as the model’s parameters can be explained) or whether the optimisation algorithm is itself transparent [36]. In addition, there has been a growing interest in post-hoc interpretability, or explainability, in the field of DL. Post-hoc interpretability usually consists of deploying another model or technics to “explain” an original black-box model. However, post-hoc interpretability is challenging [37], and also highly questionable as post-hoc explanations are not guaranteed to match the algorithm’s true decision process [35]. When available, it is often preferable to use interpretable models rather than relying on post-hoc explanations of a black-box model. This is the approach we will take in this thesis.

1.3.4 Uncertainty quantification

Another desirable characteristic for our models is the ability to quantify uncertainty, using probability distributions as opposed to point-estimates.

There are indeed multiple sources of uncertainty that we would like to keep track of and propagate, such as uncertainties in measurements (due to measurement errors, cf. Section 2.1.3; or missing values) or in parameters (as a result of learning with data of limited size). In addition, we believe it is important to quantify uncertainties in predictions, that are mainly due to the stochasticity in the dynamics of AD and the sheer complexity of predicting health in general⁸ [38].

Failing to quantify uncertainty could lead to suboptimal results and overconfident claims. Quantifying uncertainty is also critical when assessing the risks associated with making a decision. On the contrary, making point predictions would imply a forced choice imposed on the clinician or patient (e.g. dichotomising a continuous probability of disease into a 0-1 classification) [39]. Finally, quantifying uncertainty is useful for science communication, as communicating uncertainty can help maintain the public’s trust in science and the tools to be developed [40] [41].

⁸Apart from specific genetic disorders, health is typically a low signal-to-noise ratio (SNR) environment, where even identical twins can have very different medical outcomes.

1.3.5 Comparing modelling approaches

We can broadly identify three potential approaches⁹ to develop predictive models of the evolution of AD severity: mathematical modelling, statistical modelling, and Machine Learning. We argue here in favour of statistical modelling for the purpose of this thesis.

Mathematical modelling is virtually theory-driven, with the aim to model biological mechanisms of disease using a systems medicine/biology approach [42]. Mathematical models tend to be highly complex with a well-defined and sometimes highly constrained structure. They also require many parameters to be optimised and are therefore prone to overfitting. Even though mathematical models can be very valuable when trying to formalise our understanding of disease mechanisms or formulate new scientific hypotheses [43], they often do not lead to a good fit to data or accurate predictions [44], compared to other approaches. We therefore believe mathematical modelling is not the most appropriate strategy for developing predictive models of AD severity.

Machine Learning models, on the other hand, are completely data-driven and focus almost exclusively on predictions. We can locate statistical modelling somewhere in-between mathematical modelling and machine learning, although the distinction between statistical modelling and Machine Learning is often blurry. Here, we take the view that statistical modelling originates from a data modelling culture and ML from an algorithmic modelling culture [45] [46]. In that view, both statistical modelling and machine learning approaches are data-driven. However, unlike ML, statistical modelling does not treat the data-generating mechanisms as unknown, and is therefore deeply connected to causal theory [47] [48]. For example, neural networks, decision trees, ensembling methods (e.g. random forest, bagging, boosting) or Support Vector Machine can be viewed as ML; whereas we can consider regression (e.g. linear, logistic), including regularised regression (Lasso, Ridge, Elastic Net [49]) or regression splines, as statistical modelling techniques.

ML models are often believed to provide better predictive performance than statistical models, thanks to their ability to model complex and unspecified relationships in data, and that it comes at the price of reduced interpretability. However, this “accuracy-interpretability” trade-off is more of a myth than a conclusion based on empirical results [35]. It is certainly true that ML outperforms statistical models in situations with a high signal-to-noise ratio, such as computer vision or natural language processing. However, when the data has a well-defined structure¹⁰ with meaningful features¹¹, there is little evidence that ML performs better than

⁹This trichotomy is not without limitations, but we believe it can still be helpful to guide our discussion.

¹⁰Our data exhibits time-dependence, can be grouped by patients and the scoring system is structured (cf. Section 2.1.3).

¹¹Demographics, treatments, environmental factors, etc.

statistical models [35], especially for clinical prediction models [50] or time-series forecasting [51].

With their focus on data-generating mechanisms, statistical models are considered more interpretable than ML, and can be used for estimation and inference, in addition to predictions [48]. As they explicitly formulate a probabilistic model for the data, statistical models are also more suitable to quantify uncertainty [46], whereas uncertainty quantification tends to be challenging in ML, especially in DL [52].

There are also benefits in using “simpler” statistical models compared to more complex ML models or algorithms. Statistical models tend to be simpler than ML models or algorithms as they typically assume an additive structure and require fewer parameters to be optimised compared to ML models. In the absence of prior studies that investigated predictive models for AD, we can expect simple models to capture the bulk of the maximum achievable performance, as the benefit of using complex models over simple models is often marginal and sometimes illusory [53]. Using more complex approaches such as ML also typically requires a higher sample size compared to simpler approaches such as statistical modelling [46], and can pose challenges to the reproducibility of results [54]. Finally, simpler models are often easier to implement, can guide the development of more complex models and provide useful benchmarks along the way.

1.3.6 Bayesian modelling

Within statistical modelling, we believe Bayesian modelling provides a relevant framework for developing predictive models of eczema severity.

An attractive feature of Bayesian modelling is its ability to design flexible models tailored to the problem at hand, as opposed to finding the best existing algorithm/model [55]. This can be achieved by specifying the data-generating mechanisms explicitly with probabilistic graphical models and using Bayesian inference¹². With the development of powerful probabilistic programming languages such as Stan [56], this approach also allows researchers to focus more on modelling rather than inference algorithms.

A Bayesian approach also “provides a powerful way to handle uncertainty in all observations, model parameters, and model structure using probability theory” [57], and can explicitly incorporate prior knowledge in the models, a useful feature when working with small data.

¹²This approach has been coined “model-based machine learning” by Bishop [55]. This oxymoron highlights the arbitrary dichotomy between ML and statistical modelling defined above, and that statistical modelling techniques can also be used to model complex relationships in data.

Finally, Bayesian decision analysis offers a simple and principled approach to generating treatment recommendations from existing Bayesian models [58].

1.4 Structure of this thesis

This thesis is organised as follows:

- In **Chapter 2**, we review relevant preliminary information pertaining to this thesis. We start by providing a brief background on Atopic Dermatitis, and especially AD severity scores, which constitute the main source of information for our models. Then, we detail the main concepts of Bayesian modelling, which will be used in Chapters 4, 6, 7 and 8. We also introduce common time-series forecasting methods. Finally, we review the ordered logistic distribution, which we will use to model AD severity items in Chapters 5, 7 and 8.
- In **Chapter 3**, we present a computer vision pipeline, EczemaNet [59], that can automatically assess eczema severity from camera images.
- In **Chapter 4**, we develop a Bayesian model to predict the evolution of eczema severity [60], and demonstrate the possibility of predicting the short-term evolution of AD, and its challenges.
- In **Chapters 5 and 6**, we explore whether additional measurements can help predict future (daily to monthly) AD severity. We investigate the role of environmental factors (weather, pollution) in AD severity prediction in **Chapter 5** [61] and whether serum biomarkers can help predict the severity outcome of a systemic therapy in **Chapter 6** [62].
- In **Chapter 7**, we investigate to what extent using better quality measurements can improve the prediction of AD severity. We also introduce EczemaPred [63], a computational framework that provides building blocks for eczema severity models, available as a R package.
- In **Chapter 8**, we build upon the models proposed in Chapter 7 to integrate multiple sources of information, and generate treatment recommendations using Bayesian decision analysis.
- In **Chapter 9**, we present the general conclusions of this thesis and discuss future directions.

Chapter 2

Background

2.1 Atopic Dermatitis

Atopic Dermatitis (AD) is a common chronic inflammatory skin disease characterised by a dry and itchy skin (Fig. 2.1). It is usually, although imprecisely, referred to as eczema (see Section 1).



Figure 2.1: Itchy skin with signs of redness and scratch. The image is reproduced from [Wikimedia Commons](#) under the [Creative Commons Attribution/Share-Alike 3.0 Unported License](#).

AD is more frequent in children than in adults and affects up to 20% of the paediatric population and 10% of adults worldwide [1]. It usually starts during infancy with 45% of all cases beginning within the first 6 months of life, 60% during the first year, and 85% before 5 years of age [64]. AD prevalence varies in different regions of the world, and is more common in industrialised countries with a Western lifestyle, even though AD is on the increase worldwide [65].

While AD is not life-threatening, it has a severe impact on patients' quality of life, as well

as a high socio-economic impact [3]. Conservative estimates have found that the total cost of AD in the US was over 5 billion US dollars in 2015, including healthcare costs, costs associated with a lower quality of life and costs associated with a loss of productivity in the workplace [3]. The social impact of AD is also significant, with patients reporting feelings of isolation, loss of self-esteem and self-confidence [3].

The diagnosis of AD is based on the physical examination of the skin and patient history (chronic inflammation of the skin, personal or family history of atopy). Several criteria have been developed over the years, including the Hanifin and Rajka criteria [66] and the UK working party criteria [67]. The essential features of these diagnostic tools are pruritus (itch) and inflammation of the skin.

2.1.1 Pathogenic mechanisms

AD is a complex disease and its pathogenic mechanisms remain only partially understood. It has nonetheless been established that AD pathogenesis is primarily driven by immune dysregulation and skin barrier defects [68]. Understanding these mechanisms could help the design of predictive models for AD severity.

On one hand, AD is linked to the predominance of Th2 cells over Th1 cells, two types of T-helper cells (immune cells). This imbalance leads to the overproduction of Immunoglobulin E (IgE), which is a characteristic of allergic diseases [64].

On the other hand, the AD skin is characterised by abnormalities in the stratum corneum (upper layer of the skin), such as decreased hydration, increased water loss, increased skin pH and the lack of diversity in the skin microbiome with an overabundance of the bacteria *S. aureus* [69]. Alterations and deficiencies in the stratum corneum proteins have been implicated in these deficiencies. In particular, mutations in the gene coding the filaggrin protein (FLG), involved in the formation of the stratum corneum, were shown to be associated with AD [70]. Environmental factors are also responsible for the degradation of the skin barrier [1], especially airborne pollutants [71], which are associated with industrialisation and urban living.

2.1.2 Treatments

Treatments are the main interventions we will consider in this thesis.

The current main treatments of AD consist of the application of emollient creams to prevent and soothe dry skin, and topical corticosteroids for inflammatory skin [72] [73]. Calcineurin

inhibitors can also be used for AD patients who are not responsive to corticosteroids, and antibiotics are reserved for patients with bacterial infections [73]. Systemic therapies using traditional immunosuppressants can be used for severe AD patients who do not respond to topical therapies [74]. Proactive therapies have been suggested to control long-term outcomes, as opposed to traditional “reactive” therapies consisting of the application of a treatment when a flare occurs [75].

However, open questions remain about the best and safest way of using these treatments [69], mainly because of the heterogeneity in phenotypes and responses to treatments [10]. Educational aspects may also be at play, considering the problem of low adherence to treatments [9], especially due to corticosteroids phobia [76]. For example, patients often delay initiating treatments after a flare [8].

2.1.3 Severity scores

Severity scores are the primary source of information for the models presented in this thesis. In this section, we will describe the main scoring systems/instruments in length, as they guide the design of our models, and help us understand the limitations of our data.

Several tools to assess eczema severity, commonly called severity scores, have been developed in the last 30 years. Severity scores are the primary outcomes for most clinical trials. All these instruments report AD severity as a single score, obtained by combining the assessments of intensity of AD signs, the extent of eczema (area affected by eczema) or subjective symptoms. The scores thus differ by the severity items they consider and the rule used to combine them.

Recently, the eczema community formed an international focus group, the Harmonizing Outcome Measures for Eczema (HOME), to develop a consensus-based core outcome set (COS) for clinical trials and clinical practice.

Outcomes for clinical trials

HOME recommended the Eczema Area and Severity Index (EASI) [77] [78] as the core outcome instrument for measuring the clinical signs of AD [79]. SCORing AD [80] (SCORAD) and its objective component (oSCORAD) have also been validated as outcome instruments [81] [82], and other scores such as Six Area Six Signs AD (SASSAD) [83] or Three Item Severity Score (TISS) [84] are still routinely used in clinical trials or practice.

The HOME initiative also recommended the Patient-Oriented Eczema Measure (POEM)

[85] as the core outcome for measuring subjective symptoms in clinical trials [86].

Outcomes (self-assessments) for clinical practice

Self-assessments, i.e. scores measured by patients or their carers, can be useful in clinical practice as they are suitable to track the short-term (daily to weekly) evolution of the severity, compared to clinical assessments that can be performed only during clinical consultations of a limited frequency (usually monthly in a clinical trial).

HOME recommended the Patient-Oriented SCORAD (PO-SCORAD) [87] and POEM for measuring eczema severity in clinical practice [88]. However, global assessments or ad hoc measurements such as answers to the question “how much bother did your eczema cause today?” are still routinely used.

SCORAD

SCORAD was created by the European Task Force on Atopic Dermatitis in 1990 [80], in order to define a standardised way of assessing the severity of AD and has become a reference tool for AD assessment since then [81].

SCORAD was defined as the first three principal components of 16 indices of AD symptoms, obtained from a group of 88 AD patients. The three components were derived from a Principal Component Analysis (PCA) and were found to account for more than 50% of the total variance across patients. It is defined as:

$$SCORAD = 0.2A + 3.5B + C \quad (2.1)$$

- $A \in [0, 100]$ corresponds to the extent of AD (aka body surface area), i.e. the percentage area affected by eczema in the whole body. A is often estimated from a drawing of the body where the areas of different regions are pre-calculated (“rule of 9”). For example, if AD signs are only present on the head and the neck (representing approximately 9% of the body), and one arm (representing approximately 9% of the body), then $A = 9 + 9 = 18$.
- $B \in [0, 18]$ is the intensity signs component. It is calculated as the sum of the intensity of six signs (Fig. 2.2):
 - Erythema (redness)
 - Edema/papulation (swelling)
 - Oozing/crusting
 - Excoriation (scratch marks)

- Lichenification (thickening)
- Dryness

The signs are graded as “absent” (0), “mild” (1), “moderate” (2), and “severe” (3). Each sign is evaluated at its own representative area based on the rater’s judgement, where the representative area should reflect the average intensity of the sign in the patient, rather than one “target” area or the worst affected site. The same site may be chosen for two or more signs.

- $C \in [0, 20]$ represents subjective symptoms and is calculated as the sum of two symptoms:
 - Pruritus (itch)
 - Sleep loss/disturbance

The symptoms are scored by the patient using a Visual Analogue Scale (VAS) from 0 (absence) to 10 (worst case possible).

SCORAD ranges from 0 to 103. However, high values of the scores are very unlikely. For example, it is very unlikely to observe a patient with eczema covering the entirety of its body ($A = 100$) or that all intensity signs are severe. In practice, SCORAD above 50 is already considered severe or very severe [89] [90]. The intensity signs component is the predominant component of SCORAD and can contribute up to $3.5 \times 6 \text{ signs} \times 3 \text{ points} = 63$ points out of the 103 points of SCORAD. The extent ($0.2A$) and the subjective symptoms (C) components contribute at most 20 points each to the total SCORAD.

o SCORAD refers to the objective component of SCORAD (A and B terms) and ranges from 0 to 83:

$$oSCORAD = SCORAD - C = 0.2A + 3.5B \quad (2.2)$$

PO-SCORAD is a self-assessment of SCORAD and uses the same scoring system. PO-SCORAD has been shown to be strongly associated with SCORAD, even more so after a few weeks of practice [91]. PO-SCORAD can be assessed with the help of a smartphone app (<https://www.poscorad.com>), which includes illustrations of the different intensities for each of the six intensity signs, for different skin types, and a tool to calculate extent by selecting the affected body regions.

EASI

EASI was developed in 1998 based on the Psoriasis Area and Severity Index (PASI) used for assessing psoriasis severity. It is defined as:

$$EASI = \sum_{r \in \text{Body regions}} \gamma_r \cdot A_r \cdot S_r \quad (2.3)$$

Atopic dermatitis : SCORAD intensity scoring




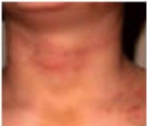


Intensity	None	Mild	Moderate	Severe
Redness	 Score 0	 Score 1	 Score 2	 Score 3
Swelling	 Score 0	 Score 1	 Score 2	 Score 3
Oozing / crusting	 Score 0	 Score 1	 Score 2	 Score 3
Scratch marks	 Score 0	 Score 1	 Score 2	 Score 3
Skin Thickening (Lichenification)	 Score 0	 Score 1	 Score 2	 Score 3
Dryness	 Score 0	 Score 1	 Score 2	 Score 3

Figure 2.2: SCORAD intensity scores. The image is supplied by [DermNet NZ](http://DermNetNZ.org) under the [Creative Commons Attribution-NonCommercial-NoDerivs 3.0 \(New Zealand\)](https://creativecommons.org/licenses/by-nc-nd/3.0/) license.

- Four body regions are considered: head and neck, trunk, upper limbs, and lower limbs.
- γ_r is a fixed multiplier representing the relative importance of each body region¹, so that the sum of the multipliers over regions is 1. For example, lower limbs and the trunk contribute more to EASI than upper limbs and the head and neck.
- A_r represents the area score for region r , on a discrete scale from 0 (no eczema in the region) to 6 (the entire region is affected by eczema).
- $S_r = \sum_{s \in \text{Sign}} I_{r,s}$ is the sum of the intensity score, $I_{r,s}$, of four signs (redness, thickness, scratching, and lichenification), each of which is assessed by none (0), mild (1), moderate (2) or severe (3).

EASI ranges from 0 to 72.

POEM

Unlike SCORAD and EASI, POEM focuses on the subjective illness experienced by the patient. The POEM score is the sum of the answers to 7 questions asking how many days *an AD symptom* was present over the last week. The answer can be “No day” (0), “1-2 days” (1), “3-4 days” (2), “5-6 days” (3) or “Every day” (4). As a result, POEM ranges from 0 to 28.

The questions (formulated for the carers of children here) are:

1. On how many days has your child’s skin been **itchy** because of their eczema?
2. On how many nights has your child’s **sleep** been disturbed because of their eczema?
3. On how many days has your child’s skin been **bleeding** because of their eczema?
4. On how many days has your child’s skin been **weeping or oozing** clear fluid because of their eczema?
5. On how many days has your child’s skin been **cracked** because of their eczema?
6. On how many days has your child’s skin been **flaking** off because of their eczema?
7. On how many days has your child’s skin felt **dry or rough** because of their eczema?

2.1.4 Properties of severity scores

In the previous section, we alluded that scores such as EASI, (o)SCORAD and POEM had been validated. Many aspects (aka psychometrics properties) are considered when evaluating a new scoring system, such as whether it measures what it is expected to measure (construct validity), its interpretability or its ease of use. Here, we detail some aspects of the validation procedure

¹The multipliers are slightly different between children (0-7 years) and adults (8 years or more).

of severity scores, that are of interest when evaluating the predictions of these scores, namely the quantification of measurement errors and what can be considered an important change.

Inter- and intra- rater reliability

Inter-rater reliability refers to the degree of agreement between raters, i.e. to what extent the scores are independent of a particular rater. In particular, if the inter-rater reliability is high, “raters can be used interchangeably without the researcher having to worry about the categorisation being affected by a significant rater factor” [92].

Intra-rater reliability refers to the degree of agreement between repeated assessments of the same item by a single rater.

Estimating the extent of eczema was found to display high inter-rater variability [93]. Moreover, the scoring of intensity signs can be challenging, for example the distinction between mild or moderate lesions is not always clear (Fig. 2.2) and mostly an artefact of the measurement process, as severity is likely a continuum in reality. Overall, the evidence is poor that SCORAD, EASI and POEM have a good inter- and intra-rater reliability [82], i.e. that the measurements are perfect².

Minimal Detectable Change (MDC)

Inter- and intra-rater variability implies the existence of measurement errors, compared to a “true” estimate of the severity, which could be obtained by averaging the repeated measurements of multiple independent raters. As a result, it can be desirable to know when a change can no longer be explained by the presence of measurement noise (at a given confidence level), i.e. when it exceeds the Minimal Detectable Change (MDC).

In the context of developing predictive models, the MDC can provide a lower bound for the minimum error it is possible to achieve (cf. Bayes error rate), when single assessments are considered as “ground truth” for prediction metrics.

To the best of our knowledge, the MDC has not been estimated for SCORAD, EASI, and POEM.

²Hence the wish to model measurement uncertainties, expressed in Section 1.3.

Minimal Important Difference (MID)

A change that is detectable statistically is not necessarily “important”³ [94]. The Minimal Clinically Important Difference (MCID) was thus introduced in 1989 [95] to determine whether a medical intervention improves perceived outcomes in patients. It is defined as the smallest change in a disease outcome measure that can be considered important/meaningful for patients, as opposed to a statistically significant change. The Minimal Important Difference (MID) was introduced as a synonym to the MCID to clarify the focus on patients’ experiences rather than clinical interpretations [96]. The MID is also referred to as “responsiveness to change” in terms of psychometric properties of severity scores.

Estimating the MID (or MDC) for a disease outcome can notably be useful in designing clinical trials, as it can inform an effect size of practical interest to compute the power (or equivalently, given the power, compute the required number of participants) of minimal effect or equivalence tests to estimate treatment effects. The MID can also put the average prediction error of models in perspective.

However, the concept of MID is also problematic in many ways⁴:

- Dichotomising a change as “important” or “not important” is rarely a good idea [97]. In particular:
 - MID is often estimated using global assessments serving as anchors, which are themselves poorly defined [98].
 - Procedures to estimate the MID often ignore the trade-off between false positives (changes classified as important when they are not) and false negatives (changes classified as not important when they are) [99] [100] [101], while this trade-off should be specified explicitly according to the research question. Indeed, classifying a change as “important” or “not important” will produce misclassification and is, therefore, a decision-making problem, which cannot be solved using data only.
- MID implicitly assumes linearity, i.e. the perception of change is the same regardless of the severity. However, it is unclear whether patients are more perceptive of absolute changes compared to relative changes, for example.
- The purpose of the MID is to quantify what is a meaningful change for patients, which can appear in conflict with the implicit search for a unique MID value that would apply to all patients.
- While the goal of “MCID studies is the search for a unique threshold value, [...] ironically, the different methods produce a variety of MCID values” [99].

³Nor an “important” change is necessarily detectable, even if this is desirable.

⁴This is also true to some extent about the MDC.

Even though there have been attempts to estimate the MID of SCORAD, EASI and POEM [102], we do not rely on these measures in this thesis. Instead, we can interpret the prediction error in light of the maximum value M that the score can take, instead of using the MID as a reference.

2.2 Bayesian modelling

Most of our analyses are probabilistic by design, especially Bayesian, notably because Bayesian modelling provides a useful framework to develop flexible models and quantify uncertainty (cf. Section 1.3). In this section, we review the main concepts of Bayesian statistics and modelling required to comprehend this thesis, assuming the readers are already familiar with basic statistical and machine learning concepts. We will refer interested readers to [103] as an introduction to statistical learning, and to [104] and [105] for more details about Bayesian modelling.

2.2.1 Bayesian statistics

Here, we introduce Bayesian statistics in contrast to the standard (i.e. usually taught and applied) frequentist approach. Philosophically, the two approaches notably differ in the way they interpret what a probability is. In the frequentist setting, a probability is the long-term frequency of an event (e.g. the frequency of heads in coin flip experiments), whereas in a Bayesian setting, a probability is a measure/quantification of the uncertainty of an event (e.g. the probability of rain tomorrow⁵).

A parameter is thus considered differently in the two settings. In a frequentist setting, a parameter is considered to have a true, fixed but unknown value (not a random variable), and its estimate is a random variable. In a Bayesian setting, a parameter is considered a random variable and can be described directly using probability distributions. The implications of this difference can be illustrated with the notion of confidence intervals (in a frequentist setting) and its Bayesian “equivalent”, the credible intervals. A (frequentist) confidence interval is an interval that is designed to include the true value of the parameter at a certain rate (e.g. 95%): if the procedure to produce a $x\%$ confidence interval is repeated a large/infinite number of times, $x\%$ of the $x\%$ confidence intervals should include the true value (this property is called coverage) [106]. However, since the true value is unknown and fixed in a frequentist setting, a $x\%$ confidence interval cannot be the interval in which the true value lies with $x\%$ probability:

⁵This probability is not meaningful in a frequentist paradigm as the event cannot be repeated.

on a given experiment, the true value is either in or out of the confidence interval. In contrast, a Bayesian $x\%$ credible interval is an interval in which the parameter lies with $x\%$ (subjective) probability.

When data y is observed, Bayes' theorem is used to update the prior $p(\theta)$ over the parameters θ , by their posterior distributions $p(\theta|y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta) \quad (2.4)$$

$p(y|\theta)$ is the likelihood (the “model”) and $p(y) = \int p(y|\theta)p(\theta)d\theta$ is the evidence term. The evidence term can be seen as a normalisation constant (it is not a function of θ) and its calculation (often intractable) is usually not necessary for Bayesian inference. The role of the prior $p(\theta)$ is to incorporate information about the problem that is not directly present in the data.

Bayesian statistics are widely used in science, but remain controversial [107], even though the criticisms often stem from misunderstandings or incorrect applications of Bayesian concepts [108] (not that fair criticisms are nonexistent [109]). For example, the Bayesian approach is often criticised by the existence of the prior $p(\theta)$, which supposedly introduces subjectivity in the analysis, while “science should be objective” [107]. However, the choice of a model/likelihood is often as subjective as the choice of the prior [110] [111], and a prior can be perfectly objective if it is based on historical data [112]. It is usually more helpful to think in terms of transparency, consensus, or awareness of multiple perspectives to evaluate a statistical approach [113], rather than introducing a false dichotomy between objectivity and subjectivity in statistics.

2.2.2 Priors

Prior specification is essential to Bayesian analysis, and nonetheless one of its challenging parts. Many types of priors are available, including conjugate priors, Jeffrey's priors, hierarchical priors, regularising priors, sparsity-inducing priors, maximum entropy priors, penalise complexity priors, etc. We do not attempt to review all these different types of priors for concision, but we can broadly classify priors as informative, weakly informative and non-informative.

Informative priors are priors that reflect substantial knowledge about the parameter values, and are usually derived from the literature. On the contrary, uninformative priors (aka improper priors, uniform priors, or flat priors) put equal weights on all parameter values, making Bayesian inference equivalent to frequentist inference ($p(\theta)$ is constant in Eq. (2.4) and the posterior is proportional to the likelihood). Even though these priors are called uninformative, a change of

variables can make a uniform prior highly informative. For example, a uniform prior on the logit space (e.g. for the intercept of a logistic regression) would result in a highly informative prior on the probability space with modes at 0 and 1. As such, “the prior can often only be understood in the context of a likelihood” [114]. Finally, weakly informative priors are priors designed to rule out implausible parameter values, without excluding values that could make sense. For example, a weakly informative prior on the height of humans would at least exclude values with an order of magnitude that is different than one meter (e.g. excluding values below 10cm and above 10m). In this thesis, we will often use weakly informative priors.

2.2.3 Inference algorithms

Once the likelihood and priors have been specified, Bayesian inference is conducted to obtain the posterior distribution. There are mainly two types of algorithms to perform Bayesian inference: algorithms relying on variational inference, i.e. approximating the posterior distribution, and algorithms relying on sampling the posterior distribution using Markov chains. In this thesis, we will focus on the latter type of algorithms, as early attempts at fitting models with variational inference methods, notably using expectation propagation [115], were unsuccessful.

Sampling algorithms are based on the Markov Chain Monte-Carlo (MCMC) method and include Gibbs sampling, the Metropolis-Hasting algorithm, and the Hamiltonian Monte-Carlo algorithm [116]. In specific cases, for example in an online learning setting when frequent updates are necessary, Sequential Monte-Carlo algorithms (SMC, particle filters) can be used. However, SMC tends to be less flexible than MCMC algorithms for model development (the main focus of this thesis, as opposed to developing computational methods), so they will not be discussed here.

Intuition behind MCMC

The purpose of MCMC is to obtain random samples from a probability distribution, the posterior $p(\theta|y)$, which is difficult to sample, especially as the number of parameters grows. For example, if we want to sample a one-dimensional cube (that is, a segment), we can use 10 evenly spaced points. If we sample a two-dimensional cube instead of a segment at the same sampling rate, we would need 100 evenly spaced points; and 1000 evenly spaced points in the case of a three-dimensional cube. This is the curse of dimensionality, as the number of samples needs to grow exponentially with the number of dimensions. However, in a high dimensional space, the probability mass of non-uniform distributions tends to concentrate in small regions. In the cube analogy, this means that we do not need to sample the whole volume but maybe only

the centre, for example. MCMC methods take advantage of this phenomenon by randomly sampling the regions of high probability mass.

With MCMC, the sampling is made by Markov Chains. A Markov Chain is characterised by a transition operator and describes how to perform random walks in a graph. The basic idea of MCMC is to let Markov Chains explore the regions of high probability mass, by generating dependent samples rather than independent samples. Briefly, if we have a “good” sample of θ , sampling around this value is likely to generate other “good” samples. The task is then to construct the transition operator, the Markov Chain, to transition from one θ to another, for which the equilibrium (stationary) distribution is the target distribution. Finally, the dependence between samples can be addressed by thinning⁶ or computing an effective sample size (details below).

Metropolis algorithm

The original algorithm for MCMC was the Metropolis algorithm (later generalised as Metropolis-Hastings algorithm). This algorithm performs a random walk and uses an acceptance/rejection rule to converge to the target distribution.

From a starting sample θ , a new sample θ' is proposed from a proposal distribution⁷. This new sample θ' is accepted with a probability equal to $\min(1, \frac{\pi(\theta')}{\pi(\theta)})$, where $\pi(\theta) = p(y|\theta)p(\theta)$ ⁸, otherwise a new sample is proposed. If the process is repeated long enough, the distribution of samples is guaranteed to converge to the target distribution.

Hamiltonian Monte-Carlo

Choosing an appropriate proposal distribution in the Metropolis (-Hastings) algorithm is key to ensure an efficient exploration of the posterior distribution. Unfortunately, relying on a random walk is usually not an efficient way to explore the parameter space, especially in high dimensions, where the algorithm either stays in the local neighbourhood of the current state, or has a low acceptance rate. The Hamiltonian Monte-Carlo (HMC) algorithm tries to overcome this issue by replacing the random walk with a Hamiltonian flow (used to describe

⁶Saving only a fraction of the samples, as the dependence decreases with the increase in the number of transitions between samples. For example, if the second sample of the Markov Chain is highly dependent on the first sample, the tenth or hundredth samples may be reasonably independent from the first sample.

⁷In the original Metropolis algorithm, the proposal distribution needs to be symmetric, and is typically a normal distribution $\theta' \sim \mathcal{N}(\theta, \sigma^2)$, where σ corresponds to a step size.

⁸Taking the ratio of π is equivalent to taking the ratio of posteriors, $\frac{\pi(\theta')}{\pi(\theta)} = \frac{p(\theta'|y)}{p(\theta|y)}$, without requiring the computation of the normalising constant.

the diffusion of particles in physics), which relies on the computation of gradients to propose better transitions.

In the Bayesian analyses presented in this thesis, we use a particular flavour of the HMC algorithm, the No-U-Turn Sampler (NUTS) [117]. NUTS is designed to avoid retracing its own steps, and is characterised by its use of adaptive hyperparameters that facilitates hyperparameter tuning. In particular, we use NUTS as implemented in the probabilistic programming Stan [56].

MCMC diagnostics

When using MCMC, it is desirable to know whether the Markov Chains have converged to their stationary distribution, i.e. whether they are sampling the posterior distribution or they are “stuck” elsewhere. Unfortunately, like many inference algorithms, it is not possible to obtain guarantees that a particular Markov Chain has converged, but instead the algorithm provides diagnostics of a lack of convergence.

Before computing any diagnostics, the first iterations of the Markov chain, corresponding to a “burn-in” (or “warm-up” in Stan [56]) period, are discarded, as they are unlikely to have converged yet.

Running the Markov chain for as long as possible is a good practice, since the longer a chain is being run, the more likely it is to converge. However, the extent to which this is possible in practice is limited. Instead, multiple chains, using different initial conditions, can be run in parallel, and convergence can be assessed by checking whether different Markov chains sample the same distribution. This can be done visually by inspecting trace plots (time-series plots of MCMC draws) or numerically using the Gelman-Rubin convergence diagnostic, \hat{R} [118] [119], for example. As a rule of thumb, $\hat{R} < 1.1$ indicates good “mixing” of the Markov Chains, i.e. the chains sample the same distribution.

In the case of HMC, numerical errors (divergences) occurring in the computation of Markov Chain transitions can also be flagged as an indicator for a lack of convergence.

Finally, the precision of parameter estimates can be assessed by computing an effective sample size N_{eff} (an efficiency diagnostic as opposed to a convergence diagnostic). Given a parameter θ whose distribution is estimated by N independent and identically distributed samples, the sample mean $\bar{\theta}$ is given by $\bar{\theta} \sim \mathcal{N}(\mu_{\theta}, (\frac{\sigma_{\theta}}{\sqrt{N}})^2)$, 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 \sqrt{N} , suggesting that 100 times more samples are required if one more digit of precision is needed for $\bar{\theta}$. 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 computed from the

autocorrelation function of the Markov chains [119].

2.2.4 Bayesian workflow

In this section, we describe the different steps of a Bayesian analysis and the corresponding Bayesian workflow we adopted in this thesis [57] [120] (Fig. 2.3). The Bayesian workflow is iterative, starting from a most simple yet relevant model for the task, and adding complexity gradually.

After initially specifying a model (likelihood) and priors, prior predictive checks are conducted to check that the chosen priors are reasonable [121]. Prior predictive checks consist of sampling parameters from the prior distribution and generating data according to the model, to obtain samples from the prior predictive distribution. The priors can then be understood in the outcome space, which is generally more interpretable than the parameter space, to examine whether the priors conform to our understanding of the problem.

Then, fake-data checks (aka fake-data simulations) are conducted by fitting the model with samples from the prior predictive distribution (the fake-data, with known data-generating mechanism), to verify that the inference algorithm is able to retrieve known parameters. Fake data simulations are a minimum requirement to validate the inference algorithm, as inference algorithms often only guarantee asymptotic results (with infinite data and computation) on toy models. Fake-data simulations are especially useful to detect problems with the computational method, but are not sufficient to guarantee that the inference on real data will be error-free (absence of evidence is not the evidence of an absence). A more exhaustive validation of the computational method can be done using Simulation-Based Calibration (SBC) [122], even though it is often computationally intensive.

Finally, the model can be fitted to real data: convergence diagnostics are computed and posterior predictive checks or out-of-sample validation (e.g. cross-validation) are conducted to assess model fit and validation. Posterior predictive checks are useful to detect failings of the model and are more qualitative (e.g. visual) than quantitative. The idea behind posterior predictive checks is to generate data from the posterior predictive distribution $p(y_{\text{new}}|y) = \int p(y_{\text{new}}|\theta)p(\theta|y)d\theta$, and check whether it is similar to the data that was used to fit the model.

2.2.5 Evaluating probabilistic forecasts

Predictions from a probabilistic model come in the form of distributions rather than point-estimates. Given the posterior distribution $p_{\text{post}}(\theta) = p(\theta|y)$ (we drop the condition on training

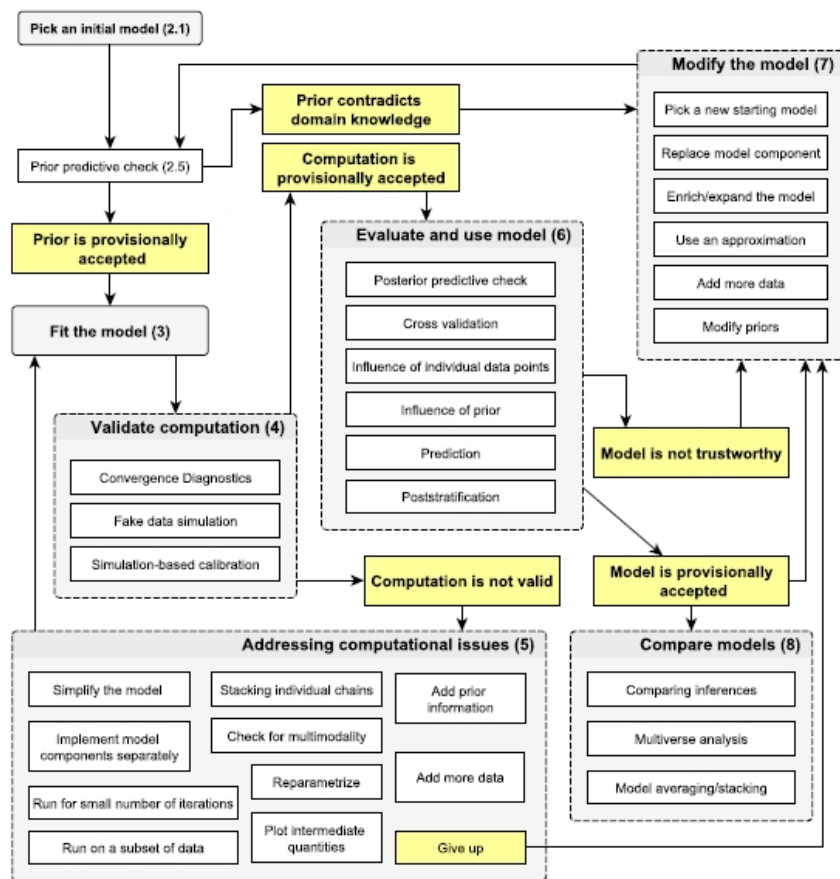


Figure 2.3: Overview of the Bayesian workflow. The figure is reproduced from [57].

data y for concision), the expected predictive distribution of a new datapoint y_{new} is given by:

$$p(y_{\text{new}}) = \int p(y_{\text{new}}|\theta)p_{\text{post}}(\theta)d\theta \quad (2.5)$$

Ideally, probabilistic forecasts should be calibrated. A classifier is said to be calibrated if, for example, an event that is forecasted with a probability of 50% (forecast probability) happens approximately 50% of the time (observed frequency). If the event happens only 30% of the time, the classifier is called overconfident. If the event happens 70% of the time, then the classifier is called underconfident. However, a perfectly calibrated classifier can be useless when the forecasted probability matches the prevalence of the event (historical forecast, cf. Section 2.3.3). The goal of probabilistic forecasts is then to maximise the sharpness of the predictive distribution, subject to calibration [123].

For discrete outcomes, calibration plots can be used to assess visually whether forecasts are calibrated [124]. More generally, scoring rules provide quantitative measures to evaluate probabilistic forecasts [125] [126]. In particular, proper scoring rules encourage honest forecasts [126]. This is not the case of improper scoring rules, such as classification accuracy, sensitivity,

specificity, etc. which can be gamed as they reduce probabilistic forecasts to all-or-none predictions [127].

In this thesis, we will mostly consider the logarithmic scoring rule, the log predictive density (lpd, aka predictive log-likelihood, which is equivalent to (minus) the Log-Loss or cross-entropy for classification problems), as well as quadratic scoring rules: the Brier Score (BS) for binary forecasts, the Ranked Probability Score (RPS) for ordinal forecasts and the Continuous Ranked Probability Score (CRPS) for continuous forecasts. Logarithmic scores tend to penalise incorrect predictions more compared to quadratic scores. Logarithmic scores are also local, which means they only focus on the density at the test datapoint, whereas quadratic scores are global and focus on the entire distribution. For example, unlike the RPS, the lpd does not consider the ordinal nature of disease signs, although it may be preferable to predict “moderate” rather than “mild” when the true label is “severe”.

In the following, we present the equations for computing the scoring rules for individual predictions, which can then be averaged for a set of predictions (e.g. forward chaining iteration).

Log predictive density (lpd)

The log predictive density (lpd) is the local strictly proper logarithmic scoring rule that is defined by the density (probability mass in the case of a discrete outcome) assigned to the true outcome. For a new datapoint y_{new} , the lpd is defined by:

$$lpd(y_{\text{new}}) = \log p(y = y_{\text{new}}) \quad (2.6)$$

$lpd \in] - \infty, +\infty[$ for continuous outcomes and $lpd \in] - \infty, 0]$ for discrete outcomes (i.e. using probability mass functions), with higher values of lpd indicating better forecasts.

Brier Score (BS)

The Brier score is a quadratic scoring rule for categorical outcomes defined as the Euclidean distance between the outcome y_{new} and its forecast. For a multi-category forecast, the Brier score is defined as:

$$BS(y_{\text{new}}) = \sum_{i=1}^R (f_i - o_i)^2 \quad (2.7)$$

- R is the number of classes
- $f_i = P(y = i)$ is the forecast probability to predict the i -th class

- $o_i = \delta_{i,y_{\text{new}}} \in \{0, 1\}$ is the actual outcome of the i -th class, with $\delta_{i,j}$ the Kronecker delta.

$BS \in [0, 2]$, where 0 corresponds to a perfectly-calibrated classifier.

Ranked Probability Score (RPS)

The Ranked Probability Score (RPS) is a quadratic scoring rule for discrete ordinal outcomes. For a single datapoint y_{new} , the RPS is defined as:

$$RPS(y_{\text{new}}) = \frac{1}{R-1} \sum_{i=1}^R (F_i - O_i)^2 \quad (2.8)$$

- R is the number of classes.
- $F_i = P(y \leq i)$ is the (expected) cumulative forecast distribution.
- $O_i = \mathbb{1}_{y_{\text{new}} \leq i}$ is the cumulative observed outcome ($\mathbb{1}$ is the indicator function).

$RPS \in [0, 1]$, where 0 corresponds to a perfect score.

Continuous Ranked Probability Score (CRPS)

For continuous outcomes, we can compute the CRPS of a single data point y_{new} for a forecast with cumulative distribution $F(x)$ by:

$$CRPS(y_{\text{new}}) = \int (F(x) - \mathbb{1}_{y_{\text{new}} \leq x}(x))^2 dx \quad (2.9)$$

A notable difference between the CRPS and the RPS is that the CRPS is not normalised to be between 0 and 1.

While the integral in Eq. (2.9) is challenging to compute analytically, efficient algorithms have been proposed to compute the CRPS from Monte-Carlo samples [128].

2.3 Time-series forecasting

In this section, we review key time-series concepts that we will use throughout this thesis, such as the validation procedure for time-series models, standard models for time-series forecasting that we will use as references to benchmark our models, and state-space models.

2.3.1 Forward chaining

When dealing with time-series data, it is not possible to use standard cross-validation methods to internally validate a model, as it is not reasonable to use future data to predict the past. Instead, we assess the predictive performance of the model in a “forward chaining” setting, aka rolling forecast origin validation [129], where, for example, the model is first trained on the first week of data and tested on the second week of data (first iteration), then trained on the first two weeks of data and tested on the third week of data (second iteration), etc.

Usually, we will start with only the first timepoint in the training set (iteration 0), so that the associated predictions correspond to the prior predictive distribution (in a Bayesian setting). Then, for the i -th iteration and an horizon h (in weeks or days depending of the context), the training set includes timepoints $t \leq i \times h + 1$ and the testing set includes $i \times h + 1 < t \leq (i + 1) \times h + 1$ (Fig. 2.4). In this setting, we generate predictions up to h -steps-ahead in the future, but the predictive performance can be evaluated for shorter prediction horizons as well.

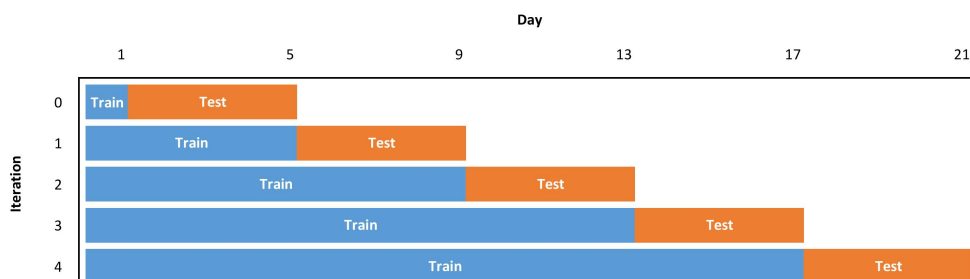


Figure 2.4: Forward chaining with a horizon of 4 days

The forward chaining procedure is computationally demanding, as it involves retraining the model multiple times⁹, and approximations have been proposed to reduce the number of training iterations [130]. However, we do not use these approximations in this thesis, as the training time of the models is not prohibitive because our datasets are small, and we have access to a large number of CPUs to parallelise the forward chaining iterations.

2.3.2 Learning curves

After computing the predictive performance for each prediction, we can compute the average predictive performance for each forward chaining iteration and plot a learning curve (perfor-

⁹Unless when proper Bayesian updating is implemented, for example when using Sequential Monte-Carlo algorithms, or conjugate priors.

mance vs forward chaining iteration or equivalently, training time or training data size). A learning curve allows us to investigate the overall trend of the performance (e.g. determined visually or by computing a smoothed average), as opposed to looking at the performance of a single iteration that may exhibit random variability¹⁰.

Because we work with longitudinal data, the average performance at a given iteration is mostly equivalent to the average performance across patients. However, the population of patients that we averaged across is not always the same for different training iterations due to missing observations. For instance, it is possible that we average across patients 1 and 2 to compute the performance at the i -th iteration and across patients 1 and 3 at the $(i + 1)$ -th iteration. This could cause an issue if the model is consistently better at predicting patient 2 than patients 1 and 3, because the performance at the i -th iteration could be higher than the performance at the $(i + 1)$ -th iteration, even though the model could be learning from the i -th to the $i + 1$ -th iteration. This phenomenon is known as Simpson’s paradox, and occurs because we fail to consider the patient IDs as a confounding factor [126] (an example is given in Chapter 4, Fig. B.1).

Another confounding factor of the predictive performance is the prediction horizon, which may differ on average from one iteration to another. For example, due to irregular measurements, the average prediction horizon may be 2 days at the i -th iteration and 3 days at the $i + 1$ -th iteration. This may result in the performance at the i -th iteration appearing higher than the performance at the $i + 1$ -th iteration, as it is usually easier to predict shorter prediction horizons, even though the model’s performance may actually improve from the i -th to the $i + 1$ -th iterations.

To address these potential confounding issues, we usually attempt to adjust¹¹ for these confounding factors using a “meta-model” of the predictive performance as a function of training iteration, prediction horizon, and patient ID, and report the average performance from the meta-model (details are given in each chapter).

¹⁰This is similar to standard cross-validation techniques such as K-fold, where the average performance across folds is deemed more informative and robust than the performance estimate from a single fold.

¹¹Another option is to use relative measures of performance (e.g. pairwise difference in performance between the model of interest and a reference model). However, these measures are improper, in the sense that an increase in the relative performance is not necessarily associated with an increase in the performance of the model of interest (it could be due to a decrease in performance of the reference model). We looked at such metrics during model development to better “triangulate” and understand the performance of the models, but do not report them in this thesis.

2.3.3 Reference models for time-series forecasting

In this thesis, we compare the performance of our severity prediction models to that of standard time-series forecasting models, including a uniform forecast, a historical forecast, a random walk model, an exponential smoothing model, an autoregressive model of order 1 and a mixed effect autoregressive of order 1 [129]. We did not consider higher-order time-series forecasting models, such as Autoregressive Integrated Moving Average (ARIMA) models, as they are not well suited and difficult to fit with longitudinal data, short time-series or in the presence of missing values (which are characteristics of the data we use in this thesis).

The reference models are reimplemented in a Bayesian context¹² so that they can provide probabilistic predictions instead of point predictions, to allow a fair comparison within Bayesian analyses. We go beyond the off-the-shelf implementation by training the models in a semi-supervised setting where missing values are treated as parameters to be inferred by the model. However, the models remain “naive” in the sense that the likelihood is defined with non-truncated distributions, even though severity scores are bounded, and implicitly treat the scores as continuous, similarly to standard off-the-shelf implementations. As a result, the predictive distributions are truncated for proper evaluation, so that the predictive distribution integrates to one over the support of the score. In the case of discrete variables, predictions are also discretised (rounded to the nearest integer).

The following describes the likelihood of the reference models considered in this thesis. Their priors are defined in the individual chapters where they are used, as they are chosen to be consistent with the other models they are compared to.

Uniform forecast

A uniform forecast assumes that observations $y \in [0, M]$ follow a uniform distribution (discrete or continuous depending on the outcome):

$$y(t+1) \sim \mathcal{U}(0, M) \tag{2.10}$$

For discrete variables, this means $\forall i \in [0, M], P(y(t+1) = i) = \frac{1}{M+1}$.

¹²Except for the uniform and historical forecasts.

Historical forecast

A historical forecast model makes heuristic forecasts based on the prevalence of past observations, at the population level¹³.

In the discrete case, a probability table can be calculated. For instance, if 10% of observations in the training set are 0, then $P(y(t+1) = 0) = 0.1$.

For the continuous case, it is not possible to compute a probability table from past observations. Instead, we compute the performance metrics associated with this forecast directly by using kernel density estimates for computing the lpd and considering the training set as Monte-Carlo samples for computing the CRPS (cf. Section 2.2.5).

Random walk model

A random walk model provides a flat forecast, i.e. a forecast centred on the last observation with the uncertainty quantified by a variance parameter σ^2 , and is described by:

$$y(t+1) \sim \mathcal{N}(y(t), \sigma^2) \quad (2.11)$$

Exponential smoothing model

An exponential smoothing model corresponds to an exponential smoothing of the data and a flat forecast, and is described by:

$$l(t) = \phi y(t) + (1 - \phi)l(t-1) \quad (2.12)$$

$$y(t+1) \sim \mathcal{N}(l(t), \sigma^2) \quad (2.13)$$

where l ¹⁴ represents the smoothed values (the level, which is also the predictor for y : $l(t) = \hat{y}(t+1)$) and $\phi \in [0, 1]$ is the smoothing factor, which can be related to the time constant τ of the process and the delay ΔT between two observations (e.g. one day) by:

$$\phi = 1 - e^{-\frac{\Delta T}{\tau}} \iff \tau = -\frac{\Delta T}{\log(1 - \phi)} \quad (2.14)$$

¹³The historical forecast could be made patient-dependent given enough data, especially if the number of categories is small (in the discrete case). However, we deemed that the patients' time-series used in this study are too short to obtain stable estimates of a patient-dependent historical forecast.

¹⁴We deviate from our convention to be consistent with the time-series literature: l is a function of data and parameters.

Autoregressive model

An autoregressive model of order 1 is an extension of the “random walk” model with an autocorrelation coefficient α , and an intercept β_0 , and defined by:

$$y(t + 1) \sim \mathcal{N}(\alpha y(t) + \beta_0, \sigma^2) \quad (2.15)$$

We assume stationarity and set $\beta_0 = (1 - \alpha)y_\infty$, where y_∞ is the expected value of the series.

Mixed effect autoregressive model

A mixed autoregressive model is an extension of the “Autoregressive model”, by assuming patient-dependence for the autocorrelation $\alpha^{(k)}$ and the intercept $\beta_0^{(k)} = (1 - \alpha^{(k)})y_\infty^{(k)}$. The model is described by:

$$y^{(k)}(t + 1) \sim \mathcal{N}(\alpha^{(k)}y^{(k)}(t) + \beta_0^{(k)}, \sigma^2) \quad (2.16)$$

The patient-dependent parameters are partially pooled using hierarchical priors (details in the relevant chapters).

2.3.4 State-space models

State-space models (SSM) constitute a class of models that are useful for time-series forecasting. We will use SSM in Chapters 4, 6, 7 and 8.

SSM are discrete-time stochastic models that assume the existence of latent (i.e. unobserved) states, which follow their own dynamics, and from which the measurements/observations are obtained (Fig. 2.5). For example, in the context of modelling the evolution of AD severity discussed in this thesis, a state-space model can be understood as if the severity scores are the imperfect measurements of a true latent severity, which follows its own latent dynamic.

SSM are thus described by two sets of equations: the state equation that specifies how the latent process transitions in time (the latent dynamic), and the observation equation that describes how the measurements are obtained from the latent states (the measurement process).

SSM are similar to Hidden Markov Models (HMM), in that HMM often refers to models where the latent states are discrete, whereas the term SSM is often used when the latent states

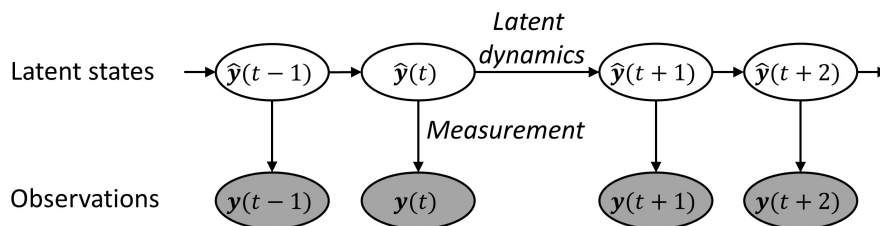


Figure 2.5: Schematic of a state-space model

are continuous. In addition, dynamic linear models¹⁵ are a special case of SSM, where the latent dynamics and measurement processes are linear and follow (multivariate) normal distributions.

SSM are especially useful when the measurements are noisy, since predictions are not directly based on the measurements (as in an autoregressive model, for example) but on the states, which can be seen as a smoothed version of the measurements and may be more reliable. In addition, assuming the data is missing at random¹⁶, missing values can be treated elegantly by specifying that the measurement at the corresponding timepoint is absent and leaving the latent dynamic untouched.

2.4 Ordered logistic distribution

In this section, we review the ordered logistic distribution, which we will use in Chapters 5, 7 and 8 to model the measurement processes of AD severity items.

For a discrete ordinal outcome $y \in \{0, \dots, M\}$ ($M + 1$ categories), the probability mass function of the ordered logistic distribution is defined by:

$$\text{OrderedLogistic}(y|\eta, \mathbf{c}) = \begin{cases} 1 - \text{logit}^{-1}(\eta - c_1) & \text{if } y = 0 \\ \text{logit}^{-1}(\eta - c_y) - \text{logit}^{-1}(\eta - c_{y+1}) & \text{if } 0 < y < M \\ \text{logit}^{-1}(\eta - c_M) & \text{if } y = M \end{cases} \quad (2.17)$$

η is the location of the distribution¹⁷ and \mathbf{c} is a vector (of size M) of cut-offs for the distribution.

The ordered logistic distribution can be understood as a logistic distribution (the distribution for which the cumulative distribution is a logistic function) with location η and scale 1, being

¹⁵Including the Kalman filter, an algorithm performing the recursive estimation of a dynamic linear model.

¹⁶For example if the missing severity is not the cause of the missingness. If the data is not missing at random, we would need to model the cause of the missingness.

¹⁷In a regression setting, the location would usually correspond to the linear predictor $X\beta$.

discretised using cut-offs \mathbf{c} . As such, the probability of observing y is equal to the area under the logistic distribution between the y -th and the $y + 1$ -th cut-offs¹⁸ (Fig. 2.6). Eq. (2.17) can thus be written as:

$$\begin{aligned}
 & y \sim \text{OrderedLogistic}(\eta, \mathbf{c}) \\
 \iff & \hat{y} \sim \text{Logistic}(\eta, 1) \ \& \ \begin{cases} y = 0 & \text{if } \hat{y} < c_1 \\ y = i & \text{if } c_i < \hat{y} < c_{i+1} \text{ for } 0 < i < M \\ y = M & \text{if } \hat{y} > c_M \end{cases} \quad (2.18)
 \end{aligned}$$

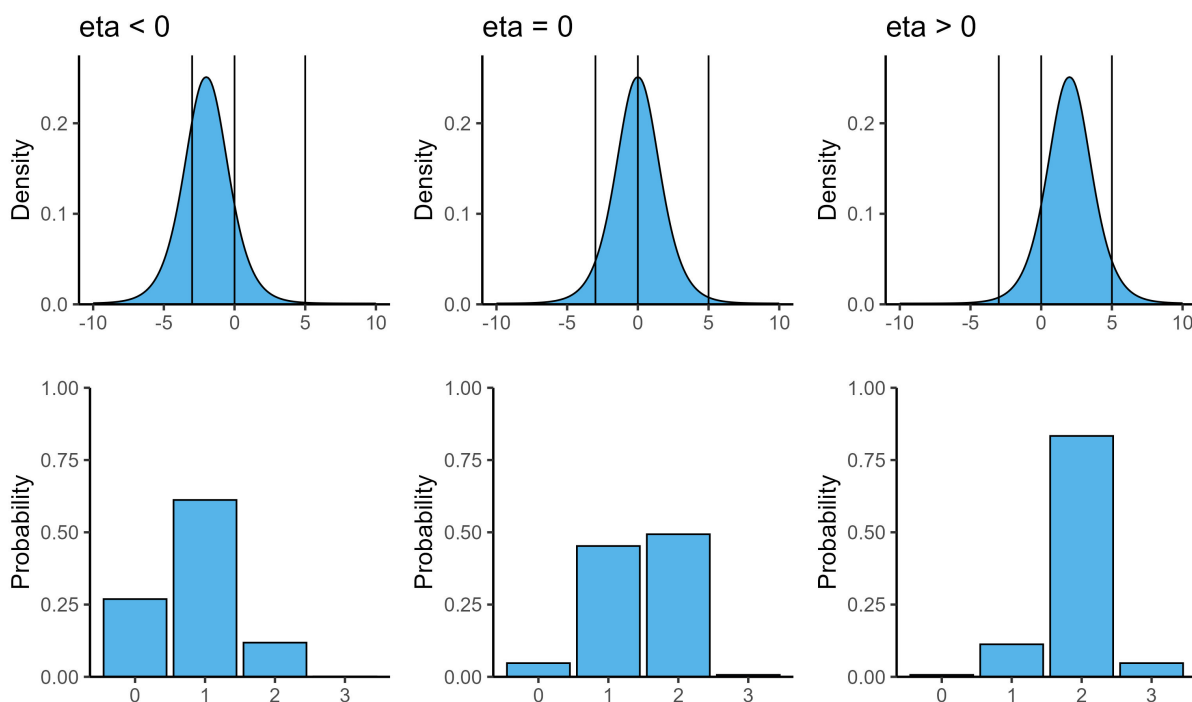


Figure 2.6: Illustration of an ordered logistic distribution. Top: logistic distribution with location η and scale 1. Vertical lines represent cut-off values. Bottom: Corresponding ordered logistic distribution.

Thresholding a normal distribution instead of a logistic distribution results in an ordered probit distribution. Since the logistic distribution has a similar shape to a normal distribution (with slightly fatter tails) (Fig. 2.6), the ordered logistic distribution is similar to the ordered probit distribution, in the same way a logistic regression is similar to a probit regression. In practice, we found that using an ordered logistic distribution was more robust computationally than using an ordered probit distribution, in our situation.

We can intuit from Fig. 2.6 that the ordered logistic distribution is invariant if η and \mathbf{c}

¹⁸The cut-offs vector \mathbf{c} can be appended with $c_0 = -\infty$ and $c_{M+1} = +\infty$ for the sake of the argument. In that view, the observations $y = 0$ or $y = M$ are considered censored, which means that we interpret $y = 0$ as $y < 1$ (e.g. for intensity signs, any intensity value that is less than mild or cannot be detected), and $y = M$ as $y > M - 1$ (e.g. for intensity signs, any intensity value that is more severe than a “moderate” severity).

are translated by a scalar λ : $\text{OrderedLogistic}(y|\eta + \lambda, \mathbf{c} + \lambda) = \text{OrderedLogistic}(y|\eta, \mathbf{c})$. As a result, if η contains an intercept, one of the c_i can be set to an arbitrary value, e.g. $c_0 = 0$. Alternatively, η does not need an intercept if \mathbf{c} is not anchored. We can also intuit a scale invariance: if we scale η and \mathbf{c} by a scalar $\frac{1}{s}$, then $\text{OrderedLogistic}(y|\frac{\eta}{s}, \frac{1}{s}\mathbf{c})$ is equivalent to thresholding a logistic distribution with location η and scale s by the vector \mathbf{c} . This also means that the variance of the ordered logistic distribution is controlled by the range of \mathbf{c} . We will use this property to obtain a more interpretable parametrisation of the distribution in Chapter 7.

It is worth noting that fitting an ordered logistic distribution is equivalent to fitting logistic regressions to the cumulative distribution function of y , assuming proportional odds¹⁹, since:

$$y \sim \text{OrderedLogistic}(\eta, \mathbf{c}) \tag{2.19}$$

$$\iff \forall i \in \{0, \dots, M\}, P(y \leq i) = 1 - \text{logit}^{-1}(\eta - c_{y+1}) = \text{logit}^{-1}(-(\eta - c_{y+1})) \tag{2.20}$$

Here the cut-offs correspond to the intercepts of the logistic regressions and the proportional odds assumption implies that η is the same for all logistic regressions. Thus, the ordered logistic distribution reduces the complexity of fitting different models for the different cumulative outcomes by having a single interpretable location parameter η . At the same time, unlike other discrete distributions such as the binomial distribution, the ordered logistic distribution has a flexible shape controlled by the vector of cut-offs \mathbf{c} , which notably controls the variance of the distribution. This is in contrast with discretising the predictions of reference models to the nearest integer, for example, which implies fixed and equally spaced cut-offs.

¹⁹An ordered logistic regression is often called a proportional odds model.

Chapter 3

Automating the assessment of AD severity

We start our journey by investigating how we can automate the collection of AD severity measurements, by relying on computer vision algorithms that could assess AD severity from camera images. Currently, AD severity is measured infrequently when patients visit a clinic, or is based on less reliable self-assessments. Collecting accurate and frequent AD severity measurements is nonetheless critical to study and develop predictive models of the evolution of AD severity, and later to deploy such tools to a wide audience.

This chapter is adapted from our paper “EczemaNet: Automating Detection and Severity Assessment of Atopic Dermatitis”, presented at the *Machine Learning and Medical Imaging* conference in 2020 and published in its proceedings [59]. Reproduction of this paper in this thesis was granted under the Springer Nature License number 5167081192515. The code written for this project is available at <https://github.com/Tanaka-Group/EczemaNet>.

This research project is a collaborative work with Kevin Pan (KP), Kai Arulkumaran (KA), and Prof. Hywel C. Williams (HW). My main contributions in this project were its conceptualisation, design of models and algorithms, and validation. KP conducted the formal analysis, wrote the computer code, and contributed to the design of models. KA helped formalise the different experiments and write the manuscript. Our clinical collaborator (HW) contributed the data and insights on the clinical relevance of our model.

3.1 Introduction

Atopic Dermatitis is characterised by recurrent skin inflammation that can severely impact patients' lifestyles, with detrimental effects on social, academic, and occupational aspects of their lives. While current treatments aim to manage dynamic and unpredictable fluctuations of AD symptoms, only 24% of patients and caregivers feel confident that they can manage AD symptoms adequately [8]. Automating the evaluation of AD severity would allow us to assist research into the disease and enable patients to become more involved in the management of their condition. Remote assessment of AD symptoms by automated evaluation would enhance data-enabled efficient clinical trials by reducing the burden of parties involved and minimising detection bias in clinical trials that test interventions.

Several clinical scores are commonly used to grade the severity of AD, including SASSAD [83], TISS [84], and EASI [78] scores, the latter of which is recommended by the HOME organisation [79]. Each of these are defined according to a combination of the severity of 7 disease signs¹ (Fig. 3.1): cracking (Cra.), dryness (Dry.), erythema (Ery.), excoriation (Exc.), exudation (Exu.), lichenification (Lic.) and oedema (Oed.). However, due to the lack of sufficient clinical training materials and the non-intuitive nature of some disease signs (e.g., “dryness” versus “cracking”), inter- and intra-rater reliability are poor [82]. Our goal is to improve the reliability of these scoring systems through computer-aided evaluation of the different disease signs.

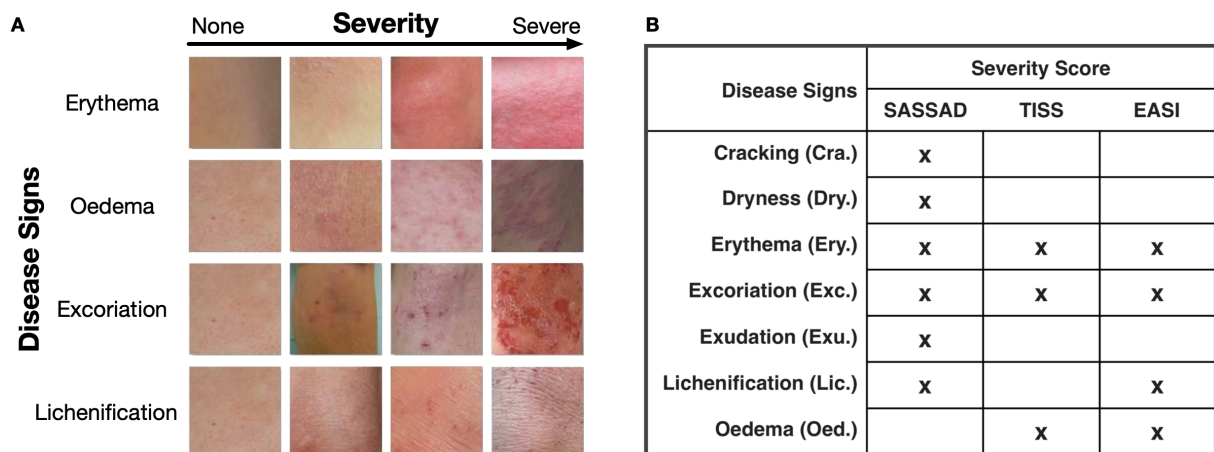


Figure 3.1: Disease signs and their relationship to severity scores. A) Examples of the 4 disease signs associated with EASI. Reproduced from [78]. B) A list of disease signs used for calculating SASSAD, TISS and EASI.

In recent years, machine-learning-based methods using convolutional neural networks (CNNs) have reached dermatologist-level performance on classifying skin cancers [131] [132]. However, due to the lack of standardised clinical datasets beyond skin cancer, applications of

¹As well as the area of the affected region in the case of EASI.

CNNs for non-cancerous diseases have mostly been limited to automatic disease diagnosis of skin lesions [133] [134] [135]. Whether a lesion can be attributed to AD is of limited value to already diagnosed patients, and does not address the important challenge of assessing the overall severity of the disease, whose lesions are spatially distributed over the entire body and can exhibit multiple signs of varying intensities.

In this paper, we introduce a novel computer vision pipeline, EczemaNet, that is capable of detecting and evaluating the severity of AD from camera images. In comparison to prior work [136], we use deep learning to learn relevant features from the data (as opposed to hand-engineered features), produce probabilistic predictions, and evaluate our method on a far larger dataset. Our pipeline uses CNNs to first detect regions-of-interest (RoI) from an image to make image crops, and then evaluate the severity of the 7 disease signs in each crop. Our input images often include background, clothes, etc. while most pipelines expect closely cropped images [137]. Similarly to recent work on psoriatic plaque severity assessment [138], we use ordinal classification to predict the severity of multiple disease signs simultaneously. However, we also propagate the uncertainties over these predictions to produce a final set of severity scores (SASSAD, TISS, and EASI) simultaneously, and show that using multiple crops and probabilistic predictions allows us to make well-calibrated predictions with low root mean squared error (RMSE). These properties make EczemaNet a promising proof-of-concept for the use of CNNs in clinical trials, with downstream applications in personalised therapies for AD.

3.2 Data

Our data originates from the Softened Water Eczema Trial (SWET), which is a randomised controlled trial of 12 weeks duration followed by a 4-week crossover period, for 310 AD children aged from 6 months to 16 years [139] [140]. The original data contains 1393 photos of representative AD regions taken during their clinic visits, along with the corresponding severity of each disease sign. During each visit, a disease assessment was made for SASSAD and TISS, using the 7 disease signs labelled for each image. The severity of each sign was determined on an ordinal scale: *none* (0), *mild* (1), *moderate* (2), or *severe* (3).

The photos vary both in resolution and subjective quality, such as focus, lighting, and blur. In addition, as the photos can contain significant areas of background or areas that are otherwise irrelevant for diagnosis, we manually curated 962 of the original photos, generating 1748 image crops of representative diseased regions by visual inspection². We used these crops

²RoI, of arbitrary size, were labelled by three non-expert volunteers given a set of 50 expert-labelled images for which the lesions were identified, where one of the volunteer was instructed directly by a clinical expert. 431 photos were deemed difficult to label by the volunteers and hence left out of our dataset.

to fine-tune an RoI detection network, and then bootstrapped our dataset by running this network on all images, extracting a further 2178 image crops. Both sets of image crops were then combined and paired with the labels for the 7 disease signs, resulting in a final dataset of 933 diagnoses from 285 patients (78% of declared “white” ethnicity), including 1237 original photos with corresponding 3926 image crops³.

This final dataset was used to train our severity prediction network (Section 3.3.2). All crops were labelled with the overall diagnosis for the entire image, as we did not have labels for the individual crops. Despite this noisy labelling, the use of RoI detection and severity prediction in EczemaNet led to better performance than using the entire image (Section 3.4.2).

3.3 Methods

Our EczemaNet pipeline consists of detecting RoI, making probabilistic predictions on all 7 disease signs over all crops simultaneously, and then combining these to predict the AD severity scores per image (Fig. 3.2). We made heavy use of transfer learning [141] to train on our medium-size dataset successfully: we fine-tuned both our RoI detection and severity prediction CNNs. The RoI detection was trained first, as otherwise it would not be able to provide relevant crops for the severity prediction network for end-to-end training. We used TensorFlow [142] for training and evaluation, starting with pretrained models in TensorFlow.

3.3.1 Region of Interest detection

Following the speed/memory/accuracy model selection guidelines from [143], we chose the Faster R-CNN model [144] to perform RoI detection for diseased areas.

3.3.2 Severity prediction

Our severity prediction pipeline is composed of a pretrained CNN base and 7 fully-connected neural networks (FCNNs), each of which predicts the severity of one of the 7 disease signs. We reflect the ordinal nature of the labels by training the FCNNs with ordinal classification. The predicted severities are averaged over all crops to calculate a probabilistic distribution of the severity of each disease sign for the image. Finally, the predictions for the disease signs are

³The full data pipeline is provided in Fig. A.1

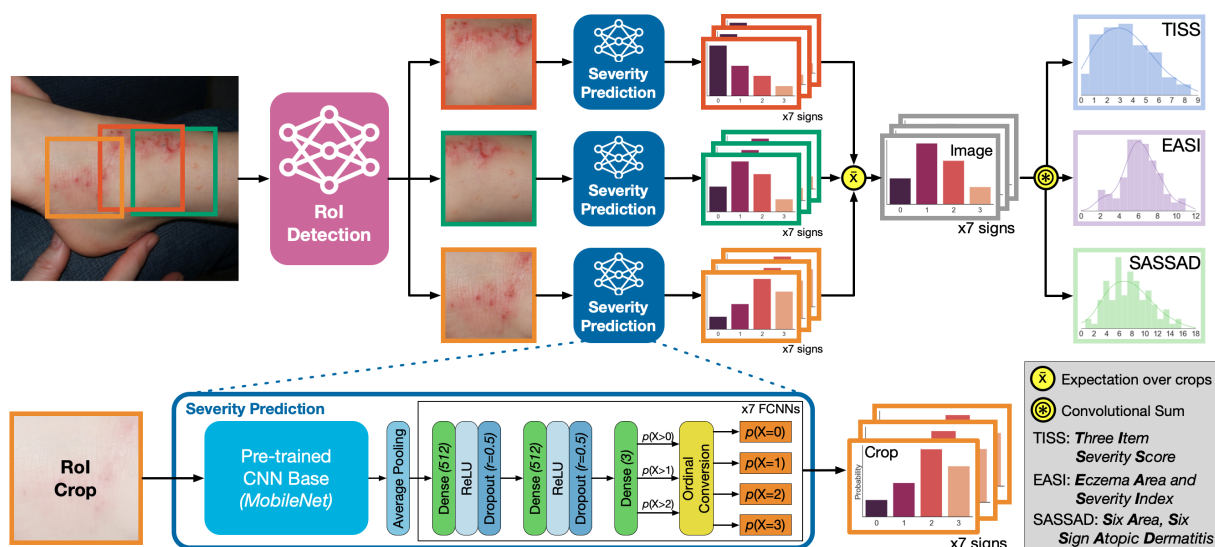


Figure 3.2: EczemaNet overview. The RoI detection network extracts crops from an image. The severity prediction network makes probabilistic predictions for each disease sign in each crop. The averaged prediction over crops are then combined to form the final probabilistic prediction of the severity scores for the image.

combined to produce a probability distribution of the regional⁴ severity scores (SASSAD, TISS, and EASI) per image. Here we describe characteristic features of EczemaNet in more detail.

Pretrained CNN base: Our base consists of all convolutional and pooling layers within MobileNet [145].

Separate FCNNs: We use separate FCNNs per disease sign, as opposed to using one FCNN to predict all disease signs simultaneously.

Ordinal Classification: Instead of predicting the 4 severities independently for each sign as a 4-way classification, as is typically done, we model them using ordinal classification, which better reflects the ordinal nature of the severity measurements⁵. To predict the classes of X for the diagnoses none ($X = 0$), mild ($X = 1$), moderate ($X = 2$) and severe ($X = 3$), we train 3 binary classifiers to output the probabilities, $p_0 = P(X > 0)$, $p_1 = P(X > 1)$ and $p_2 = P(X > 2)$. These probabilities are then converted into class probabilities for outcome X

⁴In practice, EASI and SASSAD are assessed across different regions of the body, which we do not consider in this work.

⁵AD severity is likely a continuous trait in reality, with the discrete nature of the scores being an artefact of the measurement process.

using a modification of Frank & Hall’s method [146] with dependent classifiers [147]:

$$P(X = 0) = 1 - p_0 \tag{3.1}$$

$$P(X = 1) = p_0 (1 - p_1) \tag{3.2}$$

$$P(X = 2) = p_0 p_1 (1 - p_2) \tag{3.3}$$

$$P(X = 3) = p_0 p_1 p_2 \tag{3.4}$$

Expectation over Crops: We produce a single set of severity predictions for each disease sign over the entire image, by averaging the predictions over all crops⁶. Despite the high overlap between most crops, similarly to test-time data augmentation [148], we found that averaging over crops improved both accuracy and calibration (Section 3.4.2).

Multitask prediction: All 3 regional severity scores (TISS, EASI, SASSAD) are sums of subsets of the 7 disease signs (Fig. 3.1B). While it is possible to directly predict each of the regional severity scores, we treat prediction as a multitask problem, predicting the severity of all disease signs simultaneously, and then sum them⁷ to calculate the final regional severity scores.

3.4 Experiments and evaluation

Inference for a single image on CPU (Intel i9-9980HK) took 15.6s for the detection network and 1.6s for the severity prediction network. Our work is a proof-of-concept, and could feasibly run on a smartphone in a few seconds with, e.g., model compression techniques.

3.4.1 Region of Interest detection

We fine-tuned a pre-trained Faster R-CNN model using the 962 manually curated original photos. With a train/validation/test ratio of 60:20:20, the manually curated photos were randomly split into 578:192:192 photos. It resulted in 1069:378:346 corresponding image crops, as each photo can contain a different number of image crops. The model was trained for 10^5 steps with a batch size of 1, using stochastic gradient decent (SGD) with momentum = 0.9, with an initial learning rate of 3×10^{-4} , dropped to 3×10^{-5} after 90000 steps; no data augmentation was used. We weighted the localisation loss by a factor of 1.2, as our focus was to improve detection,

⁶Crops were preprocessed by bilinearly resampling to 224×224 px.

⁷We convolve the probability mass functions of the predicted severity of the 7 disease signs, assuming that the predictions are independent random variables.

rather than classification by Faster R-CNN, which was trained to detect the presence of AD.

We evaluated our model using the average precision (AP) score, the standard measure in object detection. The AP score measures the intersection between the ground truth and predicted boundaries, with a default overlap threshold of 50%. After tuning the hyperparameters on our validation set, we tested our model using the test set of 192 images and obtained the AP score of 40.15%. We also performed a more qualitative evaluation to validate our trained model, and estimated that our model achieved a 10% false positive rate per image. We therefore concluded that our RoI detection network could generalise sufficiently well, and used it to extract more crops from the original data (Section 3.2).

3.4.2 Severity prediction

We combined a pre-trained MobileNet with 7 separate randomly initialised FCNNs (for each disease sign), and trained all parameters to predict the severity of the 7 disease signs on the final pre-processed dataset, which contained 933 diagnoses from 285 patients, including 1237 original photos with 3926 corresponding image crops. We used 10-fold cross-validation with a 90:10 train/test split, stratified on patients, to train and assess severity prediction models. The models were trained for a maximum of 50 epochs (using early stopping) with a batch size of 32, using SGD with a learning rate of 1×10^{-4} and momentum = 0.9; no data augmentation was used. Dropout with $p = 0.5$ and a max L_2 -norm weight constraint with $c = 3$ were used to regularise all fully-connected layers [149]. To combat severe class imbalance, we weighted all prediction losses by the inverse of the empirical class probabilities.

We evaluated RMSE on EASI (the recommended severity score [79]) for EczemaNet (1.929 ± 0.019) and for its variations listed below to confirm the use of each characteristic aspect of our model design (Fig. 3.3A and Table 3.1).

Pretrained CNN Base: The choice of pretrained CNN base significantly impacts the performance of the prediction model. We evaluated a range of commonly used CNN architectures for the base: Inception-v3 [150], MobileNet [145], ResNet-50 [151], VGG-16, and VGG-19 [152]. Only EczemaNet with MobileNet consistently achieved an RMSE on EASI of < 2 including standard error (Fig. A.3).

Bootstrapped Dataset: Training EczemaNet with the 1748 manually labelled crops, plus the 2178 additional crops automatically extracted by our trained RoI detection network, achieved the lowest RMSE across all of our experimental conditions (1.929 ± 0.019), compared to 2.003 ± 0.024 when EczemaNet was trained with only the manually labelled crops.

Model architectures: We used a set of baselines (baseline and intercept-only) and ablations (listed in order of performance, Fig. 3.3A and A.2):

EczemaNet	Our full model.
-Ordinal	4-way categorical classification <i>vs.</i> ordinal classification.
+Interaction	Sign interaction added by concatenating FCNN features <i>vs.</i> separate FCNN per sign.
-Separate FCNNs	A single FCNN for all 7 signs <i>vs.</i> separate FCNNs per sign.
-Crops	Using the entire image <i>vs.</i> averaging predictions over crops.
-Pretrained	Starting with random CNN weights <i>vs.</i> pretrained CNN weights.
Intercept-only	Predicting the average EASI in the training set.
Baseline	Predicting EASI from the whole image.
-Multitask	Predicting EASI directly <i>vs.</i> summing predicted disease signs.

The full EczemaNet performs best, although some components have a lesser effect on the RMSE on EASI⁸ (Fig. 3.3A). In reverse order, multitask learning is the most important modelling choice, which possibly mitigates overfitting. The baseline model, which is a naive CNN-based approach, using regression on the whole image, performs almost the same as the intercept-only model, indicating the difficulty of our problem. Using pre-trained weights and averaging over crops also play a large role in the good predictive performance of EczemaNet. Sharing FCNN parameters when modelling the 7 disease signs hurts performance slightly, perhaps due to interference between the 7 tasks. Finally, ordinal classification provides a small boost over categorical classification.

The coverage of EczemaNet (Fig. 3.3B) indicates well-calibrated prediction intervals for a neural network [153]. The performance could be further improved by post-processing, such as quantile calibration, to make the predictive distribution sharper at the mode and with longer tails.

Achieving high accuracy on the regional severity scores is a major aim of our work for clinical relevance. It is also important to examine other metrics as well, particularly because of the class imbalance in the data. We calculated F_1 scores and Ranked Probability Scores (RPS) for all disease signs for all models that predict all 7 disease signs (Table 3.1). The F_1 score is the harmonic mean of precision and recall (sensitivity, true positive rate), and hence is less sensitive to class imbalance than recall. RPS is a strictly proper scoring rule and measures the calibration of ordinal forecasts. We observed approximately the same ranking of baselines/ablations as for RMSE on EASI, with no clear outliers, supporting our earlier assessment on their relative importance.

⁸We also observed a similar ranking across models for SASSAD (Fig. A.4) and TISS (Fig. A.5), as well as across the individual signs.

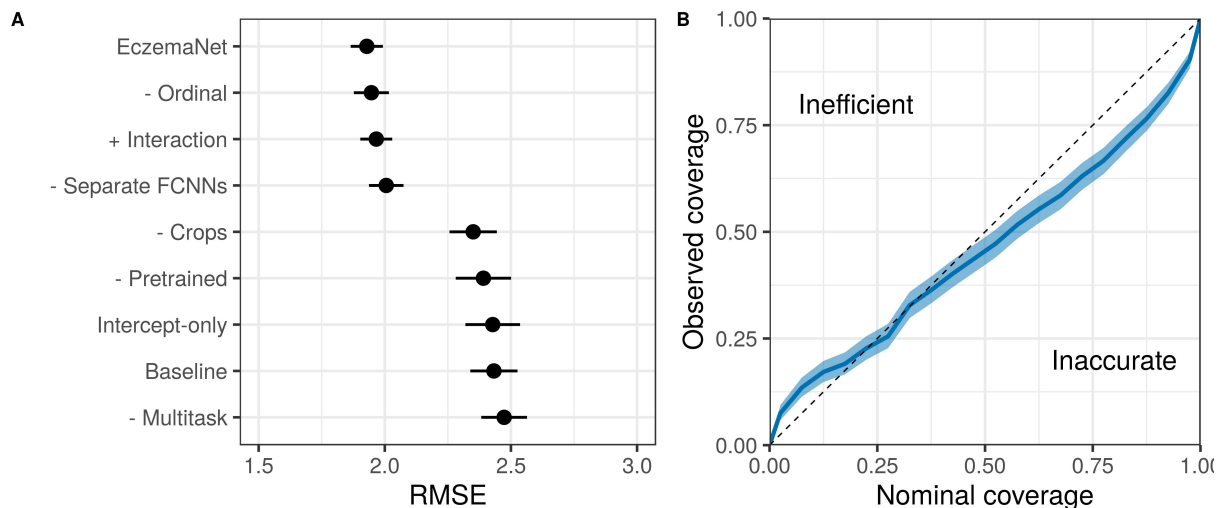


Figure 3.3: EczemaNet predictive performance. A) RMSE (mean \pm 1 standard error over cross-validation) on EASI across models. B) EASI calibration of highest density prediction intervals (coverage).

Table 3.1: Results of experiments in terms of F_1 score (top; \uparrow is better) and RPS (bottom; \downarrow is better) for all 7 disease signs. Mean \pm 1 standard error over cross-validation.

Model	Cra.	Dry.	Ery.	Exc.	Exu.	Lic.	Oed.
Full	0.707 \pm 0.013	0.443 \pm 0.006	0.419 \pm 0.004	0.480 \pm 0.007	0.769 \pm 0.008	0.404 \pm 0.005	0.694 \pm 0.007
-Pretrained	0.671 \pm 0.013	0.242 \pm 0.008	0.250 \pm 0.009	0.269 \pm 0.004	0.759 \pm 0.008	0.234 \pm 0.003	0.694 \pm 0.007
-Separate FCNNs	0.696 \pm 0.013	0.422 \pm 0.006	0.405 \pm 0.006	0.473 \pm 0.005	0.768 \pm 0.008	0.390 \pm 0.005	0.690 \pm 0.007
+Interaction	0.704 \pm 0.013	0.454 \pm 0.008	0.437 \pm 0.005	0.491 \pm 0.007	0.767 \pm 0.008	0.388 \pm 0.007	0.697 \pm 0.007
-Ordinal	0.696 \pm 0.013	0.453 \pm 0.006	0.428 \pm 0.007	0.470 \pm 0.004	0.772 \pm 0.008	0.404 \pm 0.006	0.692 \pm 0.008
-Crops	0.686 \pm 0.012	0.369 \pm 0.004	0.370 \pm 0.006	0.289 \pm 0.007	0.765 \pm 0.007	0.317 \pm 0.006	0.700 \pm 0.007
Full	0.076 \pm 0.003	0.136 \pm 0.001	0.137 \pm 0.001	0.128 \pm 0.001	0.056 \pm 0.002	0.151 \pm 0.002	0.077 \pm 0.002
-Pretrained	0.098 \pm 0.003	0.164 \pm 0.001	0.160 \pm 0.002	0.178 \pm 0.001	0.077 \pm 0.002	0.181 \pm 0.001	0.085 \pm 0.002
-Separate FCNNs	0.080 \pm 0.003	0.140 \pm 0.001	0.142 \pm 0.001	0.132 \pm 0.001	0.057 \pm 0.002	0.156 \pm 0.002	0.079 \pm 0.002
+Interaction	0.080 \pm 0.003	0.141 \pm 0.001	0.141 \pm 0.001	0.131 \pm 0.002	0.056 \pm 0.002	0.156 \pm 0.002	0.079 \pm 0.002
-Ordinal	0.079 \pm 0.003	0.139 \pm 0.001	0.136 \pm 0.001	0.130 \pm 0.001	0.055 \pm 0.002	0.149 \pm 0.001	0.079 \pm 0.002
-Crops	0.083 \pm 0.004	0.154 \pm 0.001	0.155 \pm 0.002	0.163 \pm 0.002	0.063 \pm 0.002	0.165 \pm 0.002	0.081 \pm 0.002

3.5 Discussion

This chapter presented EczemaNet, a CNN-based pipeline for evaluating eczema severity directly from camera images. EczemaNet consists of an RoI detection network, which extracts relevant crops from each image, and a severity prediction network, which predicts the severity of 7 disease signs for each crop. The probability distributions of severities are averaged over crops, and then combined to form a prediction of the 3 regional severity scores. EczemaNet achieved fair performance⁹ on a medium-size clinical dataset and demonstrated well-calibrated prediction intervals. These results present a step towards standardising the evaluation of objective AD severity scores for diverse dermatological research purposes, and could be applied to similar conditions, such as psoriasis.

⁹The RMSE for predicting regional EASI $\in [0, 12]$ was only ≈ 1.9 compared to 2.5 for an intercept-only model.

Multiple sources of systematic errors, aside from random errors, can be considered to limit the performance of EczemaNet: mislabelling due to inter- and intra-rater variability, discretisation error over a continuous outcome (severity), and errors arising from partial/noisy information (e.g., tactile diagnoses, out-of-focus images). While it is difficult to evaluate the effects of all these sources of errors, Monte Carlo simulations of the measurement process suggested that rounding error alone could account for an RMSE of 0.6, which makes it unclear how much performance could be further improved on EczemaNet trained with our data.

Limitations of this study include potential biases in the validation of the severity model due to model selection overfitting. Ideally, unbiased estimates of performance for EczemaNet should be obtained using an independent test set. This was not implemented because of the small size of the data available for training, and because this study was mostly exploratory. Nonetheless, we attempted to mitigate potential biases by quantifying uncertainty in performance estimates using cross-validation and by only considering a small number of alternative models (five pre-trained CNN bases and six ablations). EczemaNet may also suffer from skin colour biases, since the SWET dataset mostly contained images of white skin tones [154]. It would be desirable to collect images of eczema for a variety of skin tones, as it is known that the presentation of disease signs differs between skin tones. For example, erythema is more likely to appear violaceous or dark brown in darker skin tones [155].

A natural extension of the work presented here is to move beyond regional severity scores to predicting the overall severity scores. For a given area, EASI is the product of the intensity score (which we currently predict), and the area score. The area score could be predicted simultaneously with the intensity scores given the box labels identified by our RoI network. We encourage future clinical trials to collect and share richer labels, such as pixel-level segmentation, to increase the breadth of tasks, such as segmentation, that can be automated using machine learning. Future work could also involve data augmentation while tackling the issue of severity class imbalance.

3.6 Afterword

In this section, we briefly allude to follow-ups of this work after its publication, by our group and others.

At the time of writing and publication of the paper in 2020 [59], this work was the first exploring computer vision algorithms to automatically assess eczema severity, to the best of our knowledge. We have identified two similar studies, using somewhat simpler approaches, that have been published since then.

- In [156], off-the-shelf computer vision algorithms were used to diagnose different types of eczema, rather than for severity assessments.
- In [157], off-the-shelf computer vision algorithms were explored to predict the severity of erythema, papulation, excoriation and lichenification. The authors of this study reported a good performance, but it is unclear how these results would generalise to a real-world setting. For example, the authors of [157] used manually processed images (crops) as input rather than camera images, used for EczemaNet. Comparison with our results is also difficult because the severity classes were balanced before training and testing the CNNs, uncertainty was not quantified, and proper scoring rules were not calculated in [157]. In addition, the validation procedure in [157] is at times unclear. For example, we do not know whether the training and testing sets were stratified by patients.

As a follow-up to the development of EczemaNet presented in this chapter, we decided to focus on the eczema detection algorithm. First, we investigated whether it was possible to accurately identify eczema lesions in digital images. We asked four dermatologists to independently segment AD lesions in 80 images, and found that the degree of agreement between raters (inter-rater reliability) was poor (average intra-class correlation coefficient across images of 0.45, $SE = 0.04$). This result suggested that algorithms relying on AD segmentation data (including crops) may be subject to biases. At the time of the final submission of this thesis (May 2022), this result was summarised in paper (in press at *JID Innovations*), “Detecting eczema areas in digital images: an impossible task?”, by Guillem Hurault, Kevin Pan, Ricardo Mokthari, Bayanne Olabi, Eleanor Earp, Lloyd Steele, Hywel C. Williams and Reiko J. Tanaka.

Second, we explored possible improvements to EczemaNet, using data augmentation to improve the generalisability of the eczema detection algorithm, and choosing to rely on skin segmentation only rather than AD crops to avoid using unreliable segmentation data. We also investigated the interpretability and adversarial robustness of EczemaNet. While these two aspects are improved compared to the original EczemaNet presented in this chapter, we did not detect practically significant improvements in the predictive performance of AD severity scores. We interpret this result as an additional confirmation that the task of assessing the severity of eczema lesions with camera images is difficult. As such, we tend to believe that collecting more and better quality data (images and labels) would surpass the gains in performance from using a cleverer algorithm [158]. At the time of the final submission of this thesis (May 2022), this work was summarised in a paper (under review) entitled “Reliable detection of eczema areas for fully automated assessment of eczema severity from digital camera images”, by Rahman Attar, Guillem Hurault, Zihao Wang, Ricardo Mokhtari, Kevin Pan, Bayanne Olabi, Eleanor Earp, Lloyd Steele, Hywel C. Williams, and Reiko J. Tanaka.

Chapter 4

A statistical model to predict AD severity

In this chapter, we develop the first model, to the best of our knowledge, predicting the evolution of eczema severity. The model is remotely inspired by a mathematical model of AD pathogenesis previously developed by our group [159]. This proof-of-concept model allows us to explore the opportunities and challenges of AD severity prediction.

This chapter is adapted from our paper “Personalized prediction of daily eczema severity scores using a mechanistic machine learning model”, published in 2020 in *Clinical and Experimental Allergy* [60], under the terms of the [Creative Commons CC BY license](#). The code written for this project is available at <https://github.com/ghurault/mbml-eczema>.

This research project is a collaborative work with Dr. Elisa Domínguez-Hüttinger (EDH), Prof. Sinéad M. Langan (SL), and Prof. Hywel C. Williams (HW). EDH contributed her insights on the previously published mathematical model [159] and helped review and edit the manuscript. Our clinical collaborators (SL and HW) contributed the data and insights on the clinical relevance of the model. We also acknowledge Prof. Kim S. Thomas for sharing the SWET dataset and constructive comments on the manuscript.

4.1 Introduction

AD typically has a fluctuating course characterised 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. Personalised treatment strategies may be more beneficial to individual patients rather than a

“one-size-fits-all” approach to therapy [10] [11]. A first step toward developing personalised treatment strategies is to better predict the consequences of possible treatments at an individual level, rather than at a population level, to deal with the variability across patients.

Prediction of the consequences of treatments at an individual level is challenging also because of the dynamic and sudden fluctuations of AD symptoms. It can be difficult to identify reliable treatment responses, especially if a single endpoint 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 [18] and ultimately predict whether the chosen treatment is effective and whether the disease is adequately controlled at an individual level.

While ML methods have been successfully applied to prediction tasks, they often lack interpretability (cf. Section 1.3). Here we aimed to develop a biologically interpretable mechanistic machine learning model that can predict the daily evolution of AD severity scores at an individual level. We applied a model-based machine learning approach [55], 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. Such approach has already been applied to a birth cohort data on allergic sensitisation to uncover latent atopy classes [160] or to estimate asthma misclassification and risk factors in yearly questionnaire data [161]. However, it has not been applied to predict daily changes in disease outcome or in the field of AD.

We hypothesised that it is possible to decipher the apparent unpredictable dynamics of AD severity scores from each patient’s data. Our research group 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 the skin barrier, immune responses and environmental stressors [159] [162]. Our aim was therefore to adapt the structure of the published mechanistic model to real patient data (Fig. B.2), and to develop a Bayesian model that can make personalised predictions of the evolution of AD severity, given past severity scores and treatment usage data.

4.2 Methods

4.2.1 General approach

Using the longitudinal data from two published clinical studies [163] [140] (example raw data shown in Fig. B.2), we developed and validated a Bayesian model that can predict the next

day's AD severity score for each patient. Our Bayesian model explicitly describes within-patient uncertainties in disease outcomes using probability distributions, and between-patients 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 model (Fig. 4.1A), which adopted the structure of a previously published mechanistic model of AD pathogenesis [159] [162] (Fig. 4.1B). The model was tailored to the context of the clinical studies in which the data was collected. We then trained the model (fitted to the data) using Bayesian inference, i.e. 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 with a horizon of one week (cf. Section 2.3.1). 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.

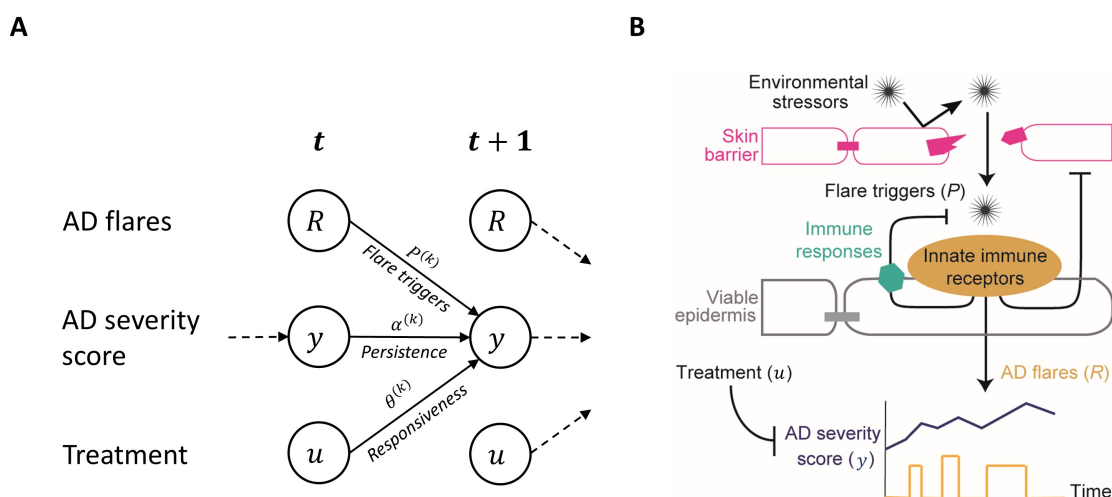


Figure 4.1: Bayesian model of AD severity dynamics. A: Schematic diagram of the probabilistic model. The arrows depict the relationships between state variables included in the model. B: A schematic diagram of the published mechanistic model of AD pathogenesis from which the structure of the proposed model was adopted. Flare triggers (P) and AD flares (R) are latent variables, and AD severity score (y) and treatment applied (u) are the measured variables. The variable u corresponds to the daily binary stepping-up variable.

4.2.2 Data

We chose two datasets that included daily recording of symptoms and treatments over a moderately long period (details in B.1).

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 [163]. 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 (Fig. B.3). The Flares dataset also included a daily binary “stepping up” variable, i.e. the answer 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 randomised 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 [140]. The SWET dataset included the individual child’s daily categorical bother score over 16 weeks with only 1.9% of the bother score missing (Fig. B.4) for a total of 35854 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.

4.2.3 Bayesian models

We developed a Bayesian model that predicts the AD severity score ($y^{(k)}(t + 1)$) for the k -th patient at day $t + 1$, given two observables, the previous day’s score ($y^{(k)}(t)$) and the treatment applied ($u^{(k)}(t)$) (Fig. 4.1A).

Our model assumes that AD severity ($y^{(k)}(t+1)$) 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 ($u^{(k)}(t)$) (Fig. 4.1B). Flare triggers ($P^{(k)}$) and the resulting flares ($R^{(k)}(t)$) were modelled as latent variables¹.

¹We deviate a bit from our convention here, as $R^{(k)}(t)$ and $P^{(k)}$ are unknown and scalar rather than observed matrices. This is to be consistent with the notation of flare triggers and flares introduced in [159].

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, $\hat{y}^{(k)}(t) \in [0, 10]$, is rounded to the nearest integer to derive the discrete severity score reported by patients as:

$$y^{(k)}(t) = \text{Round}(\hat{y}^{(k)}(t)) \quad (4.1)$$

We then described the dynamics of $\hat{y}^{(k)}(t)$ by an exponentially modified Gaussian distribution truncated between 0 and 10:

$$\hat{y}^{(k)}(t+1) \sim \mathcal{N}_{[0,10]}(\alpha^{(k)}\hat{y}^{(k)}(t) + \theta^{(k)}u^{(k)}(t) + R^{(k)}(t) + \beta_0, \sigma^2) \quad (4.2)$$

$$R^{(k)}(t) \sim \text{Exp}(\beta = P^{(k)}) \quad (4.3)$$

$\hat{y}^{(k)}(t+1)$ follows a Gaussian autoregressive process perturbed by exponentially distributed AD flares $R^{(k)}(t)$ with scale $\beta = P^{(k)}$, which reflects the assumption that flares occur more frequently in the presence of flare triggers. The autoregression is characterised by the patient-dependent autocorrelation or persistence of the severity score ($\alpha^{(k)}$), patient-dependent responsiveness to treatment ($\theta^{(k)}$), and population-level intercept (β_0) and variance (σ^2). The patient-dependent parameters, $\alpha^{(k)}$, $\theta^{(k)}$ and $P^{(k)}$, are given hierarchical priors, with population mean (μ_α, μ_θ) and dispersion parameters ($\sigma_\alpha, \sigma_\theta, \sigma_P$):

$$\alpha^{(k)} \sim \text{logit } \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \quad (4.4)$$

$$\theta^{(k)} \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \quad (4.5)$$

$$P^{(k)} \sim \mathcal{N}^+(0, \sigma_P^2) \quad (4.6)$$

We also developed an extended version of this Bayesian model for the SWET dataset (details in Appendix B.2). The extended model allowed us to analyse the effects of potential risk factors (the presence of filaggrin mutations, age, sex, white ethnicity, and sleeping away from home) on the severity score, with their respective weighting parameters, β_{FLG} , β_{Age} , β_{Sex} , β_{White} , and θ_{Home} . We also investigated heterogeneity of treatment responsiveness by replacing the term $\theta^{(k)}u^{(k)}(t)$ with $\theta_{SU}^{(k)}u_{SU}^{(k)}(t) + \theta_{CS}^{(k)}u_{CS}^{(k)}(t) + \theta_{CI}^{(k)}u_{CI}^{(k)}(t)$, where $u_{SU}^{(k)}(t)$, $u_{CS}^{(k)}(t)$ and $u_{CI}^{(k)}(t)$ are binary variables that indicate whether the k -th patient stepped-up, applied topical corticosteroids and calcineurin inhibitors, respectively, with their respective weights, $\theta_{SU}^{(k)}$, $\theta_{CS}^{(k)}$ and $\theta_{CI}^{(k)}$. The weights, $\theta_{CS}^{(k)}$ and $\theta_{CI}^{(k)}$, include dose-independent effects (intrinsic

responsiveness $\gamma_{CS}^{(k)}$ and $\gamma_{CI}^{(k)}$ and dose-dependent effects that are functions of the quantity and the potency of the treatment (Fig. B.5).

Our model did not require imputation of missing values for $y^{(k)}(t)$, since the absence of measurements is naturally accepted by the measurement process of $y^{(k)}(t)$ separately modelled from the dynamics of $\hat{y}^{(k)}(t)$. Imputation of missing values for other covariates is described in Appendix B.3. We chose weakly informative priors and confirmed that our priors were reasonable by conducting prior predictive checks and that our results were not sensitive to the choice of realistic priors (details in Appendix B.4).

4.2.4 Model fitting

Model training was performed using the Hamiltonian Monte-Carlo algorithm in the probabilistic programming language Stan [56]. 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} and computing effective sample sizes. We conducted fake-data check, by fitting the model with samples from the prior predictive distribution, to verify that the inference algorithm could retrieve known parameters.

4.2.5 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 a local logarithmic scoring rule (log predictive density, lpd). These metrics were plotted against training days (equivalently training data size) to produce learning curves (details in Appendix B.5).

We compared our model to four reference models: a discrete uniform forecast, $y^{(k)}(t+1) \sim \mathcal{U}(0, 10)$, 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, $y^{(k)}(t+1) \sim \mathcal{N}(y^{(k)}(t), \sigma^2)$, where the next score is assumed to be around the previous score; and a mixed effect autoregressive model with treatment effects (our model without flares triggers), $\hat{y}^{(k)}(t+1) \sim \mathcal{N}_{[0,10]}(\alpha^{(k)}\hat{y}^{(k)}(t) + \theta^{(k)}u^{(k)}(t) + \beta_0, \sigma^2)$.

4.3 Results

4.3.1 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 intervals (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 (Tables B.1 and B.2).

Three main model parameters that describe patient-dependent dynamics of the severity score are the autocorrelation parameter $\alpha^{(k)}$ for the short-term persistence of the AD severity score, the parameter $\theta^{(k)}$ for the responsiveness to treatment, and $P^{(k)}$ for the amount of flares triggers, of the k -th patient. $\alpha^{(k)} \rightarrow 1$ or $\alpha^{(k)} \rightarrow 0$ means that the predicted severity is close to or does not depend on the previous day's severity, respectively. $\theta^{(k)} < 0$ or $\theta^{(k)} > 0$ 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 (Figs. B.6 and B.7).

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 Fig. 4.2). Typical trajectories observed included fluctuations of the severity score with a return to a healthier state (Figs. 4.2A and 4.2C) or without (Figs. 4.2B and 4.2D). However, the model does not always capture changing data distribution over time (in Fig. 4.2A the severity fluctuates more at the beginning), which could suggest that our assumption that $P^{(k)}$ remains constant during the length of the trial is not always true.

4.3.2 Model validation

We then validated the model to assess its generalisability beyond the training data. The learning curves demonstrated an improvement in both RPS and lpd, as more data becomes available (Fig. 4.3), 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 (Fig. 4.3). For

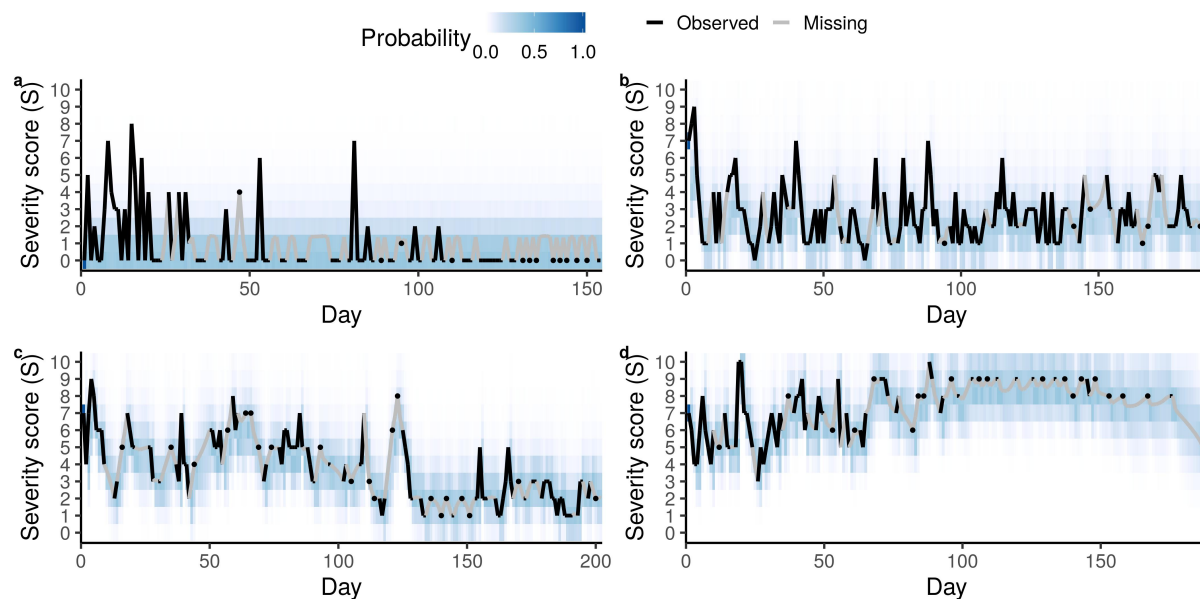


Figure 4.2: Posterior predictive distribution of AD severity scores for four representative patients from Flares dataset. (A, C) Bother score returns to a healthier state. (B, D) Bother score does not improve. The plots show the time evolution of the posterior predictive probability mass function as 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.

example, we achieved a lpd of $\log(0.25)$ with SWET dataset, which could be interpreted as if the model assigns a 25% probability to the true outcome on average, compared to 9% for a chance-level forecast. Calibration curves (Fig. B.8) suggested that the predicted probabilities were reasonably calibrated up to 30-40% in Flares dataset and up to 50-60% in SWET dataset.

We estimated that the RPS was increased (performance decreasing) by 0.001773 ($SE = 0.000299$) in Flares and 0.0057378 ($SE = 0.0001741$) in SWET when the prediction horizon (t) is increased by one day (e.g. one day forecast versus two days forecasts; the order of magnitude of the RPS is 0.1). Similarly, we estimated that the lpd was decreased (performance decreasing) by 0.016555 ($SE = 0.002372$) in Flares and 0.081664 ($SE = 0.001917$) in SWET when the prediction horizon (t) is increased by one day. These results confirm the expectation that the performance is decreasing with increasing prediction horizon. We can observe that the performance loss seems more pronounced in the SWET compared to the Flares dataset, but estimates from the Flares dataset may be over-optimistic considering that SWET data is of better quality than the Flares data. Using the SWET estimates, and with the assumption that the performance loss remains linear, we can extrapolate these results to conclude that the model performance is not much better than chance after around 10-12 days.

Similar results were obtained from the model we fitted using the daily scratch score recorded

in the observational study for Flares dataset² (Fig. B.9).

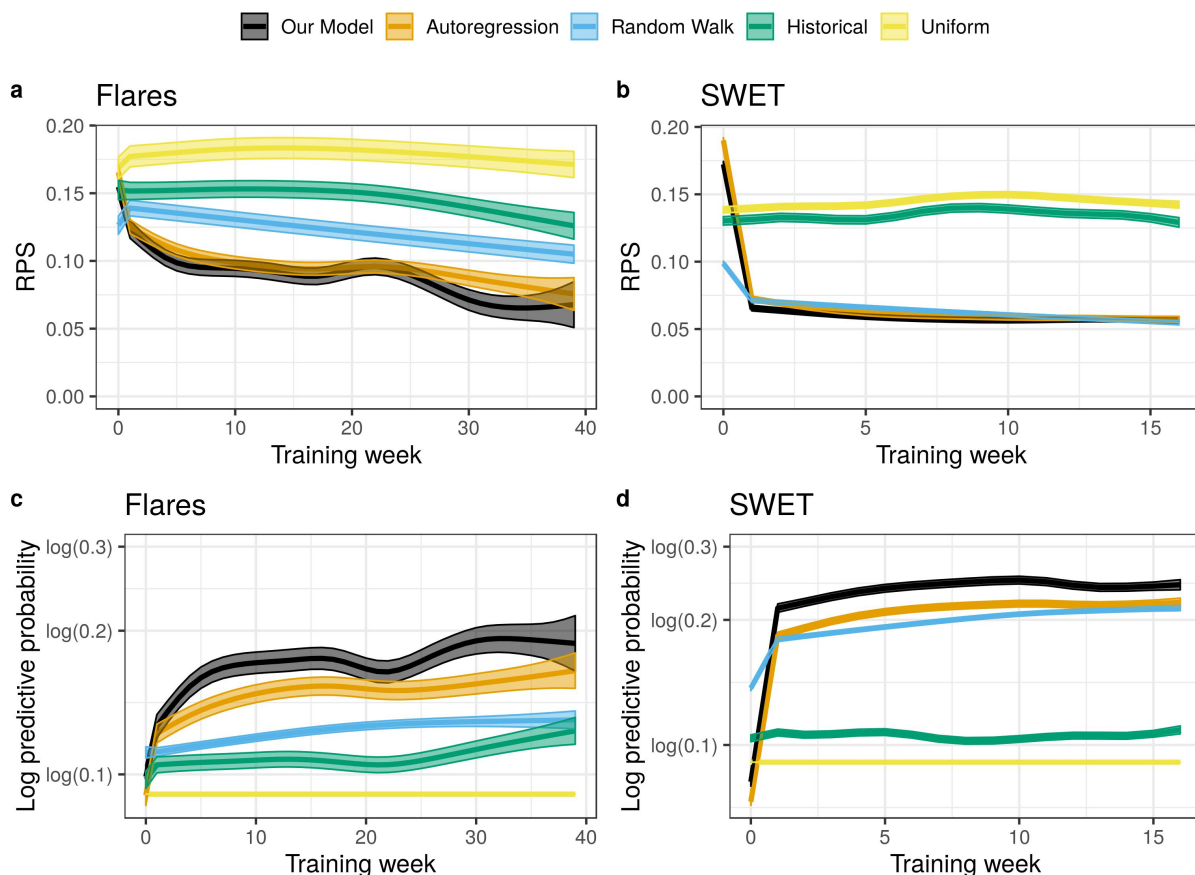


Figure 4.3: Comparison of predictive performance between the “Our Model” and four reference (“Uniform”, “Historical”, “Random Walk” and “Autoregression”) models. The performance is evaluated for one-day-ahead predictions and plotted as a function of the training week. Confidence bounds correspond to $\pm SE$. 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).

4.3.3 Effects of treatment modalities and risk factors on predictions

The extended model with additional covariates was also successfully fit to SWET dataset (Table B.3). The posterior predictive checks confirmed that the model could capture diverse patterns of the severity score trajectories, such as large and rapid fluctuations (Fig. 4.4A), large but slow fluctuations (Fig. 4.4B), and controlled AD (Fig. 4.4C). The model could not predict previously unseen patterns, such as transitions of the score from 1 to 10 in a day (Fig. 4.4D at around 70 days), as the model learned the dynamic patterns from past data.

²The scratch score was not recorded in SWET.

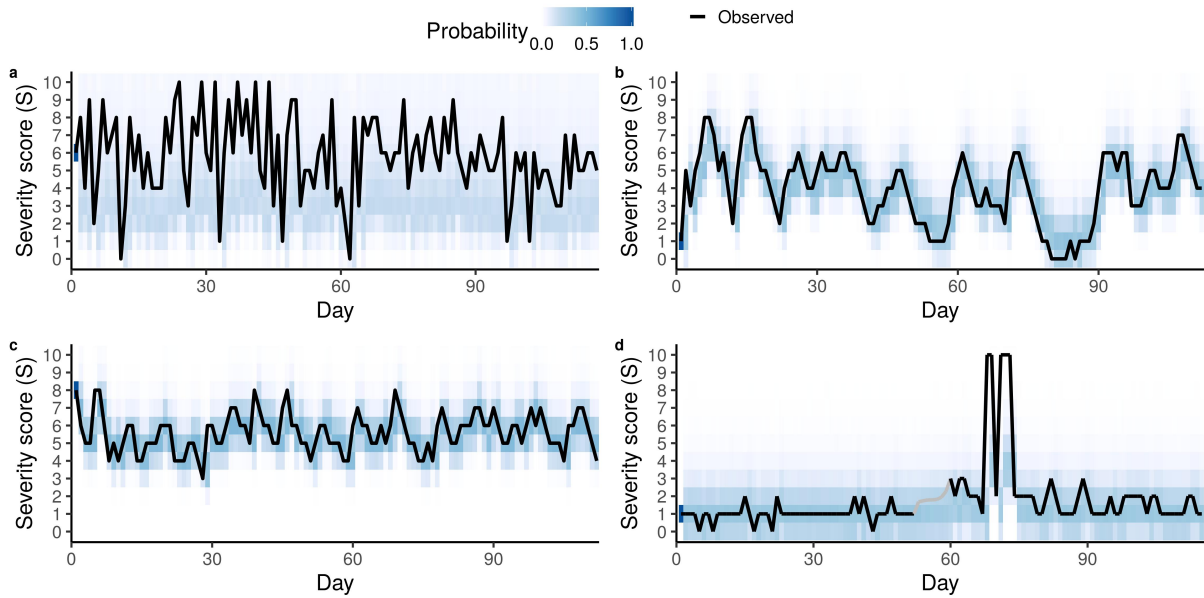


Figure 4.4: Fitting of the extended model. Posterior predictive distribution of AD severity scores for four representative patients from SWET dataset: (A) large and rapid fluctuations, (B) large but slow fluctuations, (C) controlled AD, and (D) controlled but with transitions of the score from 1 to 10 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.

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 (Fig. 4.5A), as the 95% credible interval of the relevant parameters did not contain 0 and by $\beta_{Age} < 0$ (older age decreases the severity score), $\beta_{FLG} > 0$ (the presence of filaggrin mutations increases the severity score), and $\theta_{Home} < 0$ (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 (Fig. 4.5A; 95% credible interval of β_{Sex} and β_{White} in both sides of 0, suggesting that their effects on the severity score could be both negative or positive).

Further analysis of the parameters, $\theta_{SU}^{(k)}$, $\gamma_{CS}^{(k)}$ and $\gamma_{CI}^{(k)}$, which describe the dose-independent effects of the treatment on the severity score, demonstrated that none of the treatments appears to have a significant effect at the population level (grey shaded areas in Fig. 4.5B spans from negative to positive values). However, treatments could have a significant effect at the patient-level. For example, the parameter estimates for one of the patients (orange shaded areas in Fig. 4.5B) 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

$\gamma_{CS}^{(k)}$ (the dose-independent responsiveness to corticosteroids) being negative (i.e. 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 ($\gamma_{CI}^{(k)}$) and step-up ($\theta_{SU}^{(k)}$) for the same patient. Following this criterion, we confirmed significant effects of corticosteroids in 90 individuals (out of 295 who used corticosteroids) and of step-up in 25 individuals (out of 284 who stepped-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 showed a significant dose-dependent responsiveness.

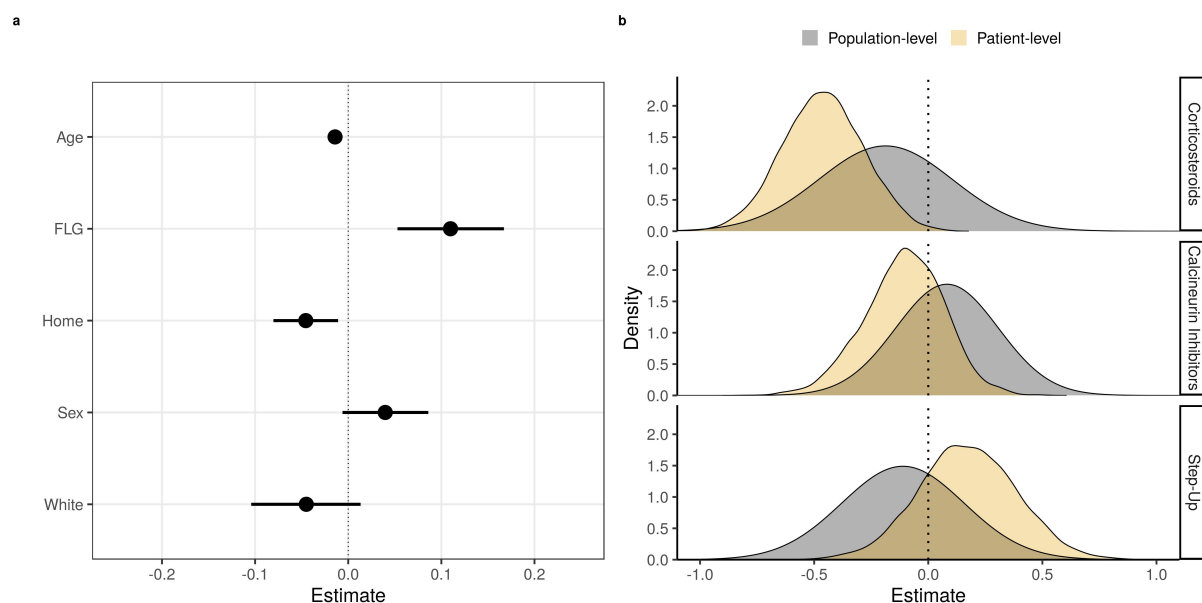


Figure 4.5: Estimated effects of potential risk factors and responsiveness to treatments on the severity score. A: Population-level estimates of the parameters (β_{Age} , β_{FLG} , θ_{Home} , β_{Sex} , β_{White}) for potential risk factors (age, presence of filaggrin mutation, sleeping at home, sex, 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 ($\gamma_{CS}^{(k)}$, $\gamma_{CI}^{(k)}$, $\theta_{SU}^{(k)}$ for corticosteroid, calcineurin inhibitors and step-up) at a population-level (grey) and for a specific patient (orange).

4.4 Discussion

4.4.1 Main findings

This study demonstrated a proof-of-concept that predicting the evolution of eczema severity is possible. We developed a novel mechanistically-inspired 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 (Fig. 4.1). The model learned rich, heterogeneous, and dynamic patterns in the daily evolution of AD severity scores that may otherwise appear random and noisy (Figs. 4.2 and 4.4). Our method extracted information on whether the chosen treatment is effective (responsiveness to treatment), whether the AD score is persistent and susceptible to flares, at an individual level (Figs. 4.5B, B.6 and B.7). 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 (Fig. 4.5A). The model was validated using the data from two published clinical studies to confirm its generalisability and the possibility to learn and predict the short-term dynamics of AD severity scores from each patient's data (Fig. 4.3).

4.4.2 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 (cf. Chapter 8). The priors will then be updated as more data becomes available, to make personalised and more accurate predictions.

In addition, our model-based Bayesian approach is appropriate to develop models for clinical use, especially when the data is not as controlled as in a clinical trial. Our model explicitly describes uncertainties in disease outcomes (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 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 the available data and study context. For example, our model was extended by incorporating the additional information (on potential risk factors and 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 (e.g. air pollution, weather, cf. Chapter 5), host factors (e.g. compliance to daily bathing, allergies)

or biological markers (cf. Chapter 6).

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 personalised predictions.

4.4.3 Limitations of the study and future directions

The datasets we used in this study contained daily measurements 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 [164] and validating AD symptom measures such as POEM [85]. 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 [78], (o)SCORAD [80], or their self-assessed versions (cf. Chapter 7). 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 horizons with weekly (instead of daily) measurements. Appropriate evaluation of the effects of data frequency on severity prediction will help designing more effective and informative clinical trials towards personalised medicine.

The predictive capabilities of the model could be potentially improved by incorporating more data, or by using better-quality data, i.e. with fewer missing values or more precise information about treatments. For example, our model assumes that the same amount 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 B.3). 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 inspired by the previously published mechanistic model of AD development [159] [162], while mechanisms of disease onset may differ from those relevant to disease persistence. As this work focuses on the evolution of AD severity, we interpret the model as describing the onset of AD flares triggered by infiltrated pathogens. The model structure was also much simpler than that of the mechanistic model. If longitudinal measurements of interactions between environmental stressors, the skin barrier and immune responses become 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 [16]. The prediction of daily evolution of severity scores could be further used to suggest optimal treatment strategies for individual patients (cf. Chapter 8), in addition to conventional computational methods using optimal control theory and bifurcation analysis [165].

Chapter 5

The role of environmental factors in AD severity prediction

In Chapter 4, we developed a proof-of-concept predictive model of AD severity and witnessed its limitations. One avenue to better predict the evolution of AD severity is to consider additional factors beyond severity measurements and treatment usage. In this chapter, we investigate whether measurements of environmental factors such as weather or air pollution can help predict AD severity.

This chapter is adapted from our paper “Impact of environmental factors in predicting daily severity scores of atopic dermatitis”, published in 2021 in *Clinical and Translational Allergy* [61] under the terms of the [Creative Commons CC BY license](#). In this paper, we reanalysed the data from a study that found that weather and air pollution were associated with AD symptoms [166], in order to investigate whether these associations are really predictive of future AD severity. The model developed in this study is designed as a tool to investigate environmental factors, and does not attempt to replicate the models developed in Chapter 4, even though they share similarities (autoregression, patient-dependence of parameters).

The code written for this project is available at https://github.com/VDelorieux/AD-environmental_factors.

This research project is a collaborative work with Valentin Delorieux (VD), Dr. Young-Min Kim (YK), Dr. Kangmo Ahn (KA), and Prof. Hywel C. Williams (HW). VD and I jointly contributed to the analysis, the computer code, design of models and writing the manuscript. YK and KA gratefully shared the data and their perspectives regarding the re-analysis of their data. HW contributed his insights on the clinical relevance of the model and helped reviewing and editing the manuscript.

5.1 Introduction

AD patients often suffer from symptoms that fluctuate every day, resulting in a decreased quality of life due to the unforeseeable dynamic nature of the symptoms. AD affects almost 20% of the paediatric population worldwide and the prevalence of AD in children is still increasing globally [69]. The rising prevalence of AD coincides with increased urbanisation and industrialisation worldwide [167], and the assessment of the effects of environmental factors on AD has gained a growing importance.

AD pathophysiology is considered to be affected by external environmental factors, such as air pollution from particulates, ultraviolet radiation, temperature and humidity - collectively known as the skin exposome [168] [169]. Environmental factors have been shown to be associated with AD development and aggravation [170] [171], as well as other aspects of AD including barrier dysfunction [71] or care visits [172]. Prior studies investigated whether environmental factors were associated with the current AD severity [166] [173] [174] [175], but none have considered the dynamic nature of the severity nor have they investigated whether the future AD severity can be predicted by environmental factors. Despite this evidence gap, a profusion of smartphone eczema apps have emerged offering to track disease severity and environmental factors with bold claims of being able to predict AD flares [12].

We have previously developed statistical machine learning models to predict daily AD severity scores at an individual level (Chapter 4). The models demonstrated that it was possible to decipher much of the apparent unpredictable dynamics of AD severity scores from each patient's longitudinal data. The models investigated the effects of age, filaggrin mutations and the treatments used, such as calcineurin inhibitors and corticosteroids, on daily changes of AD severity scores. However, environmental stressors were only modelled as latent variables due to the lack of availability of such data in the training datasets.

In this chapter, we aim to assess the impact of environmental factors in predicting future AD severity scores. We developed a statistical machine learning model to predict daily AD severity scores for individual patients using a longitudinal dataset with high-quality environmental and AD symptom data. We used that model to evaluate whether environmental factors including weather and air pollutants are important determinants in predicting the next day's AD scores from today's scores.

5.2 Methods

5.2.1 Data

We used the longitudinal data from a published panel study [166] that investigated the short-term impact of environmental factors on AD symptoms in Seoul, South Korea. The cohort included 177 Korean paediatric patients (67 girls and 110 boys) aged five or younger (average age of 2.0 years old, SD = 1.6) with mild to severe AD (mean SCORAD at enrolment of 31.1, SD = 12.8). The data contained the daily recording of the atopic dermatitis symptom score (ADSS) [176] over 17 months (Figs. 5.1 and C.1). ADSS is a sum of scores for six AD signs (dryness, edema, itching, oozing, redness, and sleep disturbance), each on a discrete scale from 0 (none) to 4 (severe). In this study, we used the six AD sign scores, rather than their sum (ADSS), to extract more information from the data. 18.9% of the daily AD sign scores were missing. We removed five patients with less than ten daily observations, resulting in a total of 34921 patient-day observations.

The use of topical corticosteroids (binary) was recorded daily. Weather variables (mean temperature, relative humidity, total rainfall, diurnal temperature range) and the concentration of air pollutants (PM10, NO2, O3) were collected daily for each patient. A binary AD symptom state was derived in [166] from the sign scores: the state was one when the sum of itching and sleep disturbance scores was greater than or equal to two, and the scores of at least two of redness, dryness, edema, or oozing were non-zero; and the state was zero otherwise (Fig. 5.1).

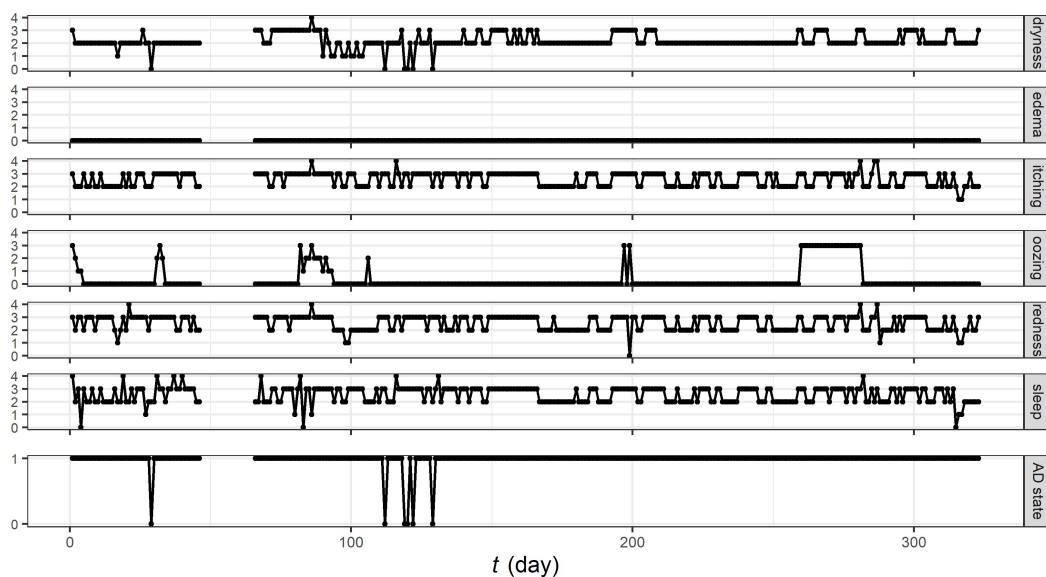


Figure 5.1: Example trajectories of the six AD sign scores and the derived AD symptom state for a representative patient.

5.2.2 Mixed effect autoregressive ordinal logistic regression model

We developed a mixed effect autoregressive ordinal logistic regression model to predict the patient-dependent dynamics for each of the six AD sign scores. The model is described by:

$$y^{(k)}(t+1) \sim \text{OrderedLogistic} \left(\beta_0^{(k)} + \sum_{i=0}^3 \alpha_i \delta_{y^{(k)}(t), i}, \mathbf{c} \right) \quad (5.1)$$

Where $y^{(k)}(t)$ is a sign score for the k -th patient at day t , $\beta_0^{(k)}$ is the patient-dependent intercept (the random effect), α_i 's are the regression coefficients, $\delta_{x,y}$ is the Kronecker delta, and \mathbf{c} is the vector of cut-off values of the ordered logistic distribution (details in Section 2.4). We also considered a model that includes all covariates of interest (environmental variables and TCS usage at t), for evaluation of the impact of environmental factors in the linear predictor, and models with one covariate each. Cross-correlation analysis did not support the inclusion of higher order time lags for sign scores or covariates in the model. The models were fitted using the “lme4” package in R¹, to pairs of successive scores $(y^{(k)}(t+1), y^{(k)}(t))$. Pairs with at least one missing value were removed from the training set.

5.2.3 Model validation

We evaluated the predictive performance of our models in a forward-chaining setting with a horizon of one day. The performance of predicting AD sign scores was quantified by the ranked probability score (RPS), a proper scoring rule for ordinal probabilistic forecasts. The performance of predicting binary AD symptom states was evaluated with the Brier score.

We compared the performance of our models to that of two benchmark models: the uniform forecast model that predicts each of the five possible outcomes of a sign score with the probability of $1/5$, and the historical forecast model where the probability of each possible outcome is equal to its occurrence in the patient's training data. We also compared our model to the logistic regression model proposed in [166] for the prediction of AD symptom states.

¹We obtained the logit to calculate the ordered logistic distribution by jointly fitting the cumulative distribution ($P(y \leq 0)$, $P(y \leq 1)$, $P(y \leq 2)$ and $P(y \leq 3)$) with logistic regressions.

5.3 Results

5.3.1 Model validation

We trained six mixed effect autoregressive ordered logistic regression models without covariates, one for each of the AD sign scores. The models learnt the patient-dependent dynamics of the sign scores as more data came in and outperformed the benchmark models in predicting the next day's score for all AD signs (Fig. 5.2). The performance of the benchmark models varied between signs, confirming that the scores of some signs are more imbalanced than others (Fig. C.1) and easier to predict. For instance, the historical forecast model (and our model) achieved an almost perfect prediction for edema, for which the outcome is 0 nearly 90% of the time. For other signs, such as dryness, the RPS of our model was about 60% lower (i.e. achieved a better performance) than that of the historical forecast model after 200 days of training.

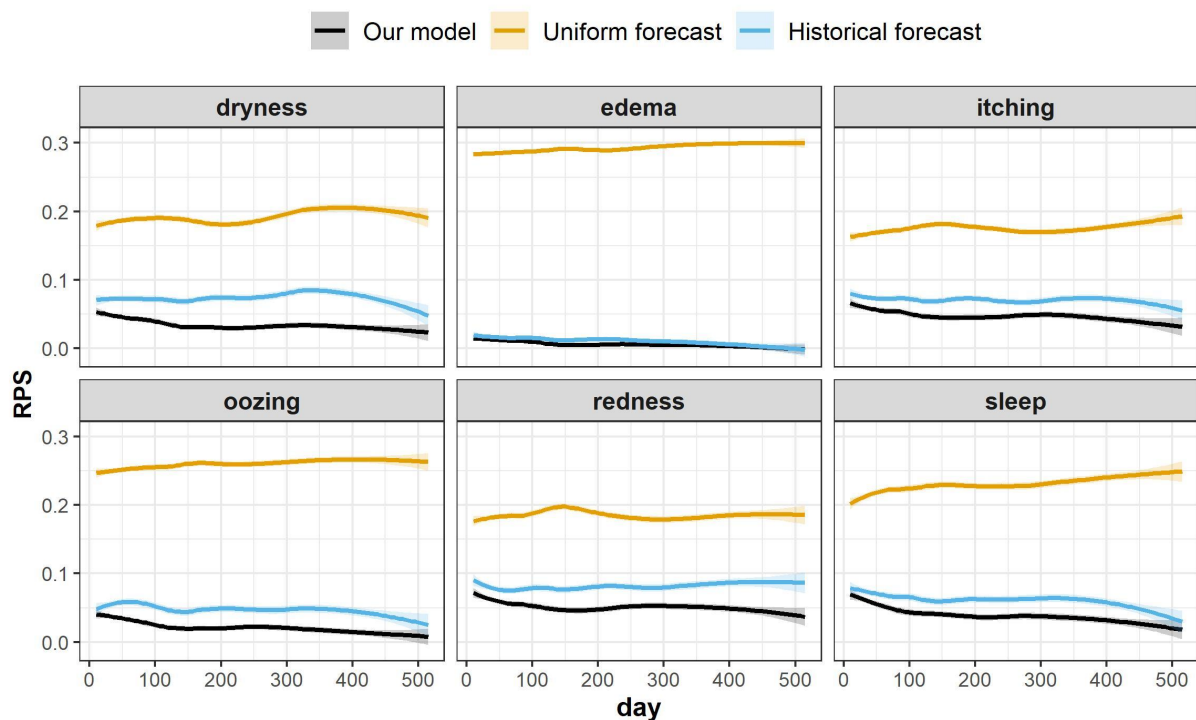


Figure 5.2: Comparison of the predictive performance of our model (the mixed effect autoregressive ordinal logistic regression without covariates) to that of the uniform forecast and the historical forecast models, for prediction of each of the six AD signs. The performance of predicting AD sign scores is measured by the RPS (the lower RPS indicates the better predictive performance). Learning curves were obtained using locally weighted scatterplot smoothing (LOWESS). Shaded areas correspond to ± 1.96 standard error.

We derived a prediction for the binary AD symptom state from the six mixed effect autoregressive logistic regression models for AD sign scores (Fig. 5.3), assuming their predictions are independent random variables. Our model outperformed the two benchmark models and the

logistic regression model proposed in [166]. The Brier score of our model was about 40% lower (i.e. achieved a better performance) than the logistic regression model, whose performance matched that of the historical forecast.

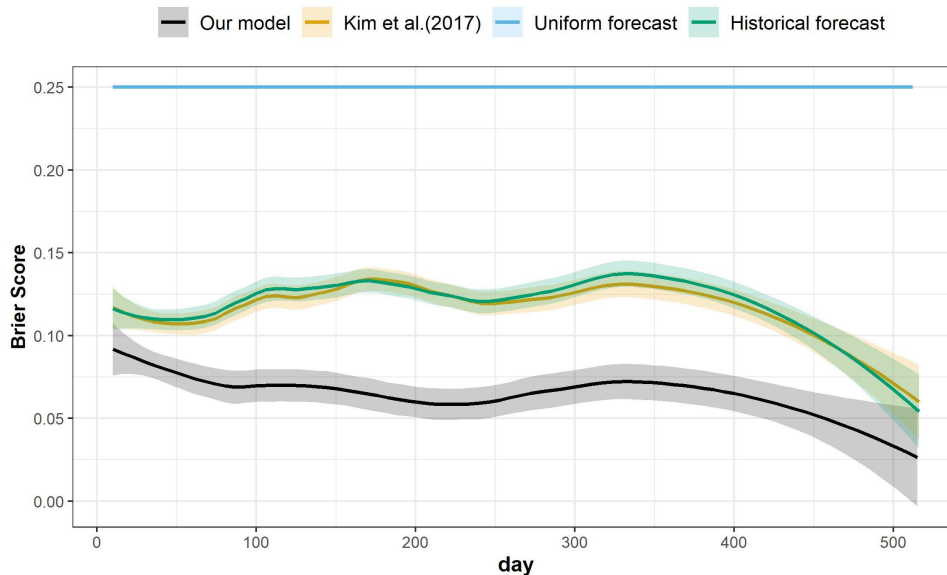


Figure 5.3: Comparison of the predictive performance for the models predicting the AD symptom state. Our model (without covariate and for which the prediction for the AD symptom state is derived from the predictions for each AD sign) is compared to the uniform and the historical forecast models, and the logistic regression model proposed in Kim et al. (2017) [166]. The performance is measured by the Brier score (the lower Brier score corresponds to the better predictive performance). Learning curves were obtained using LOWESS smoothing. Shaded areas correspond to ± 1.96 standard error.

5.3.2 Effect of environmental factors on the model's predictions

To assess the effects of exogenous factors (weather, air pollution, TCS usage) on the prediction of AD sign scores, we computed the pairwise difference in the RPS between the model without covariates, the models with a single covariate, and the model with all covariates (Fig. 5.4).

No evidence was found to support that the inclusion of exogenous factors improved the predictive performance of the model for all signs (Fig. 5.4a). Even though some of the coefficients associated with the covariates have confidence bounds that do not cross 0, all of them were small in magnitude, accounting for approximately only 1% of the linear predictor (Fig. 5.4b). These small coefficients result in the lack of a noticeable improvement in the predictive performance of the model by addition of the covariates.

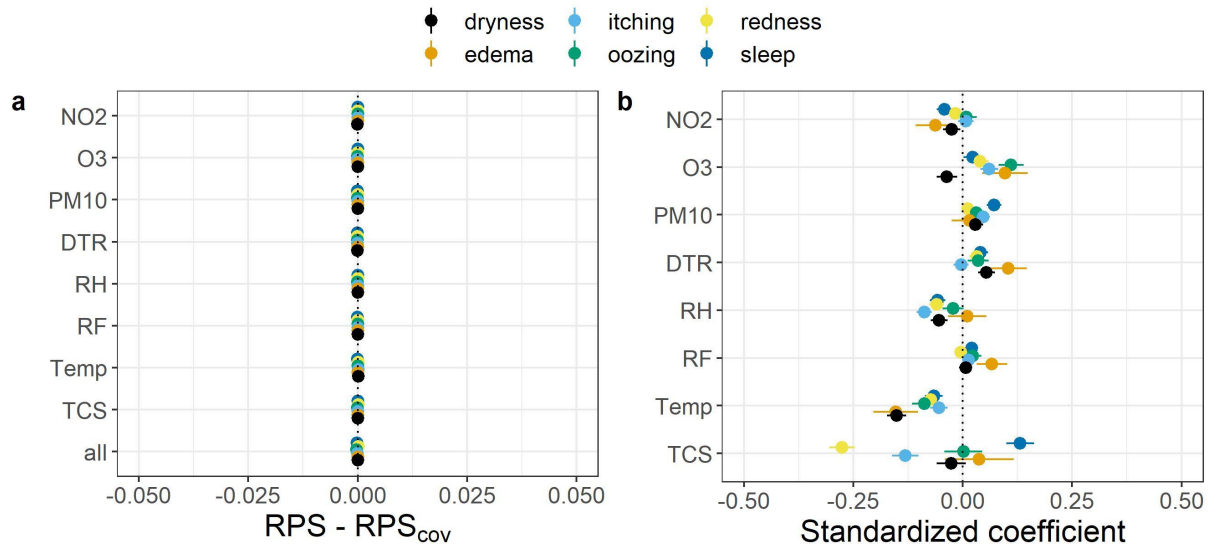


Figure 5.4: Effects of environmental factors (mean temperature (“Temp”), relative humidity (RH), total rainfall (RF), diurnal temperature range (DTR), and the concentration of air pollutants (PM10, NO2, O3)) and treatment usage (topical corticosteroids (TCS)) on AD sign score prediction. (a) The pairwise difference in predictive performance between the model without covariate (RPS) and the model with covariates (single or all, RPS_{cov}). $RPS - RPS_{cov} > 0$ indicates that the model with covariates has a higher predictive performance. (b) The coefficients for the covariates in the single-covariate models (\pm SE). A positive coefficient means that an increase in the covariate is associated with a higher probability for more severe outcomes.

5.4 Discussion

The assessment of the effects of environmental factors on AD has gained a growing importance. Many prior studies investigated whether environmental factors were associated with current AD severity, but they have failed to consider the dynamic nature of the severity nor investigated whether future AD severity can be predicted by real-time data on environmental factors.

We developed a mixed effect autoregressive ordinal logistic regression model that can predict the next day’s AD severity scores, using the longitudinal data from a published panel study [166]. Our model successfully made daily predictions of the AD severity scores: it outperformed two benchmark models for the prediction of AD sign scores (Fig. 5.2) and outperformed the benchmark models and the logistic regression model for the prediction of an AD symptom state proposed in [166] (Fig. 5.3). The inclusion of environmental factors did not improve the predictive performance of our model (Fig. 5.4).

Our results from a comprehensive dataset of South Korean children do not present any convincing evidence to support a claim that AD symptoms were associated with weather or air pollutants on a short-term basis. The short-term influence of environmental factors on AD sign scores was outweighed by the previous scores’ persistence, and the next day’s score for a

patient is more accurately predicted using the patient's today's score than using environmental data. Neglecting the time-dependence of AD severity as in previous studies [172] [173] [174] [175] may misguide inferences about the effect size of environmental associations. The extent to which AD severity can be predicted from the measurement of environmental factors remains unclear. Our results throw serious doubts into the claim of many AD apps that purport to use real-time environmental measures to inform AD users when their AD symptoms are likely to flare.

It is possible that other “internal” factors, such as the development of skin autoimmunity, may be more important than external factors in determining disease fluctuations over time [177]. Factors that determine disease incidence may also be different from those that determine disease chronicity, so it is still possible that environmental factors may be more predictive of the AD onset and long-term disease trajectories rather than short-term symptom fluctuations.

This study used a high-quality dataset on South Korean children with high rates of data completion. Modelling each of the six AD signs enabled to extract more information from the data and to generate predictions for any quantity of interest to the practitioner, be it ADSS or any combination of the sign scores. In terms of study limitations, the AD sign scores used in this study were obtained by subjective assessment by the patients (or their carers) on a discrete scale. Further investigation of the seemingly small effects of environmental factors on AD severity scores may benefit from more data or better quality data, for example by recording time-series of SCORAD or EASI, or their self-assessed version, as they are more objective and may offer better responsiveness to environmental changes. However, dichotomisation of AD sign scores into a binary AD symptom state as proposed in [166] reduces the power of the analysis [97] and is not recommended. Our model might be improved by taking measurement errors into account using state-space models or by modelling the correlations between AD signs in a multi-outcome regression. However, we believe the additional complexity in the model would only result in marginal improvements in the already solid predictive performance.

Whilst this study focused on the association between environmental factors and future AD severity scores, whether environmental factors cause a change in AD scores is of more interest to the AD community. Estimating causal effects is challenging, as causal inference methods assume the absence of unobserved confounders [178], an assumption that is deemed unrealistic. For example, “staying indoors” was not recorded in the original study [166] but could lead to reverse causation if patients decided to stay indoors during a pollution peak. Estimation of non-linear interactions may also be required, if patients react differently to environmental triggers depending on their severity: mild patients could be less sensitive than severe patients who may be subject to a “ceiling effect”. Constructing causal diagrams using specialist background knowledge could be a promising approach.

We will not consider environmental factors in the rest of this thesis, as the limited availability of such data is not outweighed by significant gains in predictive performance.

Chapter 6

The role of biomarkers in AD severity prediction

In this chapter, we continue our investigation of features that could help the prediction of AD severity by questioning whether biomarkers, measured before the start of a therapy, can help predict its outcome. In particular, we analyse whether serum cytokines and chemokines can be predictive of the outcome of a systemic immunosuppressive therapy in AD adults.

This chapter is adapted from our paper “Can serum biomarkers predict the outcome of systemic immunosuppressive therapy in adult atopic dermatitis patients?”, published in 2022 in *Skin and Health Disease* [62] under the terms of the [Creative Commons CC BY license](#). The models presented in this study are designed as tools to investigate biomarkers rather than as extensions of the models developed in Chapters 4 and 5, although they share similar aspects. The code written for this project is available at <https://github.com/ghurault/ssm-eczema-biomarkers>.

We are grateful to our clinical collaborators (Dr. Evelien Roekevisch, Dr. Mandy E. Schram, Dr. Krisztina Szegedi, Dr. Sanja Kezic, Prof. Maritza A. Middelkamp-Hup and Prof. Phyllis I. Spuls) for sharing the data and their insights on the clinical relevance of our findings.

6.1 Introduction

Atopic Dermatitis is a chronic skin disease with a considerable variation in the clinical phenotype and response to treatments among patients [69]. Current treatments aim to manage AD symptoms, such as inflammatory flares and dry and itchy skin, mainly by topical application of emollients and corticosteroids. However, systemic therapy using traditional immunosuppres-

sants is needed for patients with moderate-to-severe AD that do not respond to topical therapy. It is desirable to identify patients who are likely to respond to a systemic immunosuppressive therapy, as the decision to initiate such therapy can be difficult given its known risks [74].

It has been hypothesised that biomarker measurements could help predict therapeutic responses and be used as a tool to stratify patients [10]. Previous studies on AD biomarkers have mainly focused on severity biomarkers, i.e. biomarkers that could be used as surrogates for AD severity: thymus and activation-regulated chemokine (TARC) was suggested to be the single best biomarker to assess disease severity [179] and panels of biomarkers were proposed as “objective” substitutes for EASI [180] and SCORAD [181]. However, the reliability of “severity” biomarkers have been questioned [182] [183], and “severity” biomarkers are different from “predictive” biomarkers that are expected to be predictive of future outcomes¹.

Some previous studies aimed to explore “predictive” biomarkers for several AD treatments. In [184], predictive biomarkers for systemic immunosuppressants (methotrexate or azathioprine) were sought by investigating whether baseline levels of some cytokines/chemokines were statistically different between responders (who achieved > 50% reduction in SCORAD) and non-responders of the therapy. In [185], a high level of serum total IgE was found to be associated with poor response to the maintenance treatment by topical tacrolimus and/or corticosteroids. A clinical trial is underway to explore predictive biomarkers for dupilumab that are most strongly associated with improvement in EASI [186]. However, those studies did not investigate whether the biomarkers can predict treatment outcomes. Instead, they investigated how much the biomarkers were associated with treatment outcomes, but an association does not imply prediction since associations often do not generalise to unseen data [19]. Predictions need to be generated and evaluated on out-of-sample data, beyond quantification of associations.

In this study, we explored predictive biomarkers for systemic immunosuppressive therapy for AD (by methotrexate or azathioprine) using the same data as in [184] and investigated whether serum cytokines/chemokines measured for each patient pre-treatment can be used as predictive biomarkers. Here, biomarkers are considered predictive only if their inclusion improves the performance of the best available predictive model (without those biomarkers) for AD severity scores (the primary outcomes of clinical trials). Using model comparison to evaluate the predictive potential of biomarkers can be useful to offset the effects of other factors, such as historical data (Chapter 4), that can also contribute to the prediction of future AD severity scores. We also considered multiple biomarkers in a multivariable regression setting.

¹In fact, there is little reason to believe “severity” biomarkers are also predictive, since if a biomarker is a perfect surrogate of (current) AD severity, it cannot carry additional information about future AD severity. An alternative approach, if perfect “severity” biomarkers existed, would be to predict the evolution of these biomarkers and then transform these predictions into predictions for AD severity scores. However, we are sceptical of this approach, as biomarker data is expensive to collect compared to severity data, and may be noisier.

Specifically, we developed a statistical machine learning model that can predict the patient-dependent dynamic evolution of AD severity scores. Our model predicts continuous AD severity scores rather than arbitrary dichotomies of “responders” vs “non-responders” to avoid potential information loss that may demand more data to reach a reliable conclusion [18]. Using the model, we explored predictive biomarkers that can reliably predict AD severity scores at different timepoints, rather than a single timepoint after treatment, to reduce the impact of the variability in treatment responses at an individual patient-level. A mere comparison of AD severity scores before and after treatment is indeed unsuitable to determine patient-level treatment responses and whether biomarkers are predictive of those responses, because AD severity scores can fluctuate over time regardless of treatment or biomarkers [18]. These fluctuations can be stochastic (unpredictable), due to unobserved/unrecorded factors (e.g. environmental factors) or measurement error (cf. inter- and intra-rater variability of severity scores).

6.2 Methods

6.2.1 Data

We used longitudinal data from a published clinical study [184], in which 42 adult AD patients received systemic therapy (azathioprine or methotrexate) for over 24 weeks. The data includes the baseline concentrations of 26 serum cytokines and chemokines (listed in Fig. 6.4) measured before the start of the treatment (week 0), the status of the filaggrin gene (FLG) mutation (yes/no), age and sex for each of the 42 patients. Therapeutic responses were assessed by EASI, SCORAD, oSCORAD (the objective component of SCORAD) and POEM at weeks 0, 2, 4, 8, 12 and 24 from the start of the therapy, for each patient.

Concentrations of the serum biomarkers were log-transformed and standardised to have zero mean and unit variance for each biomarker. Three out of 1092 (= 26 x 42) measurements of the serum biomarkers were missing and imputed by the population mean of the corresponding biomarker. The missing FLG mutation status for six patients was imputed by a default status of “no mutation”. The patients’ age was standardised to have a population mean of 0 and a variance of 1. Our model (detailed below) considers the dynamics of the severity scores with a constant interval of two weeks up to week 24. Therefore, we treated the absence of AD severity measurements at weeks 6, 10, 14, 16, 18, 20, and 22 as missing. It resulted in 56% missing values for EASI, (o)SCORAD and POEM.

6.2.2 Model overview

We developed a Bayesian state-space model to make probabilistic predictions of future AD severity scores (either EASI, SCORAD, oSCORAD or POEM) for each patient. The model for each severity score assumes that the true latent (unobserved) severity score follows its own latent dynamics, and that the measured severity score is obtained as a result of an imperfect measurement of the latent severity score at each timepoint (Fig. 6.1). Missing values were treated in our model as an absence of measurement. As a Bayesian model, our model describes uncertainties in parameters and severity scores as probability distributions. Quantifying uncertainties in parameters is especially suitable when dealing with small datasets, where the estimates are likely to be noisy.

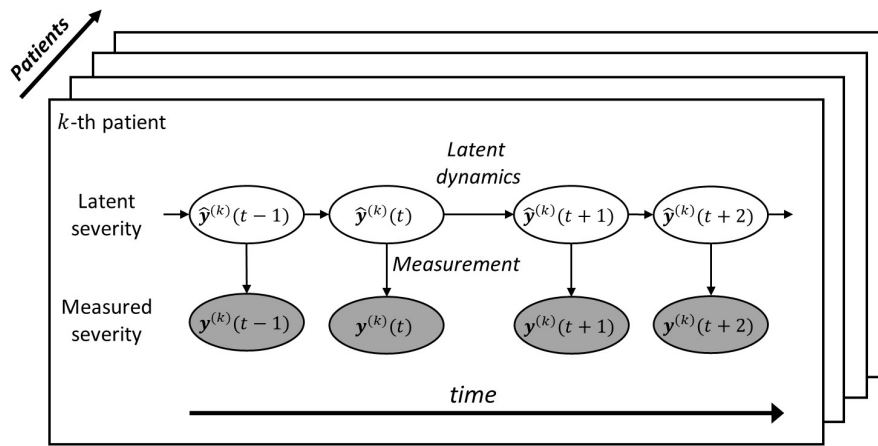


Figure 6.1: An overview of the Bayesian state-space model for probabilistic predictions of AD severity scores. The model describes the latent dynamics of a latent severity score (white ovals) and the measurement of the latent severity scores (grey ovals).

We modelled the latent dynamics of the latent score, $\hat{y}^{(k)}(t)$, for the k -th patient at time t (with a constant interval of 2 weeks) by a mixed effect autoregressive model:

$$\hat{y}^{(k)}(t+1) \sim \mathcal{N}(\alpha^{(k)}\hat{y}^{(k)}(t) + \beta_0^{(k)} + (\mathbf{x}^{(k)})^T\boldsymbol{\beta}, \sigma_1^2) \quad (6.1)$$

Where $\alpha^{(k)}$ is the autocorrelation parameter, $\beta_0^{(k)}$ is the intercept, $\mathbf{x}^{(k)}$ is an optional covariates vector for the k -th patient (including biomarkers) with their coefficients $\boldsymbol{\beta}$, and σ_1 is the standard deviation of the latent dynamics. We performed feature selection on the covariates $\mathbf{x}^{(k)}$ by assuming a regularised horseshoe prior for $\boldsymbol{\beta}$ [187]. The horseshoe prior shrinks small coefficients toward 0 while allowing strong signals to remain large, thus limiting overshrinkage, unlike L_1 or L_2 regularisations [188].

Measurement of the latent score $\hat{y}^{(k)}(t)$ is modelled by a truncated² Gaussian distribution centred around the measured severity score $\hat{y}^{(k)}(t)$ for the k -th patient at the t -th timepoint:

$$y^{(k)}(t) \sim \mathcal{N}_{[0,M]}(\hat{y}^{(k)}(t), \sigma_m^2) \quad (6.2)$$

The standard deviation of the measurement process σ_m quantifies the measurement error. Here, we use the term “measurement error” as if the proposed state-space model was the true data-generating mechanism. In practice, this is unlikely to be true and σ_m would quantify both the uncertainty due to the data collection process (aleatoric uncertainty, the measurement uncertainty as it is usually understood) and the uncertainty due to an imperfect modelling of the latent dynamic (epistemic uncertainty). In other words, if a better model of the latent dynamic were proposed, σ_m would be reduced accordingly. With that in mind, σ_m can be used to compute a (model-dependent) minimum detectable change (MDC), that is “the smallest change that can be considered above the measurement error with a given level of confidence” [99]. For the default 95% confidence level, the MDC is determined by $MDC = 1.96 \sigma_m$.

We assumed hierarchical priors for $\alpha^{(k)}$ and $\beta_0^{(k)}$ and weakly informative priors for the other parameters (detailed in Appendix D.1). Model inference was performed using the Hamiltonian Monte-Carlo algorithm in the probabilistic programming language Stan [56], with four chains and 2000 iterations per chain, including 50% burn-in. Prior predictive checks were performed to confirm our choice of priors was reasonable, and fake data checks were conducted to validate the computational method. Convergence and sampling were monitored by looking at trace plots, checking the Gelman-Rubin convergence diagnostic \hat{R} , and computing effective sample sizes.

6.2.3 Model validation

The predictive performance of our model was assessed by K-fold cross-validation ($K = 7$, stratified by patients), where we applied forward chaining to the “test” fold to reflect how the model would be used in a clinical setting³ with the model being updated after each measurement (Fig. D.1). The probabilistic predictions of AD severity scores were evaluated by a logarithmic scoring rule, the log predictive density (lpd), and compared to that of four reference models (details of the reference models in 2.3.3 and their priors in D.1): a uniform forecast model, a random walk model, an autoregressive model and a mixed effect autoregressive model. We also

²The distribution is truncated between 0 and the maximum value, M , of the severity score (72 for EASI, 83 for oSCORAD, 103 for SCORAD and 28 for POEM).

³In a practical clinical setting, we would expect the model to be pre-trained (cf. $K - 1$ training folds) and then updated as more data comes in (cf. forward chaining). We implement this validation procedure in this chapter because the small size of the dataset makes it computationally reasonable.

report the root mean squared error (RMSE) of the expected prediction for ease of interpretation.

6.3 Results

6.3.1 Model fit and validation

We first developed a Bayesian state-space model that predicts the evolution of AD severity scores without covariates (i.e. without demographics, types of treatment, cytokines/chemokines) as a baseline model. The baseline model that predicts future EASI was fitted successfully to the data without evidence of an absence of convergence (Table D.1). Population-level parameters were estimated with good precision, with posterior distributions narrower than their prior distributions (Table D.1). We confirmed that the patient-dependent parameters, $\alpha^{(k)}$ and $\beta_0^{(k)}$, vary between patients, within the range of $[0.37, 0.99]$ for the expected autocorrelation ($\alpha^{(k)}$) and $[0.03, 2.3]$ for the expected intercept ($\beta_0^{(k)}$). The measurement process is responsible for 94.7% (90% credible interval 87.3% - 99.1%) of the total variance for prediction. The posterior mean of the minimum detectable change (MDC) is 8.6 (90% credible interval 7.6-9.6). The posterior predictive distribution of EASI trajectories demonstrated that the model could capture different patterns, despite the absence of several measurements (Fig. 6.2).

Learning curves for two-weeks ahead predictions of EASI by our Bayesian state-space model (SSM in Figs. 6.3A and D.2) demonstrated that the predictive performance improved as more training data (newer measurements for the same patient) came in, and that it outperformed all reference models, supporting the model's structure. The RMSE of the mean prediction for EASI at the next clinical visit (e.g. from week 0 to 2, 2 to 4, 4 to 8, etc.) was 6.3 ± 0.62 (mean \pm SE) for our model, compared to 9.9 ± 0.43 for the random walk model. The performance of our model and the mixed autoregressive model for EASI prediction tended to improve as the prediction horizon increased (Figs. 6.3B and D.3), while we normally expect the predictive performance decreases for a longer prediction horizon. It is possible that this counter-intuitive observation is the result of most patients recovering before the end of the study, making predictions easier.

Similar results, with lower performance relative to the reference models, were obtained for the model predicting oSCORAD, SCORAD and POEM (Fig. D.2). The posterior means (and 90% credible intervals) of the MDC were 9.1 (7.4-10.7) for oSCORAD, 11.4 (9.1-13.5) for SCORAD and 7.7 (6.7-8.9) for POEM. This implies similar amount of measurement error for EASI and (o)SCORAD relative to the range of the score M (between 8% and 12%) and substantially more measurement error for POEM ($\approx 28\%$, cf. Table D.2).

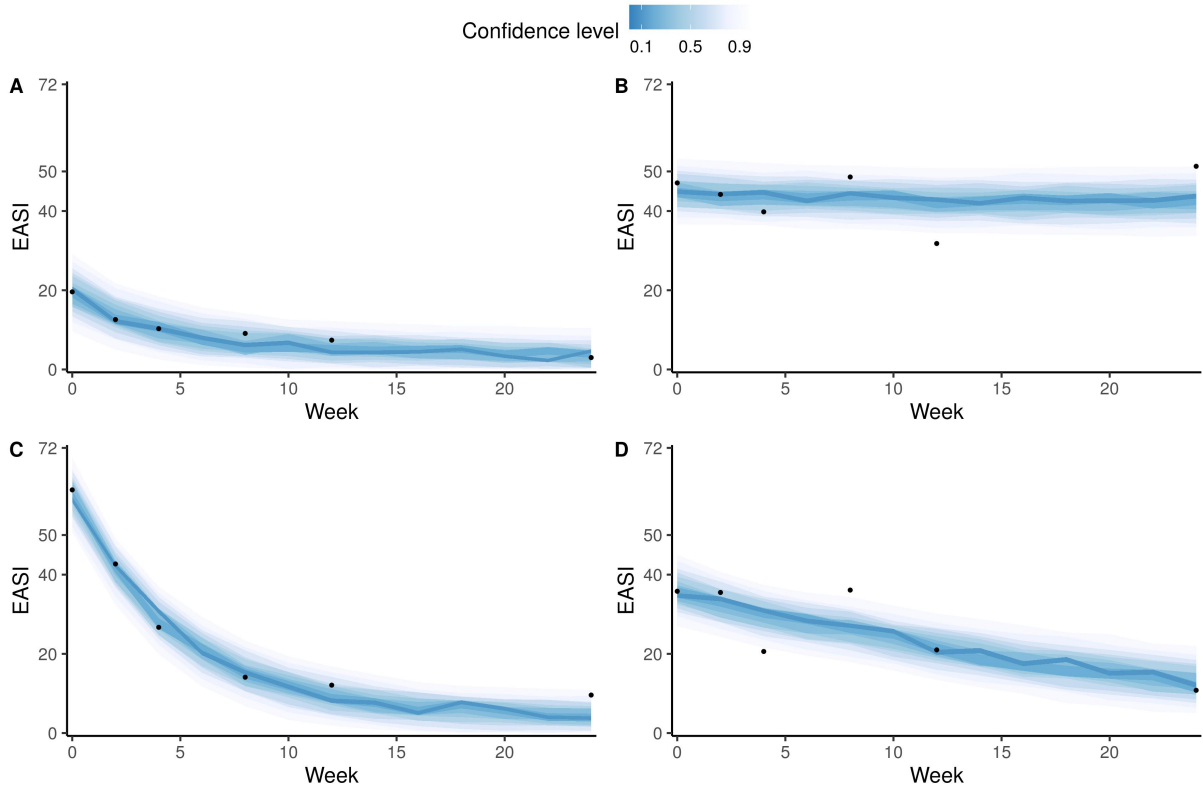


Figure 6.2: The posterior predictive distribution of four representative patients (A-D) by our model predicting EASI. Each of the representative patients demonstrates different dynamics: slow recovery from a moderate EASI (A), persistence of severe EASI (B), rapid recovery from a severe EASI (C), and slow recovery from a severe EASI (D). Dots indicate the measured EASI scores, and the coloured ribbons represent stacked credible intervals of highest density. Lighter and darker ribbons correspond to wider and narrower intervals, respectively.

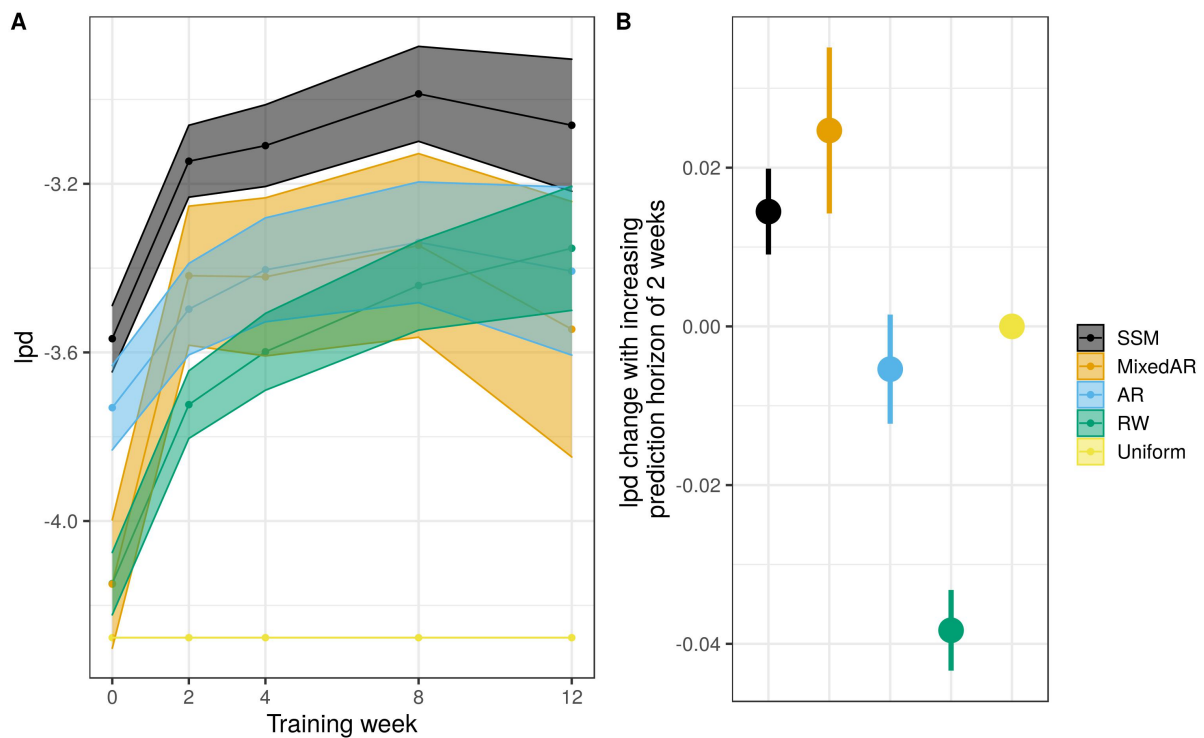


Figure 6.3: Predictive performance for EASI by our Bayesian state-space model (SSM, black) and the reference models, measured by the lpd (higher the better). A: Learning curves (mean \pm SE) for two-weeks ahead prediction after adjusting for different prediction horizons in a linear model. B: Changes in lpd as the prediction horizon is increased by two weeks. The reference models include a mixed effect autoregressive model (MixedAR, orange), an autoregressive model (AR, blue), a random walk model (RW, green), and a uniform forecast (Uniform, yellow).

6.3.2 Effects of biomarkers on the model's predictions

As our Bayesian state-space model outperformed the reference models, we used it to evaluate whether the inclusion of biomarkers improves its predictive performance, thus identifying predictive biomarkers. The covariates included were the 26 serum cytokines/chemokines measured at week 0, the status of FLG mutation, the type of systemic therapy applied (azathioprine or methotrexate), sex and age. Our analysis demonstrated that none of the covariates had a practically significant effect on the model's prediction, as indicated by a small magnitude of the posterior mean and 90% credible intervals for the coefficients β , on both sides of 0 (Fig. 6.4A), and a resulting small and not practically significant contribution of the covariates $((\mathbf{x}^{(k)})^T \beta)$ to the EASI prediction (Fig. 6.4B). As a result, the predictive performance of the model was not improved by including covariates. Similarly, we found no practically significant covariates for the predictive models of SCORAD, oSCORAD and POEM.

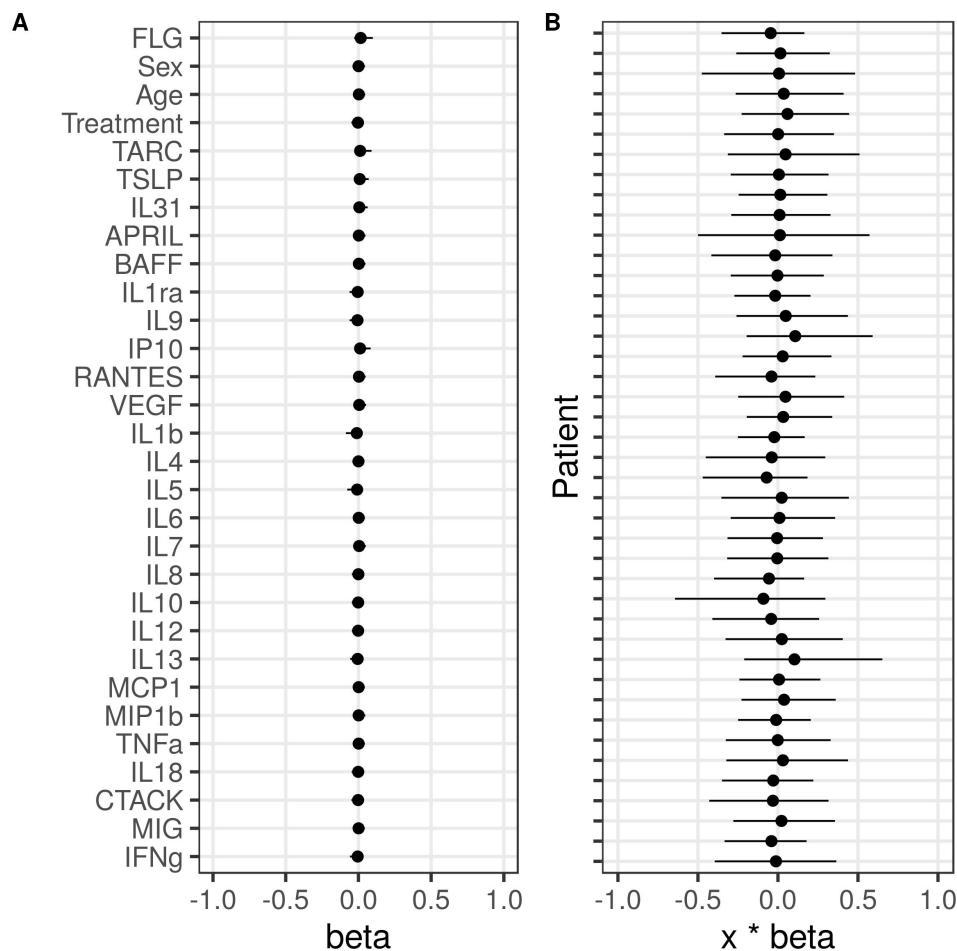


Figure 6.4: Effects of covariates in our model's predictions of EASI (mean and 90% credible intervals). A: Estimates of the coefficients for the biomarkers (26 serum cytokines/chemokines, FLG, sex, age) and the treatment applied. A change of one standard deviation in a covariate corresponds to a change of 1.0 in EASI score. B: Total contribution of all covariates $((\mathbf{x}^{(k)})^T \beta)$ to EASI prediction for each patient.

6.4 Discussion

Prediction of whether a patient is likely to respond to a specific therapy is of high clinical importance, especially if the therapy may have risks of side effects. In this study, we examined whether serum cytokines/chemokines measured for each patient before the start of the therapy can be used as predictive biomarkers for systemic immunosuppressive therapy (methotrexate or azathioprine) for AD.

We developed a Bayesian state-space model that can predict AD severity scores (EASI, SCORAD, oSCORAD, and POEM) two-weeks ahead in the future, at the individual level. The model describes the dynamics of the latent severity for each patient, and the measurement process of the severity scores (Fig. 6.1). The model was trained on the data from 42 adult AD patients who received systemic immunosuppressive therapy in a published clinical study [184] (Fig. 6.2). Our model outperformed standard reference models for time-series forecasting (Fig. 6.3) and was used for further analysis to test the predictive ability of biomarkers. The results revealed that the predictive performance was not improved by including some biomarkers as covariates (Fig. 6.4), suggesting that the biomarkers measured before the start of the therapy did not carry additional information for the prediction of future AD severity scores.

While an absence of evidence for predictive biomarkers of the therapies should not be interpreted as evidence of an absence, our results suggest that the effect of biomarkers on the prediction of severity scores, if any, is likely to be small or too subtle to be captured by our linear model. This is because the prediction errors of future scores by our model was mostly attributed to errors in the score measurement process. Further investigation of the effect of biomarkers on severity score prediction may therefore require data from a larger cohort. However, it is unclear how much new information we can expect to obtain by including more biomarkers, because the biomarkers included in this study have been claimed to be most related to AD [179] and biomarkers are often highly correlated with each other. In addition, the biomarkers' concentrations measured at a single timepoint are likely to be noisy and may not capture the dynamic heterogeneity of a complex disease such as AD. Whether the benefit of potentially more accurate predictions with biomarkers outweighs the cost of collecting data for such models remains an open question.

By explicitly describing measurement errors in severity scores in our model, we also estimated the minimum detectable change for the AD severity scores. The estimated MDC suggested that it may be easier to predict objective scores such as EASI and (o)SCORAD than subjective scores such as POEM. Even though these estimates are model-dependent, they are larger than already published estimates of the minimal important difference (MID) for EASI, (o)SCORAD and POEM [102] (Table D.2), indicating that it is possible that the changes in an

outcome that a patient identify as important are not always detectable. Further research is needed to elucidate how we ensure that clinically important changes cannot be attributed to measurement errors, and to validate the results from this study on different cohorts of patients.

While the data used in this study is from a small cohort of patients ($N = 42$), the AD severity scores were measured at six time-points for each patient. The repeated measurements of severity scores enabled us to capture the dynamic nature of the AD severity scores for each patient and to investigate consistent effects of biomarkers and treatments on AD severity scores within each patient, as it reduces the impact of the variability in treatment responses (including measurement errors).

The analysis of the data in this study did not identify any predictive biomarkers for systemic immunosuppressive therapy for AD, and validation on different cohorts of patients is still required. Until then, we will not consider biomarkers as potential predictors for AD severity in the rest of this thesis. The method proposed in this study may help to re-analyse previously collected individual longitudinal data to test the predictive ability of potential predictive biomarkers.

Chapter 7

The role of measurements in AD severity prediction

In Chapters 5 and 6, we explored factors that could help predict the evolution of AD severity, beyond previous AD severity and treatment usage. Another alternative to improve the prediction of AD severity, as noted in Chapter 4, is to rely on better quality measurements of AD severity. In this chapter, we investigate to what extent using better quality measurements can help the prediction of future AD severity. This chapter is composed of two parts, corresponding to the investigation of two severity scores: an “objective” score, PO-SCORAD, and a “subjective” score, POEM.

First, we develop models to predict PO-SCORAD and package our approach in a computational framework “EczemaPred” and its corresponding R package. This part corresponds to a paper “EczemaPred: A computational framework for personalised prediction of eczema severity dynamics” [63], which has been published in *Clinical and Translational Allergy* between the first (January 2022) and the final submission (May 2022) of this thesis, under the [Creative Commons CC BY license](#). The code is written for this part is available at <https://github.com/ghurault/EczemaPred> (package) and <https://github.com/ghurault/EczemaPredPOSCORAD> (analysis).

Then, we used the same approach for predicting another severity score, POEM, to test whether our results could be generalised to a score with different characteristics. In this second part, we also suggested improvements to the models developed for PO-SCORAD. The results for PO-SCORAD and POEM prediction are thus obtained from slightly different versions of the models¹. The code written for this part is not released at the time of writing (January 2022).

¹The next chapter will include the improvements introduced for the POEM models to the models developed for PO-SCORAD.

The first part of this project (development of EczemaPred and PO-SCORAD prediction) was done in collaboration with Pierre Fabre Laboratories, which provided the data. We thank Sophie Mery, Alain Delarue, Dr. Markéta Saint Aroman, Dr. Gwendal Josse, Dr Sébastien Barbarot, Dr. Thérèse Nocera and Yann Kling for their inputs and help editing the manuscript. We also thank our clinical collaborator, Prof. Jean-François Stalder, for sharing his clinical expertise. The second part of this project (POEM prediction) was done in collaboration with Prof. Kim Thomas and Prof. Hywel Williams who contributed the data used for fitting the models.

7.1 Introduction

We have previously developed a Bayesian model of the evolution of AD severity and demonstrated that predicting the patient-specific evolution of AD was possible (Chapter 4). The model captured the patient-specific heterogeneity in dynamic trajectories of AD severity and responsiveness to treatment. However, its predictive performance and clinical applicability were limited because the model was developed using a daily bother score, which is a subjective global measure of distress caused by AD and is not suitable to capture different aspects of AD symptoms reliably. Using a validated objective severity score that combines multiple severity items could improve the predictive performance and make predictive models more relevant for clinical practice.

The Harmonising Outcome Measures for Eczema (HOME) initiative recommended EASI as the core outcome instrument for clinical signs of eczema to be measured in clinical trials [79], and POEM to measure patient-reported symptoms [86]. SCORAD and its objective component oSCORAD have also been validated as outcome instruments [81], and other scores such as Six Area Six Signs AD (SASSAD) are still routinely used in clinical practice. All these instruments report AD severity as a single score obtained by aggregating the severity scores for multiple severity items, including intensity signs, subjective symptoms, and extent (cf. Section 2.1.3). The severity items capture different aspects of AD severity and may therefore follow their own dynamic.

In this chapter, we introduce EczemaPred, a computational framework to predict patient-specific dynamic of AD severity. It is based on the idea that modelling the evolution of each relevant severity item and aggregating the predictions could improve the performance and the clinical relevance of the prediction of AD severity dynamics. EczemaPred consists of a collection of Bayesian state-space models that describe the item-dependent dynamic of each severity item. The predictions for any AD severity score can be obtained by aggregating the predictions for the relevant severity items made by their Bayesian state-space models.

First, we developed EczemaPred using the Patient-Oriented SCORAD (PO-SCORAD) [87], a validated self-assessment of SCORAD [91] that can be recorded on a smartphone app. Self-assessments of AD severity are more suitable to track the short-term (daily to weekly) evolution of the severity dynamics compared to clinical assessments that can be performed only during clinical consultations of a limited frequency. In particular, PO-SCORAD is one of the core instruments recommended by the HOME initiative to measure patient-reported symptoms in clinical practice [88]. We validate the EczemaPred models to predict the dynamic evolution of PO-SCORAD using the longitudinal datasets from two clinical studies, a dataset from a published study [189] and another dataset we collected in an observational study.

Then, we extended EczemaPred models to predict POEM, the other instrument recommended by HOME to measure patient-reported symptoms [88], using the longitudinal data from a published clinical trial [190]. Prediction of POEM is deemed to be more challenging than of PO-SCORAD, because it is measured weekly and considered as a “subjective” score for self-assessments, whereas PO-SCORAD is considered as an “objective” score that can be recorded daily.

7.2 Methods

7.2.1 Datasets

Observational study

An observational study ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT04553224), NCT04553224) was conducted from November 2019 to February 2020 in Toulouse (France) following the approval by IEC (CPP Ile de France V, Saint Antoine Hospital, n°582211). We recruited 16 adult AD patients (mean age 25 y.o., SD=5) whose SCORAD were between 20-40 (mean SCORAD 34.6, SD=4.4 at inclusion). Patients recorded PO-SCORAD using an app (<https://www.poscorad.com>) for up to 12 weeks every day, while continuing their usual treatment. In the case of AD flare ($N = 8$ patients), medication was changed by the investigators. Informed consent was obtained from all study participants.

PO-SCORAD datasets

We used two datasets with daily to weekly measurements of PO-SCORAD and its severity items over a moderately long period. The first dataset, referred to as dataset 1, is from a published study that investigated the role of an emollient in children (mean age 3.6 y.o., SD = 1.3) with

mild to moderate AD [189]. Dataset 1 consists of PO-SCORAD recorded for 347 children approximately twice weekly (usually every 3 or 4 days) for up to 17 weeks (119 days), resulting in 9943 observations. 11 children with less than five observations in the original study were excluded. The second dataset, referred to as dataset 2, was obtained from the observational study described in the previous subsection. The data consists of PO-SCORAD recorded daily by 16 adult AD patients for up to 12 weeks (84 days), resulting in 1136 patient-day observations with 13.6% missing values.

Dataset 1 had 70.3% missing values if it was expected to have daily recordings. Compared to dataset 2, dataset 1 had more missing values due to less frequent recordings (3 to 4 times fewer observations) per patient, but contained about nine times more observations in total as it was collected from 21 times more patients (Table 7.1). The severity dynamics appeared to be relatively more stable in dataset 1 than in dataset 2 (Fig. 7.1).

Table 7.1: Characteristics of PO-SCORAD datasets

	Dataset 1	Dataset 2
Number of subjects	347	16
Age (mean \pm SD)	3.6 ± 1.3	24.7 ± 5.0
PO-SCORAD recording	Twice weekly	Daily
Duration	Up to 17 weeks	Up to 12 weeks
Missing values for daily recording	70.3%	13.6%
Observations	9943	1136
PO-SCORAD at inclusion	31.2 ± 7.7	34.6 ± 4.4
Data collection	Subject notebook	Smartphone app

We recall that PO-SCORAD is defined by $0.2A + 3.5B + C$. $A \in [0, 100]$ corresponds to the extent (the percentage of the area affected by eczema in the whole body); $B \in [0, 18]$ is the sum of six intensity signs, each of which is assessed on a representative area for that sign using an ordinal scale from 0 to 3; and $C \in [0, 20]$ is the sum of two subjective symptoms (more details in Section 2.1.3). In both datasets 1 and 2, the extent (A) takes discrete values (0, 1, ..., 100; i.e. with a resolution of 1), and each subjective symptom score takes discrete values (0.0, 0.1, ..., 10.0; i.e. with a resolution of 0.1).

We did not use demographics or treatment information in our models because our results in Chapter 4 suggested that their inclusion does not result in a noticeable improvement in performance when predicting the patient-specific daily evolution of AD severity. In this study, we aimed to develop simple models with a good predictive performance that could be extended later to investigate other factors (effects of treatment in Chapter 8).

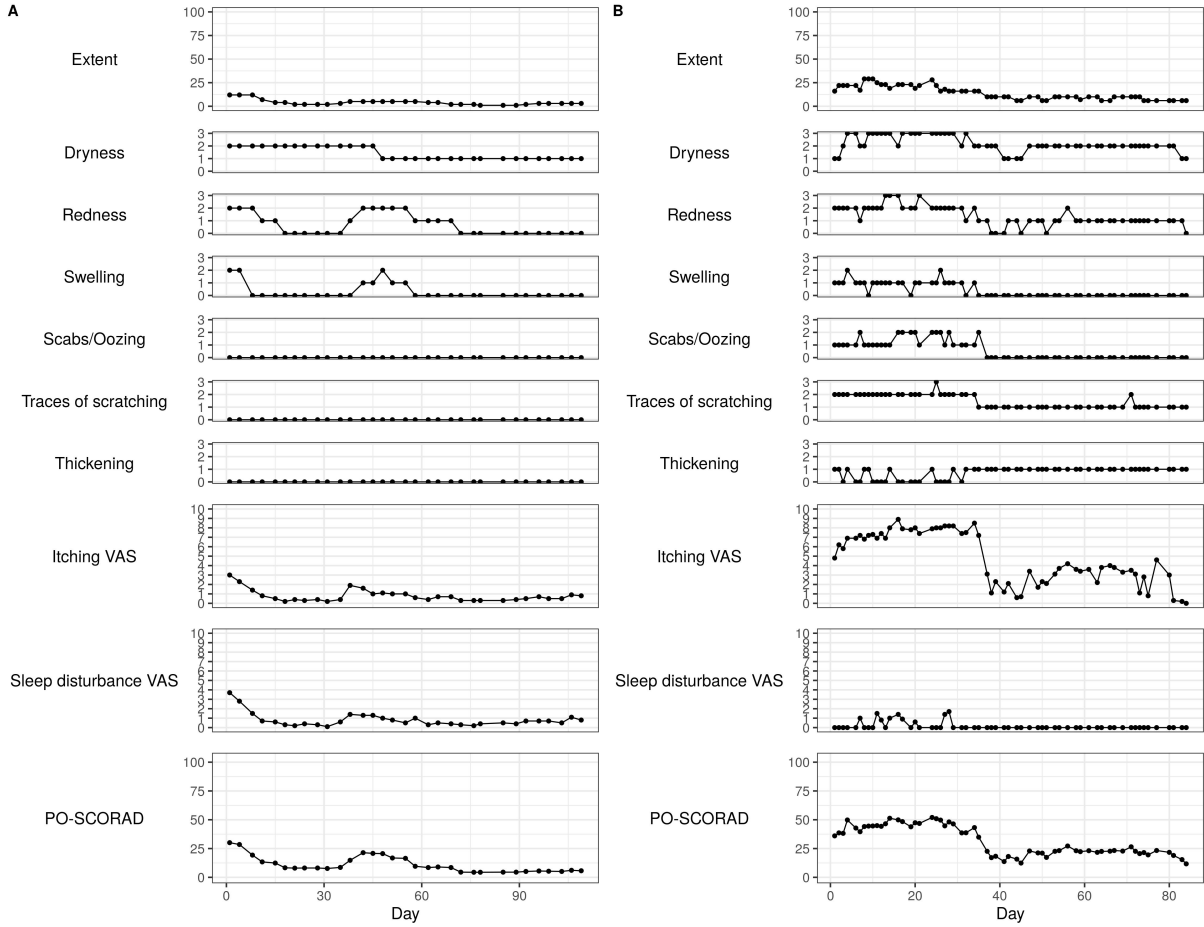


Figure 7.1: Example trajectories of PO-SCORAD and its severity items for representative patients from datasets 1 (A) and 2 (B).

POEM dataset

To develop predictive models for POEM, we used data from the published CLOTHES clinical trial [190], which investigated the effectiveness of silk garments for the management of eczema. The data consisted of the weekly recordings of POEM for up to 24 weeks by 271 children with moderate to severe AD (Fig. 7.6B). The number of missing values during the follow-up time was 11% resulting in a total of 5478 patient-week observations.

We recall that POEM is the sum of the self-reported answers to 7 questions, each of which asks how many days a symptom (itchy skin, sleep disturbance, bleeding skin, weeping skin / oozing clear fluid, cracked skin, flaking skin, dry skin) occurred in the past week, graded on a discrete scale from 0 to 4 (0 = “no days”, 1 = “1 or 2 days”, 2 = “3 or 4 days”, 3 = “5 or 6 days”, 4 = “7 days”; more details in Section 2.1.3).

7.2.2 EczemaPred

We introduce EczemaPred, a collection of (Bayesian state-space) models that can be used to describe the data-generating mechanisms of each severity item. Each model assumes the existence of a true latent (unobserved) severity that follows its own latent dynamics and that the recorded severity was obtained as a result of imperfect measurement of the latent severity (Fig. 7.2A). Predictions for the different severity items can then be aggregated to produce predictions for the severity scores. Missing values were treated as an absence of measurement in the state-space models.

PO-SCORAD model

EczemaPred to predict SCORAD (and PO-SCORAD as its self-assessment version) consists of nine independent (sub-)models, each corresponding to one of the nine severity items for SCORAD (Fig. 7.2B). We tailored the state-space models to each component of SCORAD (Fig. 7.2C).

The extent model assumes that we can subdivide the body area into 100 patches, each with a probability, $\hat{y}^{(k)}(t)$, of being classified as lesional, and that each patch can transition between lesional and non-lesional states, the transitions being described by a two-state Markov chain (the latent dynamic). Parameters of the Markov chain latent dynamic were made patient-dependent with hierarchical priors. The measurement is specified as a binomial distribution to

count the number of lesional patches to produce the extent score ($y = A$):

$$y^{(k)}(t) \sim \mathcal{B}(100, \hat{y}^{(k)}(t)) \quad (7.1)$$

We propose a general purpose state-space model for intensity signs and subjective symptoms, because we do not have much insights regarding their data-generating mechanisms. The model is described by:

$$y^{(k)}(t) \sim \mathcal{D}(\hat{y}^{(k)}(t)) \quad (7.2)$$

$$\hat{y}^{(k)}(t) = g(\tilde{y}^{(k)}(t)) \quad (7.3)$$

$$\tilde{y}^{(k)}(t+1) \sim \mathcal{N}(\tilde{y}^{(k)}(t), \sigma^2) \quad (7.4)$$

$$\tilde{y}^{(k)}(t_0) \sim \mathcal{N}(\mu_0, \sigma_0) \quad (7.5)$$

Where \mathcal{D} is the measurement distribution, \hat{y} is the latent score, \tilde{y} is a transformation of \hat{y} by the link function $g^{-1}(\cdot)$ and follows a random walk dynamic with standard deviation σ . μ_0 and σ_0 are the population mean and standard deviation of \tilde{y} at the initial condition t_0 , respectively.

For the subjective symptoms, considering the high number of categories ($M + 1 = 101$), we chose a Binomial measurement distribution with a logit link. For intensity signs, considering the small number of categories ($M + 1 = 4$), we preferred to choose an ordered logistic measurement distribution, which is more flexible than a Binomial (cf. Section 2.4) with a linear link function.

POEM model

To predict POEM, we extended the choice of measurement distribution and latent dynamics.

1. We proposed a measurement distribution (\mathcal{B}_{Day}) that is adapted to the POEM scoring system, where the number of days a symptom occurred in the past week is counted with a Binomial distribution before applying the 0-4 categorisation.
2. We introduced a novel parametrisation of the ordered logistic measurement that is more interpretable and scales well to multiple categories².
3. We modelled the evolution of the symptoms jointly by assuming correlations between changes in the latent scores. This means replacing the independent univariate random walk dynamics described in Eq. (7.4) by a multivariate random walk dynamic³ with

²This ordered logistic distribution parametrisation will allow us to fit extent and subjective symptoms in the next chapter, but will not be applied to PO-SCORAD predictions here. We refer to this parametrisation as v2 in the appendix.

³With independent dynamics, we could consider the model for POEM to be made of seven sub-models, one for

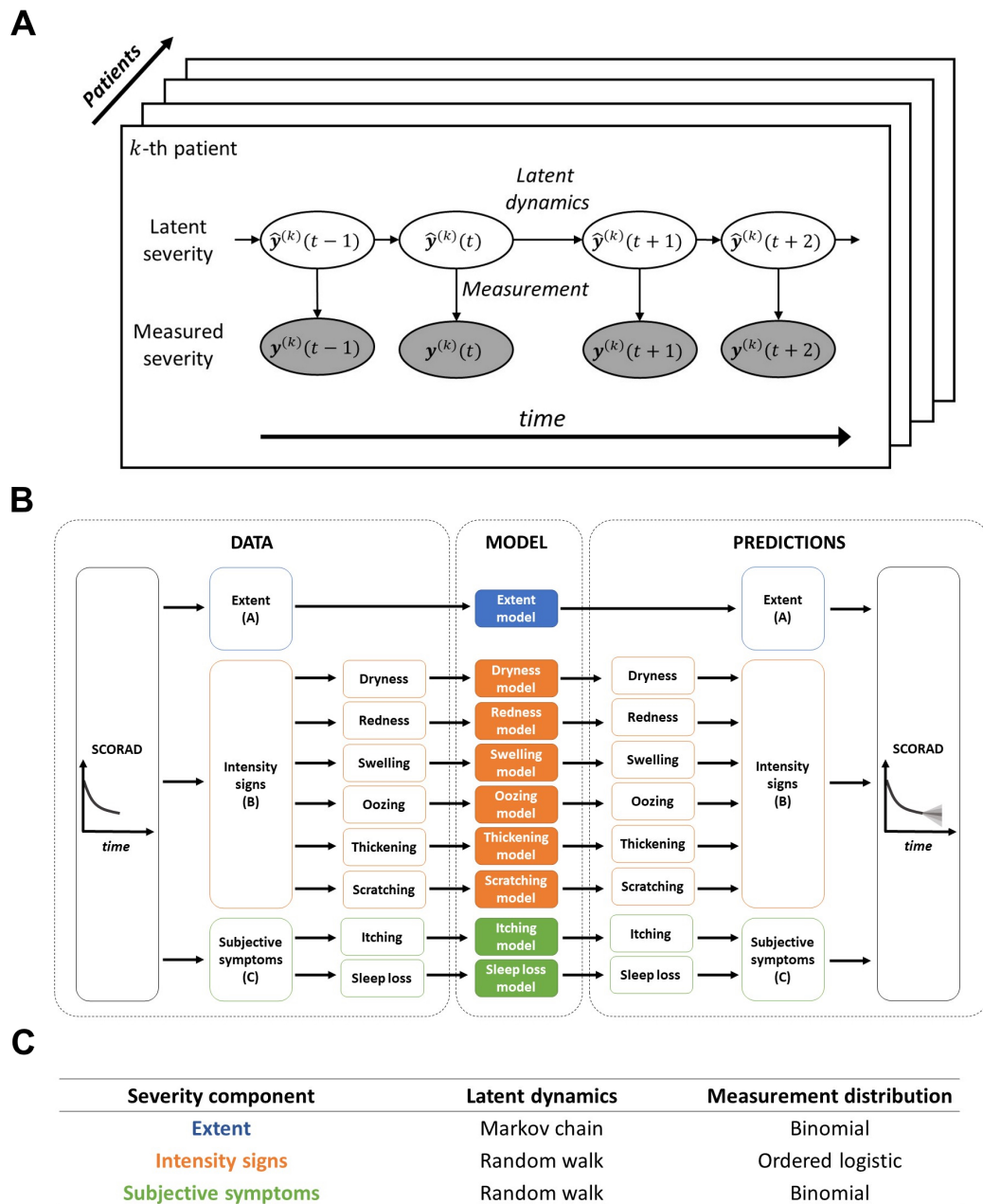


Figure 7.2: PO-SCORAD model overview. A) Bayesian state-space models in EczemaPred. Each model describes the dynamics of a latent severity (white ovals) and the measurement of the latent severity to obtain the recorded severity (grey ovals). B) Use of EczemaPred for SCORAD prediction. Predictions from nine models (coloured rectangles), each of which corresponds to one of the nine severity items for SCORAD, are aggregated to provide predictions for SCORAD. C) Latent dynamics and measurement distributions for the three severity components of SCORAD.

$$\text{covariance } \Sigma: \tilde{\mathbf{y}}^{(k)}(t+1) \sim \mathcal{N}(\tilde{\mathbf{y}}^{(k)}(t), \Sigma).$$

We considered different combinations of the measurement distributions and latent dynamics, with the most flexible model a priori being the one with an ordered logistic measurement distribution and a latent multivariate random walk dynamics (“OrderedMRW+corr”).

Priors for the PO-SCORAD and POEM models were chosen to be weakly informative. Details of the models, parametrisations, and choice of priors are described in Appendix E.1.

Model inference was performed using the Hamiltonian Monte Carlo algorithm in the probabilistic programming language Stan [56], with four chains and 2000 iterations per chain, including 50% burn-in. Prior predictive checks and fake data checks were conducted.

7.2.3 Model validation

We evaluated the predictive performance of our models in a forward-chaining setting (see Section 2.3.1), with a horizon of four days for PO-SCORAD and one week for POEM.

The probabilistic predictions of individual severity items and aggregate severity scores were evaluated using a logarithmic scoring rule, the log predictive density (lpd) [125]. We also computed an accuracy metric for PO-(o)SCORAD predictions, defined as the probability that the predictions were within 5 units of the measured score. We plotted the lpd and the accuracy as a function of the number of training observations (equivalently the number of training days/weeks) to produce learning curves. Details of the performance metrics are given in Appendix E.2.

We compared the predictive performance of EczemaPred with that of reference models, including a uniform forecast, a historical forecast, and a random walk model (which provides a flat forecast, i.e. centred on the last observed value). For the six intensity signs and the seven POEM symptoms that respectively take discrete values in $[0, 3]$ and $[0, 4]$ (4 and 5 categories), we used Markov chain models instead of random walk models as references. For PO-(o)SCORAD and POEM predictions, we also compared the performance of EczemaPred to that of standard time-series forecasting models, including an exponential smoothing model, an autoregressive model and a mixed effect autoregressive model. Details of the reference models are given in Appendix E.3.

each symptoms, that could be ran independently. This is not the case with a multivariate dynamic since all the symptoms are modelled jointly: there is only one model.

7.3 Results of PO-SCORAD models

All EczemaPred models and reference models were fitted successfully for all severity items on the two datasets. We found no evidence for an absence of convergence by monitoring trace plots and by checking the potential scale reduction factor \hat{R} . We conducted posterior predictive checks and found no clear discrepancies between the data and the models' simulations.

7.3.1 Predictions of severity items

We confirmed EczemaPred models learned the dynamics of severity items as more data came in (Figs. E.4 to E.12, top). A similar predictive performance was confirmed for the models trained with dataset 1 and those with dataset 2, supporting the generalisability of the models. However, predictions of extent and itching appeared to be more difficult with dataset 2 than with dataset 1 (Figs. 7.3, E.4 and E.11). For example, the lpd for predicting extent is much higher for the EczemaPred model trained with dataset 1 than with dataset 2 (-1.53 ± 0.07 vs -2.62 ± 0.14) after training with 80% of the data.

EczemaPred models outperformed the reference models for the two datasets in terms of predictive performance (Figs. 7.3 and E.4 to E.12, top). EczemaPred models showed only marginally better predictive performance than the historical forecasts for thickening, swelling and oozing. The lpd of the historical forecast was already close to the maximum lpd of 0 for these intensity signs that had a low prevalence (Figs. E.2 and E.3) and thus were easier to predict than other signs. A historical forecast tended to outperform a random walk model for extent and subjective symptoms, as they do not demonstrate persistent dynamics. For intensity signs whose dynamics are often persistent, a Markov chain model performed almost as well as the EczemaPred model with an ordinal logistic measurement and latent random walk.

The predictive performance decreased as the prediction horizon increased for all models. The decrease in lpd when the prediction horizon is increased by a day was similar or smaller for EczemaPred models compared to the reference models with a non-constant forecast (Figs. E.4 to E.12, bottom).

7.3.2 Predictions of PO-(o)SCORAD

Predictions for PO-SCORAD were derived by aggregating the predictions of the severity items by their respective models (example predictive trajectories in Fig. 7.4). The prediction intervals capture the evolution of the observed severity, but appear shifted/lagged compared

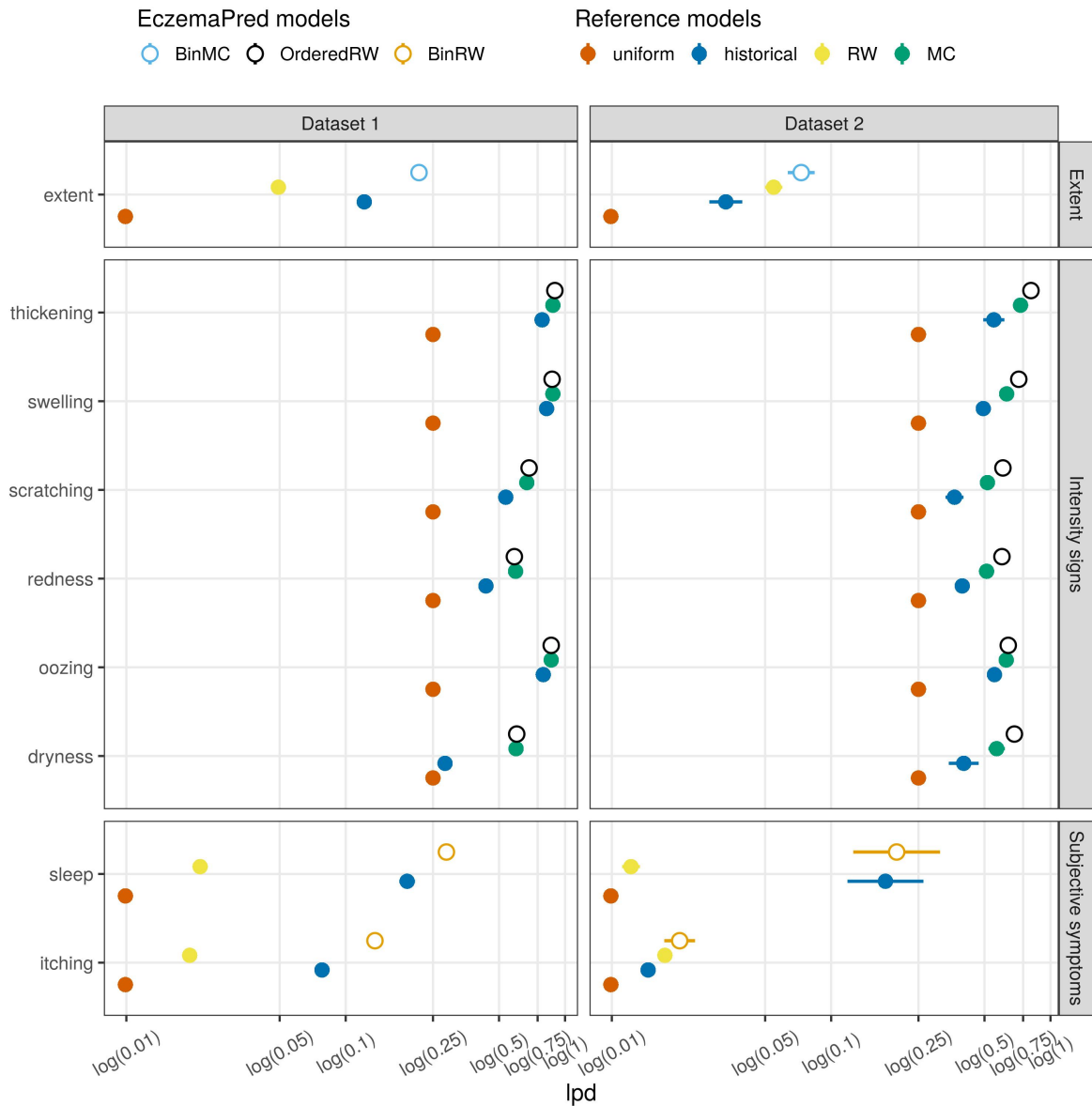


Figure 7.3: Predictive performance for 4-days-ahead forecasts by EczemaPred models (empty circles) and reference models (filled circles) measured by lpd (the higher, the better). EczemaPred models are a binomial Markov chain model (BinMC) for extent, an ordered logistic random walk model (OrderedRW) for intensity signs, and a binomial random walk model (BinRW) for subjective symptoms. Reference models include a uniform forecast (uniform), a historical forecast (historical), a random walk model (RW), and a Markov chain model (MC). The performance was calculated after training with approximately 80% of the data (77 days' data for dataset 1 and 65 days' data for dataset 2).

to the observed data. This is because the model is only updated every four days, there are no mechanisms to predict short-term trend (the model is more reactive than proactive), and the prediction intervals likely lack coverage because correlations between severity items were not modelled (see Section 7.4).

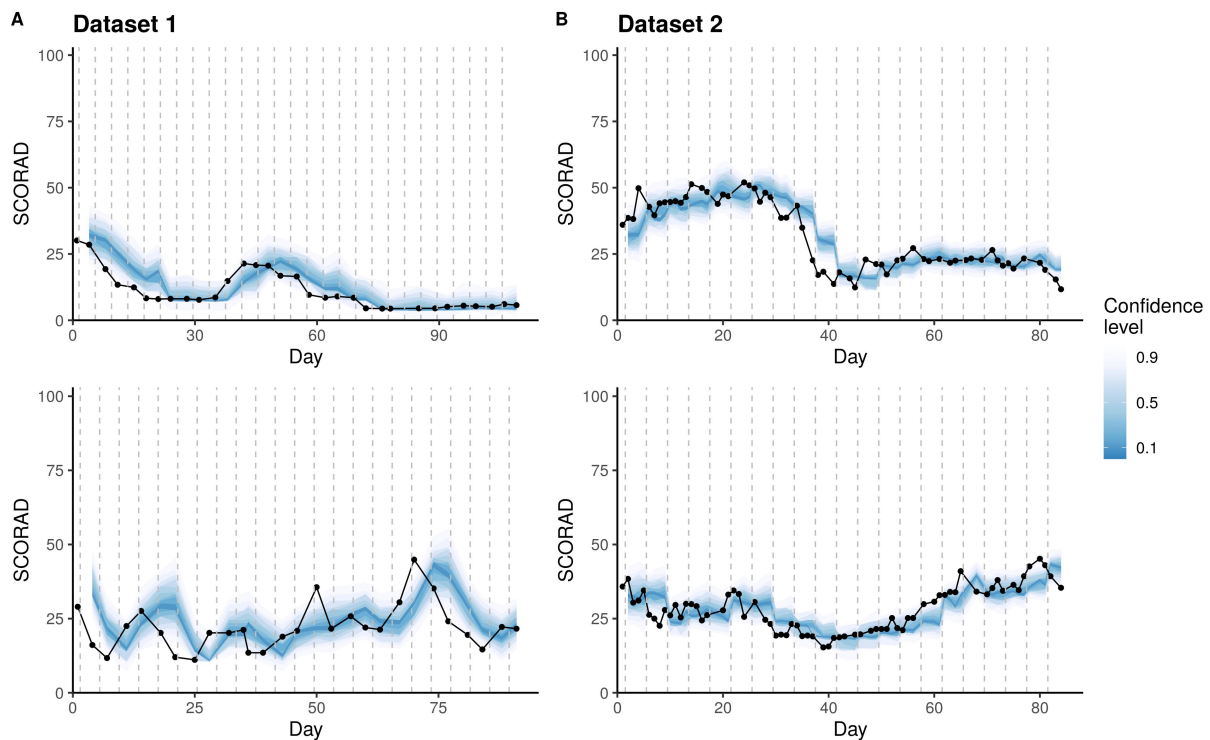


Figure 7.4: PO-SCORAD prediction by EczemaPred for four representative patients from dataset 1 (A) and dataset 2 (B). The patients whose data is depicted on the top plots are the same as the patients in Fig. 7.1. Coloured ribbons correspond to stacked prediction intervals of highest density (darkest ribbon corresponds to the mode), and black dots are the recorded PO-SCORAD. The model is updated, and new predictions are issued every four days (vertical dashed lines).

We confirmed that the performance of PO-SCORAD prediction by EczemaPred improved as more data came in but did not plateau (Fig. 7.5). It suggests a possibility of further improvement of the performance if more training data were available and a need for more accurate estimation of some model parameters. In contrast, the performance of the reference models stopped improving, suggesting that the improvement observed for EczemaPred was not due to a change in the data distribution (e.g. due to patients dropping the study early).

EczemaPred outperformed the reference models that predict PO-SCORAD directly (rather than aggregating the prediction of severity items as in EczemaPred), supporting our approach. The difference in the lpd between EczemaPred and the reference models is less evident in dataset 2 than in dataset 1 for PO-SCORAD prediction. However, the difference is more evident for PO-oSCORAD prediction (Fig. E.13), because of the lower predictive performance for subjective symptoms with dataset 2 than dataset 1. Otherwise, similar results with comparatively better predictions were obtained for PO-oSCORAD.

The exponential smoothing and the (mixed) autoregressive models achieved similar predictive performance to the random walk model (Fig. 7.5) as they tend to emulate a random walk behaviour. That is, the exponential smoothing model has a smoothing factor of 1, and the autoregressive models have an autocorrelation parameter of 1 and an intercept of 0. The fact that complex models emulate a simpler random walk model highlights the difficulty of developing accurate predictive models using only the aggregate PO-SCORAD data.

The learning curves (Fig. 7.5) indicate that EczemaPred achieved the accuracy of $71.8 \pm 1.0\%$ (mean \pm standard error) after training with 77 days' data from dataset 1 ($60.2 \pm 2.8\%$ with 65 days' data from dataset 2). That is, the 4-days-ahead prediction by EczemaPred is within 5 units of the measured PO-SCORAD with 71.8% probability, on average. The accuracies of the reference models were much lower, with $39.4 \pm 0.5\%$ for the historical forecast with dataset 1 ($34.9 \pm 1.4\%$ with dataset 2), $47.9 \pm 0.5\%$ ($43.1 \pm 1.2\%$) for the random walk model, and $51.9 \pm 0.7\%$ ($49.5 \pm 1.9\%$) for the mixed effect autoregressive model. It is worth noting that the improvement of the accuracy from the random walk model to EczemaPred (+23.9%, +17.1%) is much larger than that from the historical to the random walk model (+8.5%, +8.2%), although “the marginal gain from complicated models is typically small compared to the predictive power of the simple models” [53].

The predictive performance of EczemaPred with dataset 1 appeared to be better than that with dataset 2 (Fig. 7.5), although the predictions do not always appear qualitatively different between the two datasets (Fig. 7.4). Several data characteristics (e.g. dataset size, frequency of measurements, demographics) may explain the difference, but it is difficult to pinpoint the main factors without a meta-analysis. It is also possible that the performance with dataset 2 becomes comparable or superior to that with dataset 1, if we allow for a more prolonged training phase, given that the performance did not plateau.

The predictive performance of EczemaPred models and the reference models decreased as the prediction horizon increased (Figs. E.14 and E.15), similarly to what was observed for individual severity items. The accuracy of EczemaPred was estimated to decrease by approximately $3.0 \pm 0.2\%$ on average when the prediction horizon was increased by a day in dataset 1 ($3.9 \pm 0.5\%$ in dataset 2). It leads to the accuracy of $80.7 \pm 1.2\%$ and $62.8 \pm 1.1\%$ for one-day-ahead and one-week-ahead forecast, respectively, for dataset 1 ($71.9 \pm 2.9\%$ and $48.6 \pm 3.4\%$ for dataset 2). These results suggest that EczemaPred performance would become equivalent to a historical forecast for 15.2 days-ahead predictions and 10.4 days-ahead predictions for datasets 1 and 2, respectively, assuming that the extrapolation of accuracy loss is valid. A similar decrease in accuracy was estimated for the reference models.

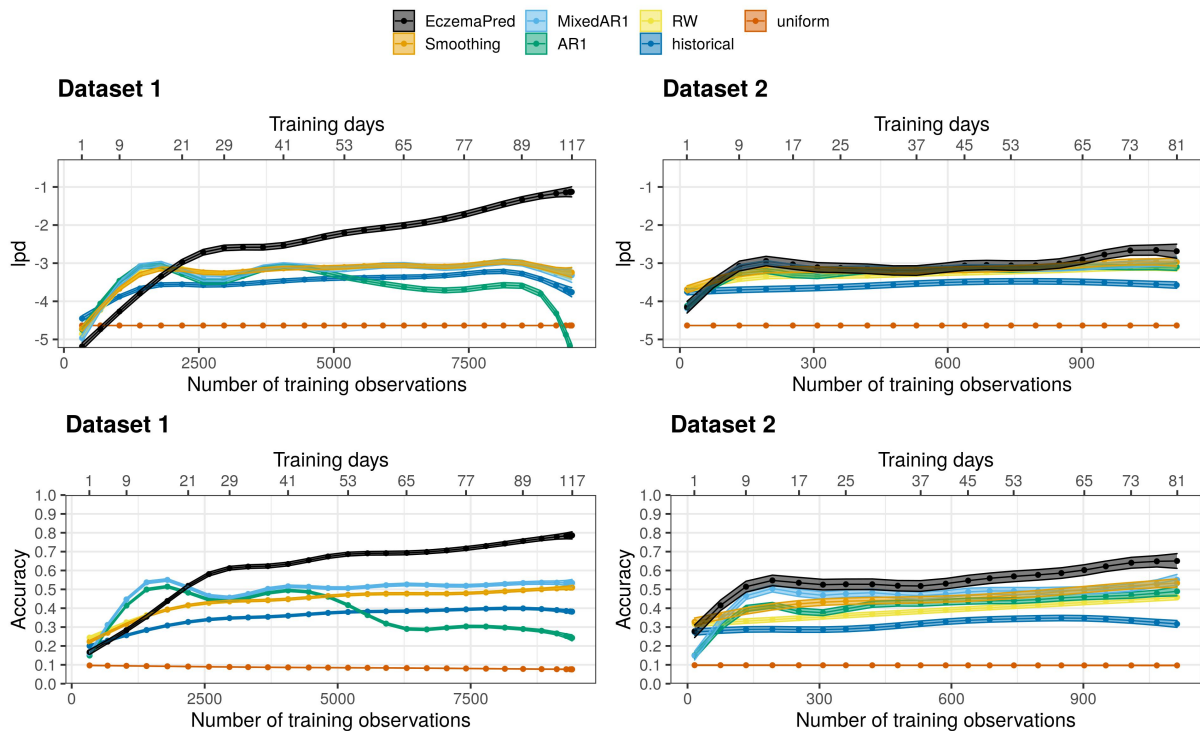


Figure 7.5: Learning curves for 4-days-ahead forecasts of PO-SCORAD, evaluated by lpd (top) and accuracy (bottom), as a function of the number of training observations (training days), for datasets 1 (left) and 2 (right). EczemaPred models perform better than reference models, including an exponential smoothing model (Smoothing), a mixed effect autoregressive model (MixedAR1), an autoregressive model (AR1), a random walk model (RW), a historical forecast (historical) and a uniform forecast (uniform).

7.3.3 Decomposition of prediction uncertainty in EczemaPred

We investigated which of the three components of PO-SCORAD (0.2A, 3.5B or C) contributed to the uncertainty in PO-SCORAD prediction the most, by computing the proportion of the variance of each component to the variance of the PO-SCORAD, for each prediction. On average, 7% of the uncertainty in PO-SCORAD prediction could be attributed to the extent (0.2 A), 79% to the intensity signs (3.5 B) and 14% to the subjective symptoms components (C) for dataset 1 (5%, 72% and 23% for dataset 2). In contrast, the intensity signs component is $63/103 \approx 61\%$ of the total SCORAD with extent and subjective symptoms each contributing to $20/103 \approx 19\%$. Accordingly, improving predictions for intensity signs is the most promising option to improve predictions for PO-SCORAD. The intensity signs that contribute to the prediction uncertainty the most were calculated to be dryness, redness, and scratching, the other signs being less prevalent.

7.4 Results of POEM models

All models were fitted successfully, except for the models with the \mathcal{B}_{Day} measurement distribution, for which numerical errors (divergences) were observed during inference. This may suggest that the \mathcal{B}_{Day} distribution does not describe the measurement process well for this dataset, although it was tailored to the a priori data-generating process of the symptom scores for POEM. Instead, an ordered logistic measurement distribution may be preferred. By using multivariate models for POEM, we inferred that changes in the latent symptom scores were strongly positively correlated (Fig. 7.6A), which implies that when a symptom increases or decreases, it is likely that the intensities of the other symptoms change in the same direction. Posterior predictive checks did not indicate clear failings in the data models (Fig. 7.6B for the model with ordered logistic measurement distributions and latent multivariate random walk).

The predictive performance of the EczemaPred models increased as more data came in (Fig. E.16). EczemaPred models outperformed or performed similarly to all reference models when the predictive performance was stable (Fig. 7.7). However, the difference in performance between EczemaPred models and reference models is small and does not always appear practically significant.

Nonetheless, we observe a systematic benefit of modelling correlations between symptoms when predicting POEM, regardless of the measurement distribution⁴. This is likely because

⁴While modelling the correlations between changes in latent scores is of interest for inference purposes (quantifying correlations) and results in a small gain predictive performance, it comes at a high computational cost (50 times longer than running independent models for each symptom, more if we consider that independent

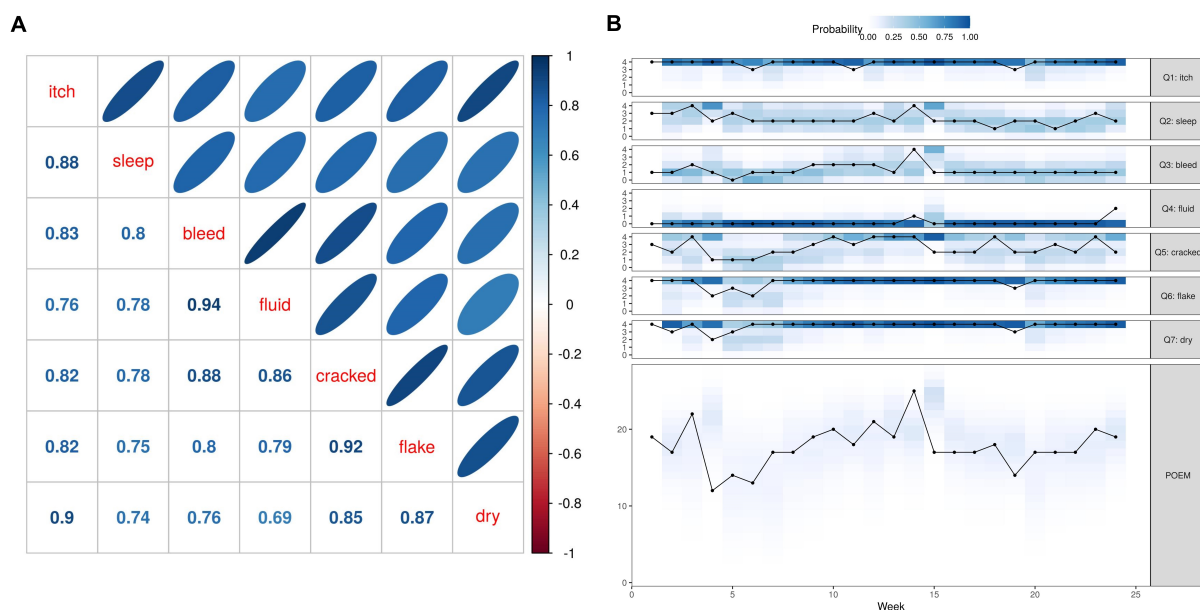
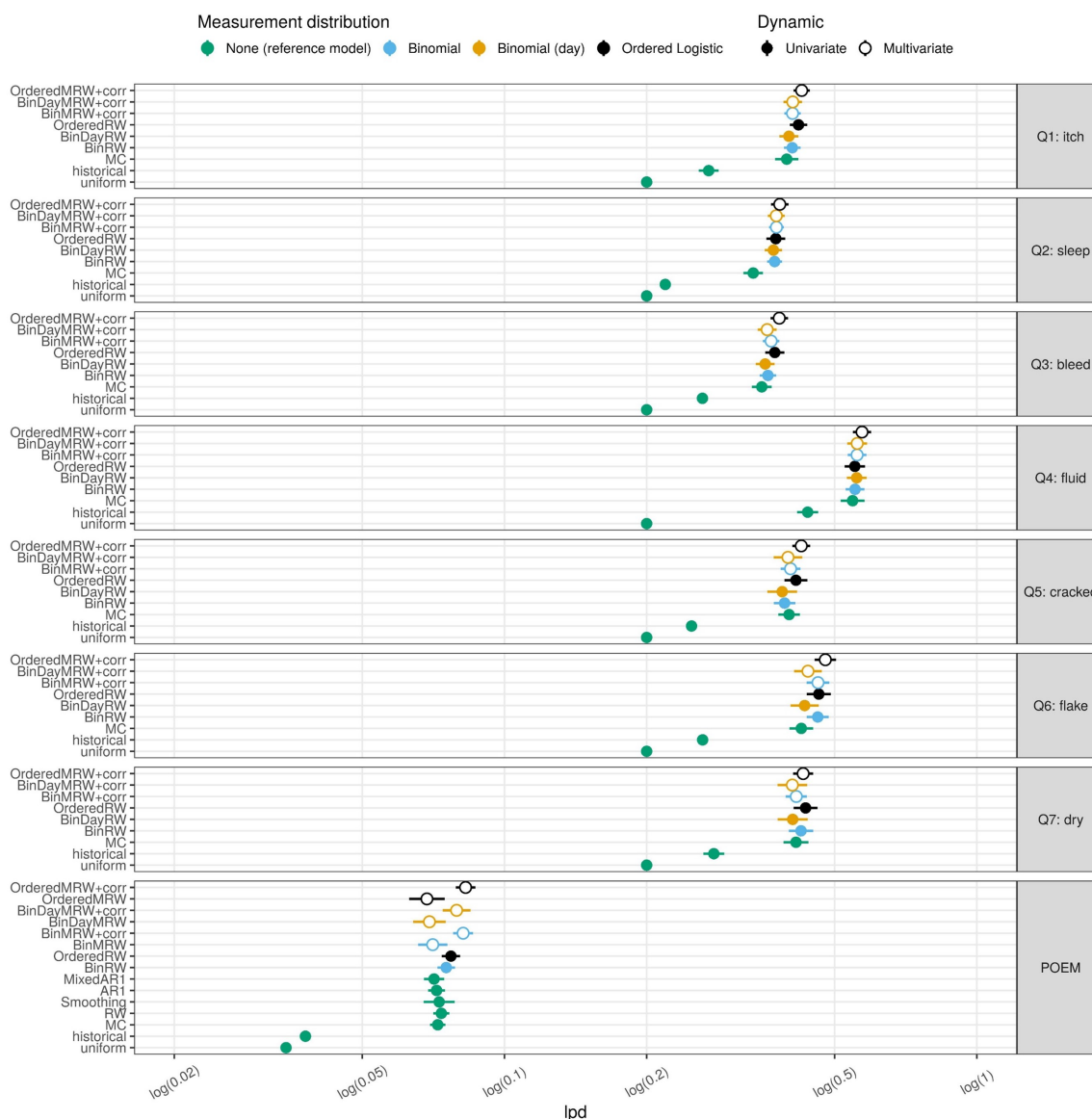


Figure 7.6: Results of the model with an ordered logistic distribution and multivariate latent random walk dynamic. (A) Correlation of changes between latent symptoms scores. The lower part of the diagonal indicates the expected correlation coefficients, and the upper part illustrates the correlation as an ellipsoid. (B) Data of a representative patient and corresponding predictive distribution (colour) for one-week-ahead forecast, for all symptoms and the aggregate POEM score. For this patient, even though 4/7 symptoms are easy to predict because they do not change a lot, the fluctuations of the 3 other symptoms make the prediction of POEM challenging. Symptom changes appear correlated, for example at week 4, or at week 14, when a few symptoms decrease, resp. increase, in intensity.

modelling correlations between symptoms, rather than assuming they are independent, results in better uncertainty estimates (calibration). More specifically, the variance of POEM with positively correlated errors between symptoms is higher than the variance of POEM assuming independent errors, suggesting that the latter model was overconfident in its uncertainty estimates. This is likely why POEM prediction intervals in Fig. 7.6B appear to have better coverage than PO-SCORAD prediction intervals in Fig. 7.4, because changes in PO-SCORAD are often due to changes in multiple severity items (cf. Fig. 7.1). We also observe that the models with ordered logistic measurement distributions tend to perform better than those with Binomial distributions, probably because the former has a more flexible shape than the latter. We do not observe a difference in performance between the models with a Binomial measurement distribution on days (\mathcal{B}_{Day} , prefix “BinDay” in Fig. 7.7) and the models with a Binomial measurement distribution on the symptom scale (prefix “Bin”), further suggesting the additional complexity of predicting the number of days may not be useful here.

The predictive performance is roughly similar across the seven symptoms, except for weeping skin/oozing clear fluid (question 4) for which predictions are more accurate because this is the least prevalent symptom (absent in 75% of observations, resulting in a higher historical forecast). All symptoms have a similar marginal contribution to the POEM prediction uncertainty.

models can be parallelised), which may be prohibitive if the model is deployed “as is” in the real world.



Reference models: “uniform” (uniform forecast), “historical” (historical forecast), “MC” (Markov Chain model), “RW” (Random Walk model), “Smoothing” (exponential smoothing model), “AR1” (autoregressive model of order 1), “MixedAR1” (mixed effect autoregressive model of order 1).

EczemaPred models prefixes: “Bin*” (Binomial measurement), “BinDay*” (Binomial measurement on the number of days), “Ordered*” (ordered logistic measurement).

EczemaPred models suffixes: “*RW” (univariate latent random walk), “*MRW” (multivariate latent random walk), “*+corr” (for multivariate models only, indicates that the correlation between symptoms is modelled). NB: Multivariate models where the correlation between symptoms is not modelled are equivalent to their corresponding univariate model. For example, for symptoms prediction “OrderedMRW” is equivalent to “OrderedRW” and “OrderedMRW” predictions for POEM are equivalent to combining predictions of the “OrderedRW” for each symptom.

Figure 7.7: Predictive performance estimates (lpd, \pm SE, x-axis, the higher the better) for one-week-ahead of the different symptoms and POEM (facet) by several models (y-axis). The models are broadly classified by their measurement distribution (colour), with green estimates corresponding to the reference models and other columns to different setups of EczemaPred models; and whether their dynamic is univariate or multivariate (filled or unfilled circle). The performance was calculated after training with the data up to 23 weeks (96% of the data).

7.5 Discussion

7.5.1 Main findings

This chapter introduced EczemaPred, a computational framework to predict the patient-dependent dynamic evolution of AD severity (Fig. 7.2). We used EczemaPred to predict PO-SCORAD using two independent datasets with different characteristics, and to predict POEM using a third dataset. Our approach consisted in modelling the evolution of individual severity items (9 items for PO-SCORAD: extent, 6 intensity signs and 2 subjective symptoms; 7 symptoms for POEM) using Bayesian state-space models. Predictions for the severity scores (PO-SCORAD and POEM) could then be produced by aggregating the predictions of the constituent severity items.

For PO-SCORAD, EczemaPred models outperformed the reference models we considered for all the severity items and the aggregate PO-SCORAD (Figs. 7.3 and 7.5). The prediction accuracy was approximately 72% and 60% for 4-days-ahead forecasts for datasets 1 and 2, respectively. Most of the prediction uncertainty in PO-SCORAD (79% and 72% for datasets 1 and 2, respectively) could be attributed to the intensity signs component, suggesting that improving predictions of the intensity signs is the most promising approach to improve PO-SCORAD predictions. In contrast to our model predicting PO-SCORAD, the performance of our EczemaPred model for predicting POEM was low, relative to the reference models (Fig. 7.7). This may suggest that POEM is harder to predict than PO-SCORAD, and if it is predictable, more sophisticated models would be required for POEM compared to PO-SCORAD.

7.5.2 Choosing the right score for severity prediction

Our results highlight that not all severity scores are equivalent when it comes to predictions. In particular, PO-SCORAD may be a better alternative than POEM for developing accurate prediction models using observational data. PO-SCORAD may nonetheless come at a cost to patients, as it is more time-consuming to assess compared to POEM, although assessment duration tend to decrease with experience [91].

We can speculate on why POEM appears harder to predict than PO-SCORAD. First, POEM consists of subjective symptoms that do not necessarily correspond to a true clinical disease state, whereas PO-SCORAD, or at least PO-oSCORAD (the objective part of PO-SCORAD), consists of objective physical signs. Supporting this, we have observed limited performance when predicting a subjective “Bother” score (Chapter 4) and the subjective symptoms (itch and sleep disturbance) of PO-SCORAD. The difficulty in predicting POEM could also be explained

by the characteristics of the scoring system. POEM is less responsive to changes [102] and subject to more measurement errors than other scores such as SCORAD (Chapter 6). POEM is a weekly score and its one-week ahead predictions are more challenging than predictions of only a few days ahead for PO-SCORAD. POEM can also be subject to recollection bias as it is assessed based on the presence of symptoms in the last seven days. In addition, POEM assigns equal weights to the presence of symptoms during the past week, even though recent measurements (e.g. yesterday) tend to be more informative than old measurements (e.g. seven days ago) for prediction. By compressing the daily presence or absence of symptoms into a weekly average, potentially useful information for prediction is lost in POEM. The severity of symptoms is also missing in POEM, even though distinguishing mild and severe symptoms, for example, could be useful for predictive models.

7.5.3 Strengths of our approach

Modelling the dynamics of each severity item has several advantages when the breakdown of the aggregate severity score is available. It enables us to extract more signals from the data, as the AD severity dynamics for each patient are described by multiple time-series, one for each severity item (nine with PO-SCORAD, seven with POEM) instead of one for the aggregate score. This approach also reduces the uncertainty in the aggregate score prediction when some severity items are easier to predict than others (e.g. when they are not very prevalent or do not vary much over time). The models can be tailored to each severity item to reflect the item-dependent data generating mechanisms with relevant measurement processes and latent dynamics (cf. extent). The models are thus more interpretable and transparent, as predictions for aggregate severity scores can be decomposed into predictions for their components [36]. The models could be used to predict any combination of the severity items (e.g. PO-oSCORAD) without potential inconsistencies in predictions that could arise if each severity score of interest (e.g. oSCORAD and EASI with overlapping severity items) is modelled separately. EczemaPred can thus be applied to develop predictive models for other AD severity scores, such as EASI.

EczemaPred has some further advantages, especially for clinical use. The Bayesian framework enables us to make probabilistic predictions by explicitly quantifying uncertainties in parameters and predictions. The state-space models explicitly describe potential and often inevitable errors in the measurement of the severity items. For example, the estimation of the body area affected by eczema is subject to high inter-rater variability [93], potentially even more so when it is self-assessed as in PO-SCORAD [87]. Modelling the measurement processes separately from the latent dynamics of the disease severity also allows us to deal with missing values efficiently as an absence of measurement, while still inferring the latent dynamics. In a practical application of the model, the posterior distributions obtained in this study could be

used as a prior for new patients to “pre-train” the model, shortening the training phase to only a few measurements (this will be done in Chapter 8).

7.5.4 Limitations and future directions

Limitations of this study include the subjective assessment of PO-SCORAD and POEM by patients. For instance, the reliability of PO-SCORAD assessment was shown to improve with experience, as patients may need time to learn how to use the PO-SCORAD instrument properly [91]. The severity item models may therefore benefit from specifying a time-varying measurement error (this will be done in Chapter 8). Computational limitations also have prevented us from modelling the correlations between PO-SCORAD items⁵. Even though the components of SCORAD (extent, intensity signs and subjective symptoms) are thought to be uncorrelated by design [80], the six intensity signs may be correlated. For instance, dryness, thickening and scratching may covary as they mainly characterise the chronicity of the disease; and redness, swelling and oozing may covary as they represent acute flares [80]. Validation of EczemaPred in a real-world evidence study is also required, as the data used in this study were taken from patients involved in a clinical study, where they may have had a better follow-up than usual.

In summary, this chapter introduced EczemaPred as a computational framework to predict the patient-dependent dynamic evolution of AD severity. Patients could benefit from EczemaPred in managing their disease and anticipate the change of their symptoms. For example, the models could be extended to quantify patients’ responsiveness to treatment and suggest personalised treatment strategies using Bayesian decision theory, as shown in Chapter 8.

⁵For instance, dataset 1 is much bigger than the POEM dataset, making inference longer. Correlations will be modelled using dataset 2 in Chapter 8.

Chapter 8

Towards generating treatment recommendations

In Chapter 7, we presented EczemaPred, a principled approach to developing predictive models of AD severity scores. After having successfully validated the predictive models for PO-SCORAD, in this chapter, we turn our focus towards generating treatment recommendations and use the EczemaPred models as a starting point to estimate treatment effects. In this chapter, we also aim to provide useful inferences and make the models more relevant to clinical practice, by integrating other sources of information beyond severity scores.

We are grateful to Pierre Fabre Laboratories for sharing the data used in this study. The code written for this project is not released at the time of writing (January 2022).

8.1 Introduction

AD cannot be cured, but the condition can be managed using treatments such as topical corticosteroids and emollients. Yet, responses to treatment vary from patient to patient and more research is needed to go beyond the “one-size-fits-all” approach to therapy and towards designing personalised treatment strategies for AD [11].

In Chapter 7, we presented EczemaPred, a computational framework to predict the evolution of eczema severity, and used it to develop models for PO-SCORAD and POEM. EczemaPred consists of a collection of Bayesian state-space models that are used to predict individual severity items, which can then be aggregated to produce consistent predictions for different severity scores. EczemaPred models address multiple challenges of working with eczema

severity data, that are usually longitudinal data of discrete scores, and contain irregular and imperfect measurements (including missing values). The models can quantify uncertainty and are modular, with the possibility to specify different combinations of latent dynamics and measurement distributions. As such, the models can be used as a starting point to analyse time-series of eczema severity data and then be adapted to the specific aims of an analysis.

Here, we extend EczemaPred models to investigate treatment effects and generate personalised treatment recommendations. Making predictions, treatment effect inferences and recommendations with the same model has the benefit of producing consistent results¹, which would not necessarily be the case if we used different models/analytical methods, or interpreted a posteriori the decisions of a black-box treatment recommendation algorithm [35]. In the context of making decisions (treatment recommendations) under uncertainty, where the decisions cannot be postponed until more evidence is collected, it is also desirable for models to integrate and combine all available information to support decision-making [191].

In this chapter, we build upon the EczemaPred model to predict PO-SCORAD shown in Chapter 7, to demonstrate how to generate treatment recommendations for AD, while integrating multiple sources of information into a comprehensive and coherent Bayesian model. In particular, we estimate the efficacy of topical corticosteroids and emollient cream and produce treatment recommendations using Bayesian decision theory. We also incorporate evidence from prior studies into our model to reach more robust conclusions and kickstart model training. Finally, we use clinical assessments (SCORAD) to calibrate self-assessments² (PO-SCORAD), to improve the quality of training data and build trust in the model's output. While PO-SCORAD (assessed by patients) can be measured daily, it is supposedly less accurate than SCORAD (assessed by clinicians) which can only be measured infrequently when patients visit a clinic. Calibrating PO-SCORAD with SCORAD measurements can thus get the best of high quality (SCORAD) and high-frequency measurements (PO-SCORAD), providing estimates of measurement biases and what would have been SCORAD if measured daily.

8.2 Methods

We used data containing PO-SCORAD, SCORAD and treatment usage as well as knowledge from a prior study to fit a model using EczemaPred (Fig. 8.1A). The model produces inferences about treatment effects, the dynamics of AD severity items, measurement biases, as well as

¹For example, if we investigate two treatments A and B and found that A is more effective than B on average, we would expect a better prognostic for A than B, and that A is more likely to be recommended compared to B. This consistency of results can be seen as a minimum requirement for the reproducibility of results.

²Here we use the term “calibration” to refer the calibration of measurements, as opposed to calibration of probabilities, which refers to accuracy of probabilistic forecasts.

predictions for future PO-SCORAD and SCORAD scores. Predictions can then be used to generate treatment recommendations. A summary of this pipeline is given in Fig. 8.1B.

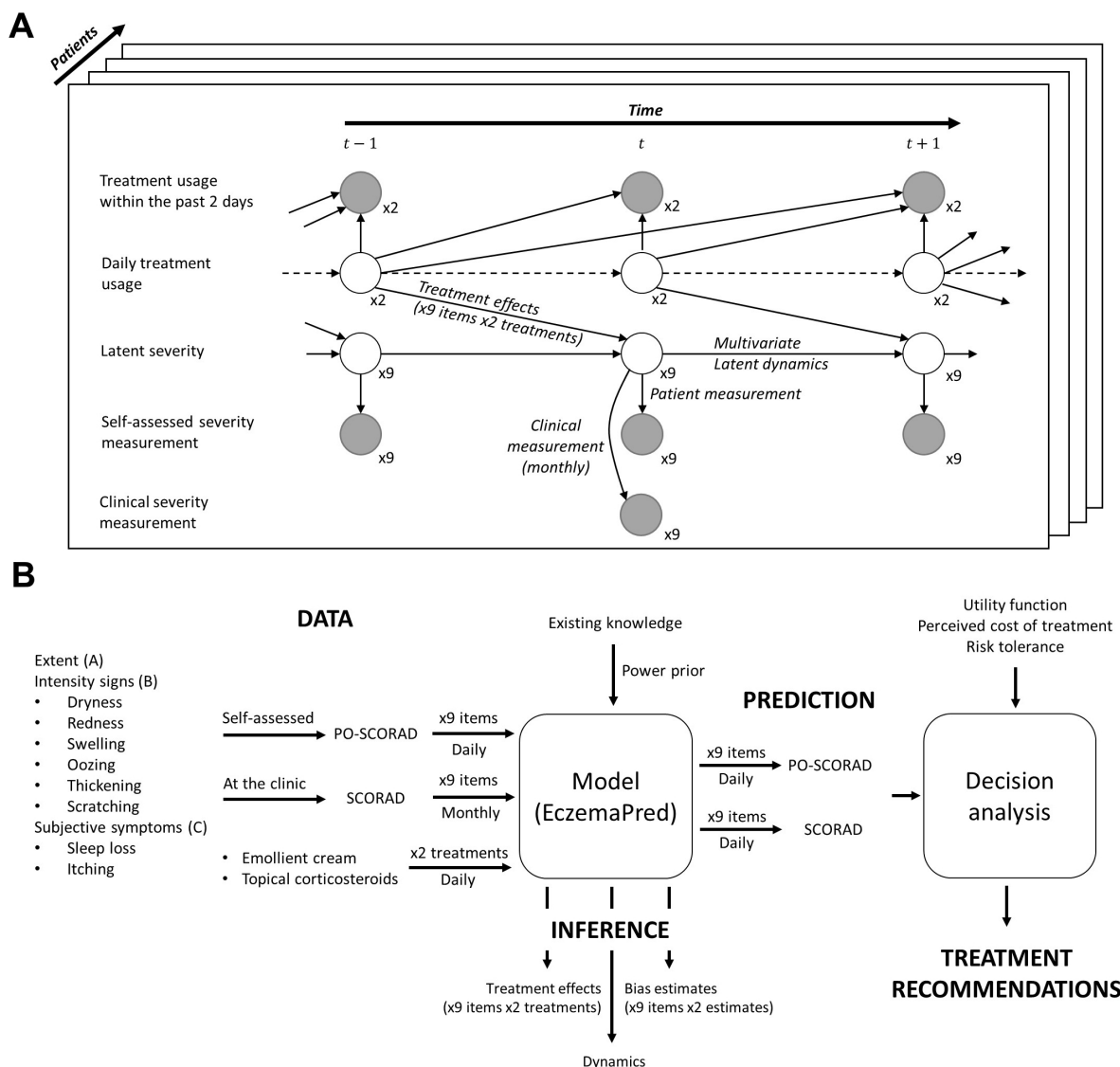


Figure 8.1: A) Schematic of the Bayesian state-space model. Grey and white circles correspond to observed and latent variables, respectively. B) Overview of the method. The data (PO-SCORAD, SCORAD, treatment) and existing knowledge, in the form of a power prior, are given as inputs to the model to produce inferences about treatment effects, the dynamics and biases between SCORAD and PO-SCORAD. Predictions can also be generated, and when a utility function is supplied, they can be used to generate treatment recommendations.

8.2.1 Data

In this chapter, we used “dataset 2” from Chapter 7, which has been described in Section 7.2.1. Briefly, in this dataset, 16 patients recorded PO-SCORAD daily for up to 12 weeks (84

days), resulting in 1136 patient-day observations (13.6% missing values between the first and last observations). The dataset also includes SCORAD, measured monthly by trained clinical staff. Whether treatment was used within the past two days was also recorded daily in the PO-SCORAD application, for topical corticosteroids and emollient cream. Example data from a representative patient is shown in Fig. 8.2.

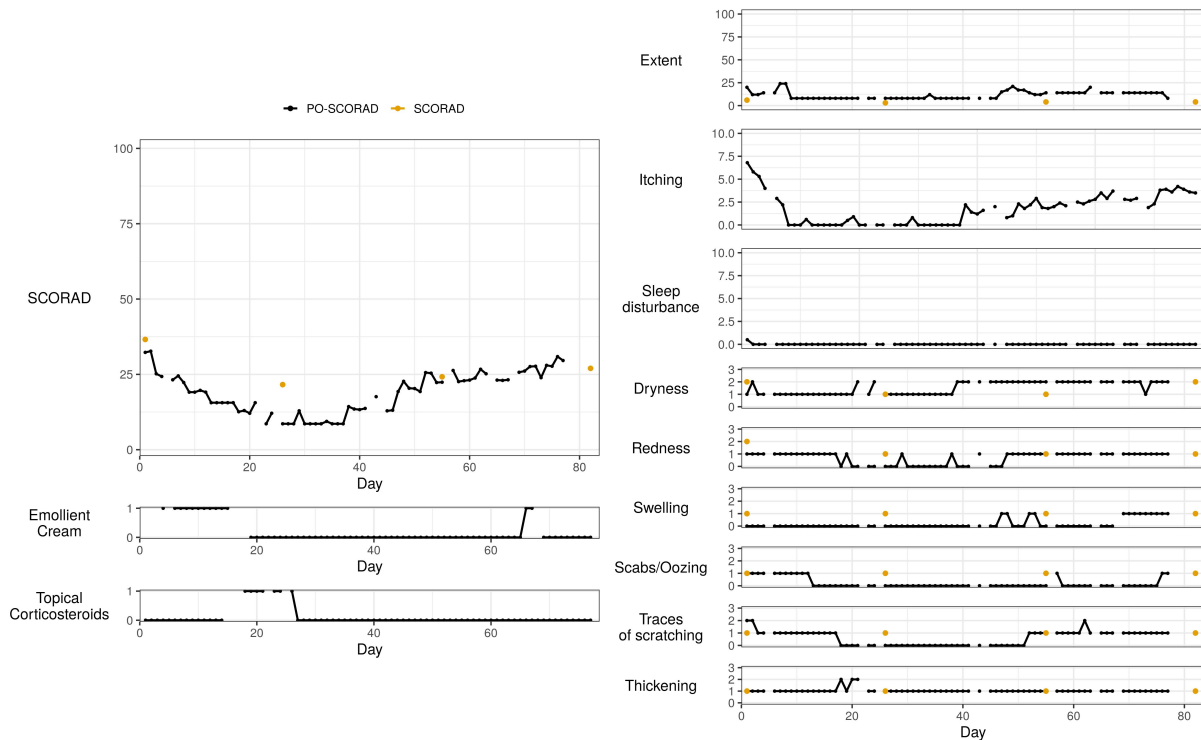


Figure 8.2: Data from a representative patient. On the left the trajectories of the aggregate PO-SCORAD severity score, emollient cream and topical corticosteroids usage within the past two days. On the right, the trajectories of the different severity items of PO-SCORAD. Orange dots correspond to the monthly SCORAD measurements. The lines are broken when the measurements are missing.

8.2.2 Model

We used the previously developed EczemaPred PO-SCORAD model (Chapter 7), and extended it by integrating clinical (SCORAD) measurements and treatment usage data, and modifying the latent dynamics of AD severity items (Fig. 8.1A).

The original EczemaPred PO-SCORAD model consisted of nine independent Bayesian state-space models, one for each severity item, where the observed severity items are the imperfect measurement of a latent score (representing a “true severity”) for which we assumed a latent random walk dynamic. In this chapter, we modelled the severity items jointly (there is one model) by assuming a multivariate latent dynamic, where changes in the latent severity items

are correlated (as we did for POEM prediction in Chapter 7). We also decided to model the measurement of all severity items using ordinal logistic distributions, as it provides a way to control the variance/precision of the measurements (cf. Section 2.4), without degrading performance. This was not the case when using Binomial distributions for extent and subjective symptoms in Chapter 7. We adopted the second parametrisation of the ordinal logistic distribution used for POEM prediction in Chapter 7, as the parametrisation scales well to a higher number of categories, a requirement for modelling extent and subjective symptoms.

We integrated SCORAD measurements in the model by assuming that they are derived from the same latent score as PO-SCORAD items ($\hat{y}_i^{(k)}(t)$), but are biased compared to PO-SCORAD (the location of the distribution $\hat{y}_i^{(k)}(t)$ is shifted by the bias term). This bias depends on the severity item and can decrease with time, since it was found that PO-SCORAD assessments become closer to SCORAD's with experience [91]. Our model also assumed SCORAD measurements are more precise than PO-SCORAD measurements, by specifying a smaller variance for the measurement of SCORAD compared to PO-SCORAD. We did not calibrate PO-SCORAD subjective symptoms as they are the same as for SCORAD.

We included a trend and treatment response components to the latent dynamic of each severity item. The trend component corresponds to an exponential smoothing of the difference between consecutive latent severities. To obtain the treatment response component, first, we deconvolved the time-series of usage of corticosteroids and emollients within the past two days, using deterministic and probabilistic inference, to obtain the respective time-series of daily treatment usage. Then, we used the deconvolved time-series of daily treatment usage to model item-dependent treatment effects for corticosteroids and emollients, assuming treatment usage at t only influences the severity at $t + 1$.

Details of the model are available in Appendix F.1.

8.2.3 Priors

We used a power prior for the parameters corresponding to the measurement and latent dynamics [112]. The power prior is an informative prior constructed from historical data. Informative priors are useful when working with small data and can help kickstart the training of the model (the model is “pre-trained”). We used “dataset 1” described in Chapter 7 as historical data, which originates from an already published study investigating the role of an emollient in 337 children with AD [189]. The power prior was derived from the marginal posterior estimates of the population parameters of independent state-space models, with ordinal logistic measurement and latent random walk for each severity item. The power prior requires the selection of a discounting parameter $a_0 \in [0, 1]$, which quantifies how much information is

borrowed from the historical data (0 means no borrowing and 1 full borrowing). Considering that the historical data is much bigger (9943 patient-day observations) than the dataset in this study (1136 patient-day observations), we chose a small value $a_0 = 0.04$ to ensure that the final posterior is mostly determined by the data used in this study rather than the historical data (Fig. 8.3A).

Treatment parameters and those corresponding to the calibration of PO-SCORAD with SCORAD were given weakly informative priors. The correlation matrix and trend parameters were given priors that penalise the model complexity. Details of the priors, including the power prior, are given in Appendix F.2.

8.2.4 Treatment recommendation

Having developed a Bayesian model that can make predictions under different treatment conditions (actions), Bayesian decision analysis is the natural approach to balance the costs and benefits of using treatments and make optimal recommendations under uncertainty [58] [105]. Bayesian decision analysis consists of choosing a utility function that quantifies the “value” of taking a particular action, making predictions corresponding to different actions and recommending the action that maximises the expected utility (objective function) of the associated predictions³. The objective function can also include a risk-sensitive criterion to balance the benefit of the action (expected utility) and its risk (variance of utility, i.e. its uncertainty), which is also a way to balance the exploration-exploitation trade-off. A patient can be risk-averse (penalising uncertainty, or pessimistic), risk-neutral, or risk-seeking (welcoming uncertainty, or optimistic).

We used a simple utility function:

$$U(\hat{y}, a) = -(\hat{y} + \text{cost}(a)) \quad (8.1)$$

Here, \hat{y} is the predicted SCORAD, a corresponds to the action of using/not using topical corticosteroids/emollient cream and $\text{cost}(a)$ corresponds to the “perceived” cost of action a . The “perceived” cost could represent the fear of side-effects [76], the inconvenience or monetary cost of using treatment, and more generally any mechanisms that drive poor adherence. For example, if the cost of using no treatment is 0 and the cost of using corticosteroids is 1, a risk-neutral patient would only use corticosteroid if the expected improvement after taking corticosteroids is at most 1 point of SCORAD.

³The objective function is not the utility function itself, because the predictions are probabilistic, which implies a distribution of utilities.

We generated next-day treatment recommendations by successively training the model every day, and considered different decision profiles corresponding to different “perceived” costs of using treatment (no cost, normal cost, or high cost) and risk tolerance (risk-averse, -neutral or -seeking). More details are given in Appendix F.3.

8.2.5 Inference and validation

Model inference was performed using the Hamiltonian Monte Carlo algorithm in the probabilistic programming language Stan [56], with four chains and 2000 iterations per chain, including 50% burn-in. Prior predictive checks and fake data checks were conducted.

We consider as a base model the original EczemaPred model with ordinal logistic measurement distribution for all severity items (i.e. our model described in Section 8.2.2 but with independent dynamics and without calibration trend, treatments and power prior). In this study, we evaluate the contribution of each new model component in a stepwise approach, starting from the base model and successively adding the power prior, the correlation between severity items, the calibration of PO-SCORAD with SCORAD, treatment effects and the trend⁴. We report the performance of uniform and historical forecasts to serve as references, but do not repeat the comparison with standard time-series forecasting models, as it was already conducted in Chapter 7 for all severity items and aggregate scores. Predictions are generated in a forward chaining setting where the model is retrained every four days, and are evaluated with the logarithmic scoring rule (log predictive density, lpd).

8.3 Results

When fitting the models, we found no evidence for an absence of convergence by monitoring trace plots and checking the potential scale reduction factor \hat{R} . We conducted posterior predictive checks and found no clear discrepancies between the data and the models’ simulations.

8.3.1 Multivariate dynamic

By fitting the nine severity items of PO-SCORAD jointly, we can estimate how the severity items covary. We found that changes of severity items were positively correlated (Fig. 8.3B).

⁴Although subjective, this order is not arbitrary and corresponds to the iterative improvements of our Bayesian workflow, where we deemed the power prior to be the first improvement to consider, and including the trend the last one.

This implies more uncertainty in the prediction of PO-SCORAD, as changes accumulate rather than cancel out, compared to a situation with independent severity items. In particular, the changes in scratching, oozing, and redness appear strongly correlated, and itching is moderately correlated with all intensity signs. However, extent is only mildly correlated with subjective symptoms or intensity signs.

The uncertainty in the evolution of the latent dynamic was found to be strongly item-dependent, implying that some items are easier to predict than others (Fig. F.1). For instance, the evolution of oozing is more uncertain than the evolution of thickening. We also note from Fig. F.1 that the uncertainty of the measurement process is always bigger than the uncertainty of the latent dynamics, meaning that most of the prediction uncertainty can be explained by the uncertainty of the measurement process. This highlights the difficulty in extracting signal from the data. Similarly, no trend was detected for any of the severity items (Fig. F.2). This means that if the severity increases or decreases at a given time, there is no indication that it will move in the same direction in the near future.

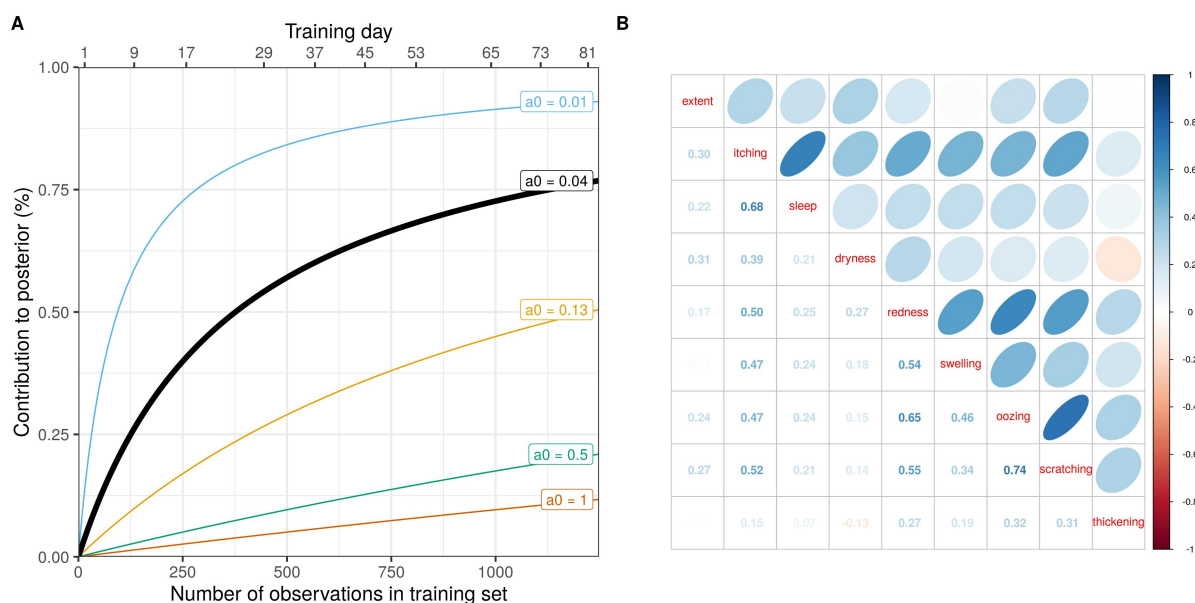


Figure 8.3: A) Approximate contribution of our dataset to the posterior distribution as a function of the number of observations in the training set, for different values of the power prior forgetting parameter a_0 . We use $a_0 = 0.04$ (in black), which is as if the model was pre-trained with $a_0 \times N_{\text{historical}} = 0.04 \times 9943 \approx 400$ observations. B) Visualisation of the expected correlation matrix of the changes between the nine latent severity items. The strength of the correlations is represented by an ellipse in the upper diagonal matrix and the lower diagonal matrix displays the expected correlation coefficients.

8.3.2 Calibration of PO-SCORAD with SCORAD

By integrating SCORAD measurements into the model, we can estimate the difference (bias) between patient assessments and clinical assessments. We found that the direction and amplitude of the biases were strongly item-dependent (Fig. 8.4A). For example, patients tend to overestimate extent and scratching, but underestimate dryness, redness, swelling and oozing, compared to clinicians, on average. The biases mostly stayed constant over time, except for scratching for which the bias is nearly 0 after the second measurement at week 4 (Fig. F.3). We would have expected the biases to decrease with time if patients were becoming better at assessing the severity of their symptoms. However, it is possible that the patients in this dataset were already familiar with PO-SCORAD assessments, or that learning does not happen if no feedback is given.

With the estimation of the measurement biases, we can convert PO-SCORAD predictions into SCORAD predictions (i.e. forecasting SCORAD), and infer SCORAD values if it had been measured daily (i.e. backcasting and nowcasting SCORAD, cf. Fig. 8.4B). For example, the severity trajectory in Fig. 8.4B demonstrates that the expected value of SCORAD would have been higher than the observed PO-SCORAD, for this patient. This is consistent with our estimates that clinicians tend to score intensity signs higher than patients (Fig. 8.4A), and the fact that intensity signs are the predominant component of SCORAD (cf. Section 2.1.3).

8.3.3 Treatment effects and recommendations

We estimated the parameters corresponding to treatment effects to be negative, confirming that treatment tends to improve severity (Fig. 8.5A). We found that topical corticosteroids were more effective than emollient cream, but estimates of treatment effect were small in absolute values. Treatment effects are also uncertain, as nearly half of patients always/never use treatment. In addition, treatment effects were highly heterogeneous across severity items. For example, all else being equal, a patient with severe scratching but no thickening of the skin would tend to respond more to corticosteroids than a patient with no scratching but severe thickening.

Using the model, we generated treatment recommendations for different decision profiles (Fig. 8.5B). We confirmed that a high perceived cost of treatment is associated with no treatment being recommended. Conversely, a null perceived cost of treatment is associated with both treatments being recommended, which was anticipated considering the expected treatment

⁵In practice, there are non-linearities between the latent and measurement spaces, even if they have approximately the same range. For example, a change of -13 in the latent space does not necessarily correspond to a change of -13 in the measurement space. If the latent score is 0, meaning that the corresponding measurement is likely 0 as well, a change of -13 in the latent space would only reduce the uncertainty that the measurement is 0.

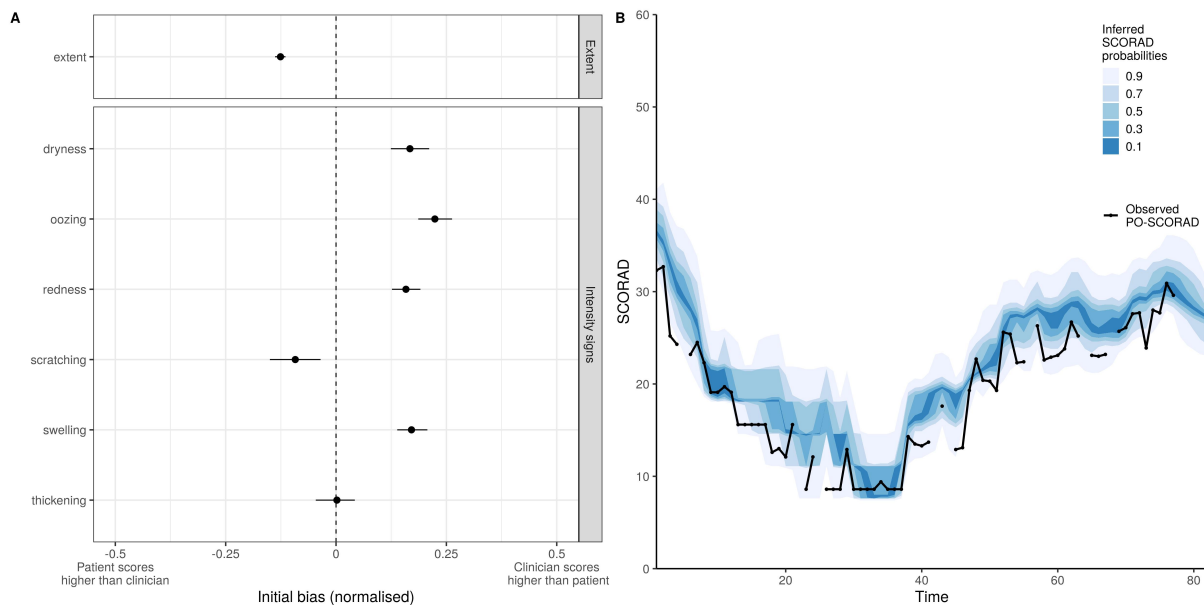


Figure 8.4: Calibration of PO-SCORAD measurements using SCORAD. A) Estimates of the initial bias (at day one) between the SCORAD and PO-SCORAD (mean and 90% CI), normalised by the range of the score. For example, the mean extent bias is -0.13 and extent takes values in $[0, 100]$, therefore patients are expected to overestimate by $0.13 \times 100 = 13$ the score given by the clinician⁵. B) PO-SCORAD trajectories and the corresponding estimates of SCORAD a posteriori (the distribution is represented by stacked credible intervals in shades of blue). For this representative patient, SCORAD (i.e. what a clinician would have measured) would be higher than PO-SCORAD (assessed by the patient), on average.

effects are all negative. For a “normal” perceived cost of treatment, if the patient is risk-averse (pessimistic), it is more likely that the algorithm recommends “using both treatments”, compared to when the patient is risk-seeking (optimistic), where the algorithm recommends “using no treatments” more. In any case, for a “normal” perceived cost of treatment, the most recommended action is using corticosteroids but not emollient. This is consistent with the result that corticosteroids are more effective than emollients (Fig. 8.5A): the additional benefit of using emollients may not be worth their “perceived” cost, in the “normal” cost scenario.

Recommendations can change depending on how much the model has learnt from the data, as illustrated by the fact that the algorithm recommends less of the “using both treatment” action and more of the “using corticosteroids and not emollient” action, as more data comes in (Fig. 8.6A). It is worth noting that recommendations can be “personalised”, even though treatment parameters are not patient-dependent. For example, recommendations can be different even for patients with the same SCORAD (Fig. 8.6B), probably because the same SCORAD can correspond to different values of the severity items, which are associated with different treatment responses. In addition, more treatment tends to be recommended when the severity is high, even though a severity dependence is not explicitly assumed in the utility function. This may be a side effect of more signs being present for severe AD, resulting in

more potential for improvement and better responses to treatment. The severity dependence of the recommendations could also be confounded by the fact that patients tend to have a higher severity at the beginning of the study, when the algorithm recommends “using both treatments” more often. These possible explanations illustrate the difficulty of interpreting the descriptive summaries of recommendations post-hoc. Nonetheless, every recommendation can be transparently explained by examining the utility function of the patient and their tolerance to risk.

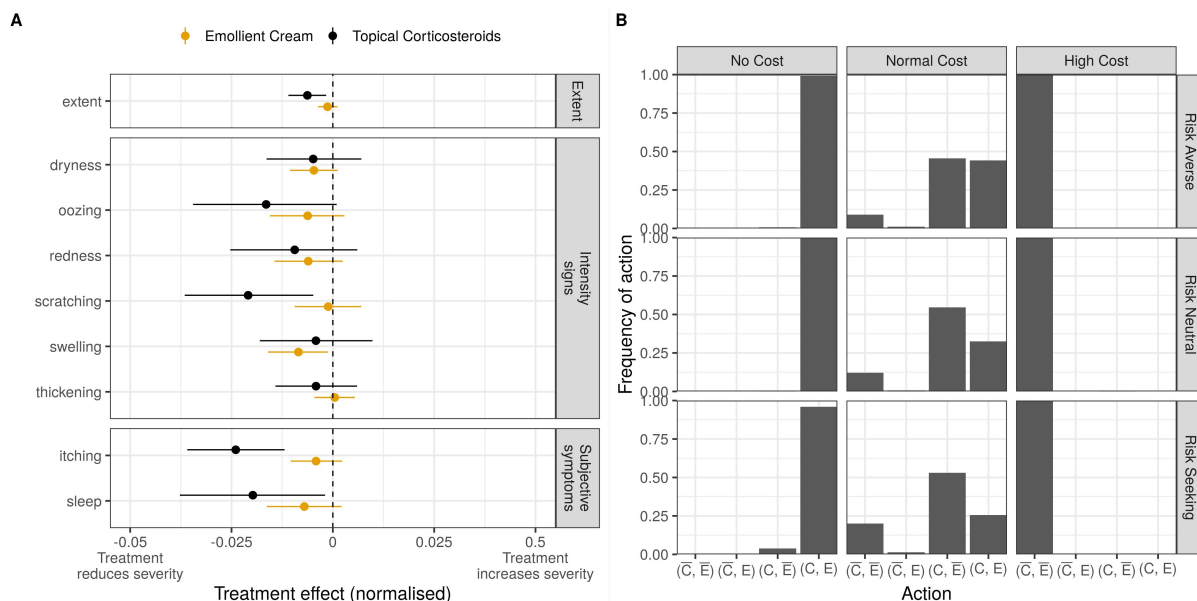


Figure 8.5: A) Average treatment effects on each severity item for topical corticosteroids and emollient cream, normalised by the range of the score (mean and 90% CI). For example, the expected effect of corticosteroids on the latent sleep loss item, defined in $[0, 10]$, is -0.02 , meaning that taking corticosteroids will on average decrease the sleep loss score by 0.20. B) Distribution of “next-day” recommended actions for different decision profiles corresponding to a no/normal/high “perceived” cost of treatment (vertical facets) and a risk-averse/neutral/seeking patient (horizontal facets). C and \bar{C} correspond to the action of using and not using corticosteroids, respectively; E and \bar{E} correspond to the action of using and not using emollients, respectively.

8.3.4 Model validation

We validated the model to assess whether its new features were associated with improvements in predictive performance (Fig. F.4). We did not find evidence that using extra information (power prior, SCORAD measurements, treatments) was associated with noticeable long-term improvements in predictive performance. This may not be surprising as the main contribution of the power prior is to accelerate the learning process; the addition of at most four SCORAD measurements is unlikely to change the performance of twelve-week-long daily time-series by much; and treatment effects were expected to be small (cf. Chapter 4). Similarly, modelling the

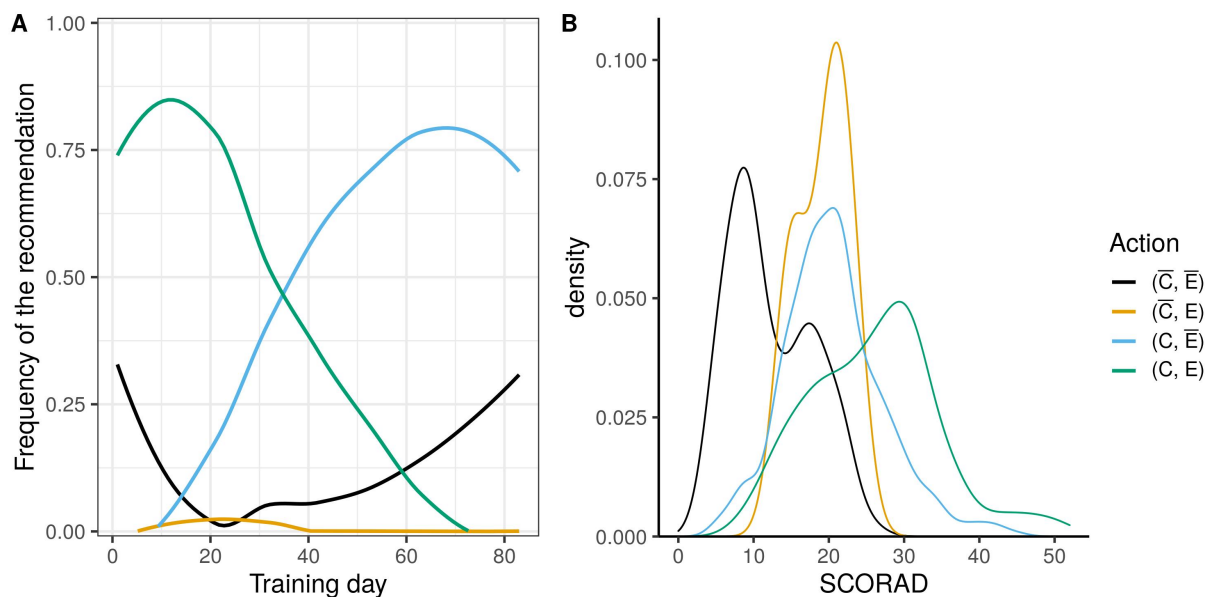


Figure 8.6: Analysis of treatment recommendations for a risk neutral patient and a “normal” perceived cost of treatments. C and \bar{C} correspond to the action of using and not using corticosteroids, respectively; E and \bar{E} correspond to the action of using and not using emollients, respectively. A) Frequency of recommended actions as a function of training day, smoothed by LOWESS. B) Distribution of SCORAD at the time of the recommendation, for each action.

trend in the latent dynamics did not improve performance, as the trend was estimated to be null. Modelling correlations did not change the predictive performance as measured by the lpd ⁶. The benefit of modelling correlations was also small for POEM prediction in Chapter 7.

8.4 Discussion

In this chapter, we used EczemaPred (Chapter 7) to predict the evolution of eczema severity while integrating complex multi-dimensional data (Fig. 8.2) to make inferences, predictions and treatment recommendations (Fig. 8.1). We leveraged existing knowledge about the dynamics of the disease by designing an informative prior derived from historical data, to accelerate model training (Fig. 8.3A). We showed that changes in eczema severity items are positively correlated (Fig. 8.3B), but found no evidence that eczema severity trajectories exhibit any short-term trend (Fig. F.2). We calibrated self-assessed PO-SCORAD using SCORAD measurements, assessed by clinical staff, to adjust for the patient biases in the measurement process (Fig. 8.4). Finally, we estimated the effects of using topical corticosteroids and emollient creams, and demonstrated a proof-of-concept for generating personalised treatment recommendations (Fig. 8.5).

⁶The lpd of individual predictions differs for the model with correlated severity items and the model with independent severity items, but when averaged over predictions, the lpd of the two models are similar.

One of the main insights from this study is the strong heterogeneity across severity items when estimating parameters of the latent dynamics, PO-SCORAD measurement biases and treatment effects. For example, we found that patients are expected to overestimate extent but underestimate redness, and that topical steroids are more effective to reduce scratching than the thickening of the skin (lichenification). As a result, the treatment effects measured with an aggregate severity score such as SCORAD may appear different between two patients with the same SCORAD but different clinical phenotypes (e.g. patients with different intensity of scratching and thickening), even if the treatment effects are not patient-dependent [18]. The fact that an effect may be confounded by the clinical phenotype of patients highlights the importance of modelling severity items rather than the aggregate scores.

By integrating multiple sources of information in a single model, we show how EczemaPred can provide a unified framework for predictions, inference and decision-making under uncertainty. The developed model can then be used to simultaneously answer multiple research questions in a consistent way. For example, our treatment recommendations are consistent with the inferred treatment effects and the model's predictions for PO-SCORAD and SCORAD, take into account the correlations between severity items, and benefits from the existing knowledge of prior studies, while quantifying uncertainty in measurements, parameters and predictions. Having a flexible model that can integrate all the available information is clinically relevant for medical decision-making, can provide more precise inference, reduce confounding, and build a more comprehensive understanding of the disease from imperfect data.

Our study is not without limitations. First, our proposal to generate treatment recommendations is very much a proof-of-concept. The estimated treatment effects are small and uncertain, so even if the recommendations were optimal and reliable, they would only be associated with a small improvements in severity. More importantly, the model is not causal, and this is why we did not attempt to evaluate the quality of the recommendations. Potentially many confounding factors are indeed missing, making counterfactual inference not possible. For example, better quality treatment data would be required, such as the daily usage, potency, and quantity of treatment applied. The suggested treatment recommendations are also illustrative, as the utility function and decision parameters would need to be adjusted to match patients' preferences⁷. Second, we did not detect any improvement of the predictive performance despite the additional complexity implemented in the model, highlighting the difficulty of predicting the evolution of eczema severity accurately. A larger cohort would also be required to investigate patient-dependence (in treatment effects, measurement biases, or even dynamical parameters) and ensure our results can be generalised.

⁷Patient's preferences could also change with time. For example, a risk-seeking behaviour (welcoming uncertainty) at the beginning of a study will encourage the exploration of new treatments. Then, the behaviour could be gradually changed, as more data comes in, to become risk-averse (penalising uncertainty), to encourage the "exploitation" of the acquired knowledge that some treatments are more effective than others.

A potentially interesting use of the model would be to reduce the dimensionality of the latent space to search for common patterns in the severity trajectories and cluster patients into different endotypes [10]. We can indeed hypothesise that the latent trajectories associated with different severity items may be somewhat redundant and the manifestation of a few, potentially independent mechanisms, that could stratify patients. This type of model-based clustering has notably been applied to investigate whether the “atopic march” hypothesis was supported by data [192].

Chapter 9

Conclusion

9.1 Summary

This thesis aimed to explore the feasibility of a data-driven personalised management of Atopic Dermatitis severity, including the automatic assessment of AD severity from camera images, the prediction of the short-term evolution of AD severity, and the generation of treatment recommendations. Our results have highlighted the challenges of this endeavour.

9.1.1 Collecting AD severity data

In Chapter 3, we developed a computer vision pipeline to automatically detect and assess eczema severity from camera images. While promising, our proof-of-concept only achieved a fair performance for predicting regional severity scores. In particular, our results have highlighted the marginal gain in predictive performance achieved by changing the model architecture and the limitations of our medium-size dataset of images of varying quality. More work would be required to enable reliable machine assessments of AD severity (see Section 9.2).

9.1.2 Predicting AD severity

In Chapters 4, 5, 6 and 7, we developed models and investigated requirements for predicting the short-term evolution of AD severity.

Our first approach (Chapter 4) was to focus our modelling efforts on the dynamic of AD

severity, by taking inspiration from an existing mathematical model that describes the biological mechanisms of AD [159]. However, we quickly realised that the severity measurements were noisy, and that it was difficult to achieve good prediction and extract signal from the data. In particular, additional features beyond severity measurements were not found useful to predict future AD severity, with only small and practically not significant effects of treatments (Chapter 4), environmental factors (Chapter 5) and biomarkers (Chapter 6) on predictions. While “negative”, our results illustrated and addressed common pitfalls in AD research in the context of personalised medicine, such as conflating predictions and associations, dichotomising continuous variables or failing to consider uncertainty in measurements and outcomes [18]. Our results suggested that it would be more effective to focus our modelling efforts on the measurement processes, as opposed to the dynamics of AD severity. This idea led to modelling and quantifying measurement errors explicitly using state-space models¹ (Chapters 6 and 7). For example, we achieved a good predictive performance with a more faithful representation of the measurement processes despite assuming a naive latent dynamic in the form of a random walk (simply assuming the future score is around the present score in Chapter 7). In particular, we found that the ordered logistic distribution was an efficient way of modelling the measurement of severity items (Chapters 5 and 7).

We also highlighted the importance of using high-quality measurements for prediction (Chapter 7), although we acknowledge a trade-off between the easiness of collecting the measurements and how much information they carry (notably to what extent they are predictable). On one hand, patient global assessments (e.g. bother score in Chapter 4) are usually easy and fast to collect for patients, but are subjective and do not differentiate the multiple symptoms or signs of AD². On the other hand, severity scores such as SCORAD (used in Chapters 6 and 7) have been validated and are based on multiple severity items, but are more time-consuming to record (even with the help of computer vision algorithms). Other scores such as EASI or SASSAD potentially carry even more information than SCORAD as they assess the severity of AD on different body regions rather than on a representative site, and could therefore be useful to integrate spatial information into predictive models, at the cost of an even longer collection time.

We also demonstrated the benefits of working with individual severity items as opposed to aggregate severity scores (Chapters 5 and 7). Severity scores such as SCORAD, EASI, or POEM were originally designed as summary tools for clinicians and patients. However, aggregating multiple severity items in a single score is inefficient for fitting models due to information loss. This has implications beyond predictive modelling, for example to optimise mathematical

¹Measurement error in Chapter 4 was limited to the rounding process, to model how the discrete score was generated.

²Although the AD symptom state in Chapter 5 can be seen as a global assessment, it is derived from multiple severity items and is therefore not as fast to collect as a score like bother.

models using aggregate severity score data. In particular, it would be challenging to fit SCORAD to the previously developed mathematical model of AD pathogenesis [159], because SCORAD aggregates severity items that relate to potentially different aspects of the model. For example, the extent of AD may be seen as a proxy for the skin barrier integrity, whereas intensity signs would correspond to the accumulation of flares, and subjective symptoms may not be well represented in any state variables.

Some scores may thus be more appropriate for developing models than others. In Chapter 7, we found that objective scores that measure physical signs of eczema (SCORAD) led to models with better predictive performance, compared to subjective scores that measure symptoms as experienced by patients (POEM).

9.1.3 Generating treatment recommendations

In Chapter 8, we made the model developed in Chapter 7 more relevant to clinical practice, and proposed a proof-of-concept for generating personalised treatment recommendations. Specifically, we demonstrated how to integrate information from previous studies to pre-train the model, therefore avoiding the cold-start problem. We also showed how to use clinician measurements to calibrate patients self-assessments, thus building more trust in the model predictions. We estimated treatment effects and demonstrated a method to generate personalised treatment recommendations. However, evaluating the validity of these recommendations is challenging, as it requires causal considerations (cf. Section 1.3.2). Even if the treatment recommendations were proven to be valid, they would result in marginal improvements of AD severity considering the small treatment effect size.

We found that treatment effects could be confounded by the clinical phenotype of patients, i.e. different patients with the same aggregate severity but with different physical signs could exhibit different treatment responses. This further highlights the importance of severity measurements in data-driven AD research, and that using aggregate scores could result in biased inferences. Choosing an appropriate score is also important when designing treatment recommendations, as different patients may have different views on which signs/symptoms are the most bothersome to them (e.g. some patients may prefer to reduce itch and others redness).

9.2 Future directions

We can identify several areas for future work, which we broadly classify into research and engineering. In this dichotomy, research work would aim to better understand the disease

dynamics. Conversely, engineering work would aim to complete the data-driven personalised pipeline for managing AD severity, with which patients would take photos of their eczema, receive the likely evolution of their condition, as well as recommendations on the best course of actions to take. In the longer term, the pipeline will also have to be validated in a real clinical setting to demonstrate its clinical utility.

9.2.1 Research

We believe there is little gain to be made in the predictive performance of our models describing the evolution of AD severity. We have indeed shown throughout this thesis that making good predictions of future AD severity is challenging, and that a priori important factors such as treatments, environmental factors or biomarkers do not have practically significant contributions to the predictive performance. As a result, we also believe the potential for making treatment recommendations is currently limited.

However, we do not claim that environmental factors, treatments, or biomarkers are irrelevant for AD research. It would be valuable to better understand their role in the disease dynamics and pathogenesis, even if they are not predictive of AD severity. We thus believe it is more promising to focus on inference, including causal inference, rather than prediction or recommendations³.

For example, it would be interesting to formally investigate the causal relationship between environmental factors and AD severity. While causal treatment effects can be estimated in randomised controlled trials, where patients are randomised into different treatment groups, it is more difficult to organise similar interventions for studying the causal effects of environmental factors, as it would be unethical to deliberately expose some patients to high levels of pollution, for example. Causal effects of environmental factors on AD severity must then be studied using observational data. This would require the explicit and transparent specification of a causal diagram, for example in the form of a directed acyclic graph (DAG), beyond relying on statistical significance (or predictive criteria) and a convincing causal story [193]. The causal diagram can be used to identify which variables need to be controlled for and which ones should not [194] [195], and serves as a guide for data collection. Once the confounding variables are identified and measured, flexible techniques such as Bayesian additive regression trees (BART) could be used to estimate causal effects, without the need to specify how these variables are parametrically related [196]. The causal diagram can also serve as a basis for a generative model to conduct full Bayesian inference [197], even in the presence of missing

³We may reassess the potential for personalised recommendations, if we can reliably estimate causal effects and find a disagreement with the estimates from the predictive models.

confounders, where missing variables are described by probability distributions (e.g. flares in Chapter 4) or inferred using proxies (e.g. daily quantities of treatment from the total reported quantities in Chapter 4) [193].

Studying the dynamics of AD severity could also be useful for patient stratification, i.e. identifying subgroups of patients with similar disease characteristics. We are not convinced by the idea that biomarkers measured at a single timepoint could help uncover subgroups of patients with different outcomes (endotypes, investigated in Chapter 6), nor that they are causally related to changes in severity (cf. Section 1.3.2). Instead, we hypothesise that severity trajectories are more likely to identify patient subgroups. For example, the evolution of severity items may exhibit similar patterns across patients, which could be a manifestation of different mechanisms of AD. Unsupervised learning, including clustering analysis, could help uncover such patterns using time-series of the short-term evolution of AD severity, similarly to what has been done to uncover developmental profiles of eczema [192] [198]. The models presented in this thesis could serve as a basis for model-based clustering. Other approaches for clustering AD severity trajectories could also be explored, for example using Gaussian Processes that require fewer assumptions about the data-generating mechanisms [199] [200].

Better understanding the long-term dynamics of AD could be another promising area of research, considering that we have only studied the evolution of AD severity over a few months, in this thesis. For example, dynamical parameters are likely to change on a long timescale, and may reflect changes in the disease phenotype. Studying these long-term dynamic changes would require the analysis of long time-series (e.g. daily to weekly measurements over a few years), which may be challenging to collect. However, the length of the time-series may balance the need to collect data from many patients. In particular, valuable preliminary insights could potentially be found by analysing the severity trajectory of even a single patient. This could be achieved using change-point detection methods [201], regime-switching models [202], or dynamical model averaging [203].

9.2.2 Engineering

We believe the main limitation for developing accurate algorithms to automatically assess AD severity from camera images does not lie in the architecture of the neural network, but in the quality and quantity of available data [158]. In the current absence of publicly available labelled images of AD severity, our efforts should thus be directed towards organising data collection and promoting data sharing. In particular, it is desirable to collect images of eczema for diverse skin tones, as the presentation of disease signs can be different in dark compared to light skin, for example.

In parallel, it may be interesting to investigate at what frequency patients should collect severity measurements to ensure good predictive performance, while minimising the inconvenience of collecting such data. We can imagine that patients would first record their severity regularly (e.g. every day) for a short period (e.g. a few weeks) to tune the patient-dependent parameters of the model, and then collect measurements only when the time to the last measurement becomes too long to ensure a minimal predictive accuracy. Investigating the frequency of data collection would ideally include elements of experimental design, to maximise the learning rate and the accuracy of the predictive model, as well as choice modelling to understand the patients' perceived trade-off between predictive performance and the inconvenience of data collection.

Finally, a computationally efficient implementation of the predictive models would be required for their large-scale deployment. Such implementation may involve simplifying the models, making approximations, deriving analytical solutions to speed up the computation (cf. Kalman filter), or using faster inference methods for time-series models and online training settings, such as Sequential Monte-Carlo (SMC, aka particle filters), variational inference or expectation-maximisation algorithms.

Our vision for a data-driven personalised management of AD severity offers many opportunities and challenges, and there is still a long way to go before it becomes a reality in clinical practice. By laying the first stones towards this goal, we hope this thesis will inspire future work in this area.

References

- [1] S. Weidinger and N. Novak, “Atopic dermatitis,” *The Lancet*, vol. 387, no. 10023, pp. 1109–1122, 2016, ISSN: 1474547X. DOI: [10.1016/S0140-6736\(15\)00149-X](https://doi.org/10.1016/S0140-6736(15)00149-X).
- [2] S. Nutten, “Atopic Dermatitis: Global Epidemiology and Risk Factors,” *Annals of Nutrition and Metabolism*, vol. 66, no. Suppl. 1, pp. 8–16, May 2015, ISSN: 0250-6807. DOI: [10.1159/000370220](https://doi.org/10.1159/000370220).
- [3] A. M. Drucker, A. R. Wang, W. Q. Li, E. Sevetson, J. K. Block, and A. A. Qureshi, “The Burden of Atopic Dermatitis: Summary of a Report for the National Eczema Association,” *Journal of Investigative Dermatology*, vol. 137, no. 1, pp. 26–30, Jan. 2017, ISSN: 15231747. DOI: [10.1016/j.jid.2016.07.012](https://doi.org/10.1016/j.jid.2016.07.012).
- [4] J. I. Silverberg, J. P. Thyssen, A. S. Paller, *et al.*, “What’s in a name? Atopic dermatitis or atopic eczema, but not eczema alone,” *en, Allergy: European Journal of Allergy and Clinical Immunology*, vol. 72, no. 12, pp. 2026–2030, Dec. 2017, ISSN: 13989995. DOI: [10.1111/all.13225](https://doi.org/10.1111/all.13225).
- [5] T. Bieber, “Why we need a harmonized name for atopic dermatitis/atopic eczema/eczema!” *Allergy: European Journal of Allergy and Clinical Immunology*, vol. 71, no. 10, pp. 1379–1380, 2016, ISSN: 13989995. DOI: [10.1111/all.12984](https://doi.org/10.1111/all.12984).
- [6] C. Frainay, Y. Pitarch, S. Filippi, M. Evangelou, and A. Custovic, “Atopic dermatitis or eczema? Consequences of ambiguity in disease name for biomedical literature mining,” *Clinical and Experimental Allergy*, vol. 51, no. 9, pp. 1185–1194, Sep. 2021, ISSN: 13652222. DOI: [10.1111/cea.13981](https://doi.org/10.1111/cea.13981).
- [7] S. G. Johansson, T. Bieber, R. Dahl, *et al.*, “Revised nomenclature for allergy for global use: Report of the Nomenclature Review Committee of the World Allergy Organization, October 2003,” *Journal of Allergy and Clinical Immunology*, vol. 113, no. 5, pp. 832–836, May 2004, ISSN: 00916749. DOI: [10.1016/j.jaci.2003.12.591](https://doi.org/10.1016/j.jaci.2003.12.591).
- [8] T. Zuberbier, S. J. Orlow, A. S. Paller, *et al.*, “Patient perspectives on the management of atopic dermatitis,” *Journal of Allergy and Clinical Immunology*, vol. 118, no. 1, pp. 226–232, Jul. 2006, ISSN: 00916749. DOI: [10.1016/j.jaci.2006.02.031](https://doi.org/10.1016/j.jaci.2006.02.031).

- [9] J. Krejci-Manwaring, M. G. Tusa, C. Carroll, *et al.*, “Stealth monitoring of adherence to topical medication: adherence is very poor in children with atopic dermatitis,” *Journal of the American Academy of Dermatology*, vol. 56, no. 2, pp. 211–6, Feb. 2007, ISSN: 1097-6787. DOI: [10.1016/j.jaad.2006.05.073](https://doi.org/10.1016/j.jaad.2006.05.073).
- [10] T. Bieber, A. M. D’Erme, C. A. Akdis, *et al.*, “Clinical phenotypes and endophenotypes of atopic dermatitis: Where are we, and where should we go?” eng, *Journal of Allergy and Clinical Immunology*, vol. 139, no. 4, S58–S64, Apr. 2017, ISSN: 10976825. DOI: [10.1016/j.jaci.2017.01.008](https://doi.org/10.1016/j.jaci.2017.01.008).
- [11] S. J. Galli, “Toward precision medicine and health: Opportunities and challenges in allergic diseases,” *Journal of Allergy and Clinical Immunology*, vol. 137, no. 5, pp. 1289–1300, May 2016, ISSN: 00916749. DOI: [10.1016/j.jaci.2016.03.006](https://doi.org/10.1016/j.jaci.2016.03.006).
- [12] L. S. van Galen, X. Xu, M. J. Koh, S. Thng, and J. Car, “Eczema apps conformance with clinical guidelines: a systematic assessment of functions, tools and content,” *British Journal of Dermatology*, vol. 182, no. 2, pp. 444–453, Feb. 2020, ISSN: 13652133. DOI: [10.1111/bjd.18152](https://doi.org/10.1111/bjd.18152).
- [13] Y. Bengio, Y. Lecun, and G. Hinton, “Deep learning for AI,” *Communications of the ACM*, vol. 64, no. 7, pp. 58–65, Jul. 2021, ISSN: 0001-0782. DOI: [10.1145/3448250](https://doi.org/10.1145/3448250).
- [14] G. D. Magoulas and A. Prentza, “Machine Learning in Medical Applications,” in *Machine Learning and Its Applications, Advanced Lectures*, London, UK, UK: Springer-Verlag, 2001, pp. 300–307, ISBN: 978-3-540-42490-1. DOI: [10.1007/3-540-44673-7{_}19](https://doi.org/10.1007/3-540-44673-7_{_}19).
- [15] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, *et al.*, “Opportunities and obstacles for deep learning in biology and medicine,” *Journal of The Royal Society Interface*, vol. 15, no. 141, p. 20 170 387, Apr. 2018, ISSN: 1742-5689. DOI: [10.1098/rsif.2017.0387](https://doi.org/10.1098/rsif.2017.0387).
- [16] K. Eyerich, S. J. Brown, B. E. Perez White, *et al.*, “Human and computational models of atopic dermatitis: A review and perspectives by an expert panel of the International Eczema Council.,” *The Journal of allergy and clinical immunology*, vol. 143, no. 1, pp. 36–45, Jan. 2019, ISSN: 1097-6825. DOI: [10.1016/j.jaci.2018.10.033](https://doi.org/10.1016/j.jaci.2018.10.033).
- [17] J. Schmitt, M. Meurer, U. Schwanebeck, X. Grählert, and K. Schäkel, “Treatment following an evidence-based algorithm versus individualised symptom-oriented treatment for atopic eczema: A randomised controlled trial,” *Dermatology*, vol. 217, no. 4, pp. 299–308, 2008, ISSN: 10188665. DOI: [10.1159/000151355](https://doi.org/10.1159/000151355).
- [18] S. Senn, “Statistical pitfalls of personalized medicine,” *Nature*, vol. 563, no. 7733, pp. 619–621, Nov. 2018, ISSN: 0028-0836. DOI: [10.1038/d41586-018-07535-2](https://doi.org/10.1038/d41586-018-07535-2).
- [19] D. Bzdok, G. Varoquaux, and E. W. Steyerberg, *Prediction, Not Association, Paves the Road to Precision Medicine*, Feb. 2021. DOI: [10.1001/jamapsychiatry.2020.2549](https://doi.org/10.1001/jamapsychiatry.2020.2549).

- [20] G. Shmueli, “To Explain or to Predict?” *Statistical Science*, vol. 25, no. 3, pp. 289–310, Aug. 2010, ISSN: 0883-4237. DOI: [10.1214/10-STS330](https://doi.org/10.1214/10-STS330).
- [21] S. Wongvibulsin, B. K.-T. Ho, and S. G. Kwatra, “Embracing machine learning and digital health technology for precision dermatology,” *Journal of Dermatological Treatment*, pp. 1–2, Jun. 2019, ISSN: 0954-6634. DOI: [10.1080/09546634.2019.1623373](https://doi.org/10.1080/09546634.2019.1623373).
- [22] J. C. Wyatt and D. G. Altman, “Commentary: Prognostic models: Clinically useful or quickly forgotten?” *BMJ*, vol. 311, no. 7019, p. 1539, Dec. 1995, ISSN: 14685833. DOI: [10.1136/bmj.311.7019.1539](https://doi.org/10.1136/bmj.311.7019.1539).
- [23] R. Szeliski, *Computer Vision*, ser. Texts in Computer Science. London: Springer London, 2011, ISBN: 978-1-84882-934-3. DOI: [10.1007/978-1-84882-935-0](https://doi.org/10.1007/978-1-84882-935-0).
- [24] R. S. Sutton and A. G. Barto, *Reinforcement Learning: an introduction - second edition*. The MIT Press, 2018, ISBN: 9780262039246.
- [25] Q. Huang, “Model-Based or Model-Free, a Review of Approaches in Reinforcement Learning,” in *2020 International Conference on Computing and Data Science (CDS)*, IEEE, Aug. 2020, pp. 219–221, ISBN: 978-1-7281-7106-7. DOI: [10.1109/CDS49703.2020.00051](https://doi.org/10.1109/CDS49703.2020.00051).
- [26] J. M. Bernardo and A. F. Smith, *Bayesian Theory*, J. M. Bernardo and A. F. M. Smith, Eds., ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., May 1994, pp. 1–595, ISBN: 9780470316870. DOI: [10.1002/9780470316870](https://doi.org/10.1002/9780470316870).
- [27] T. Le and B. Clarke, “A bayes interpretation of stacking for m-complete and m-open settings,” *Bayesian Analysis*, vol. 12, no. 3, pp. 807–829, 2017, ISSN: 19316690. DOI: [10.1214/16-BA1023](https://doi.org/10.1214/16-BA1023).
- [28] B. Clarke, J. Clarke, and C. W. Yu, “Statistical Problem Classes and Their Links to Information Theory,” *Econometric Reviews*, vol. 33, no. 1-4, pp. 337–371, Feb. 2014, ISSN: 07474938. DOI: [10.1080/07474938.2013.807190](https://doi.org/10.1080/07474938.2013.807190).
- [29] J. Pearl, *Causality: Models, reasoning, and inference, second edition*. Cambridge University Press, Jan. 2011, pp. 1–464, ISBN: 9780511803161. DOI: [10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161).
- [30] B. Scholkopf, F. Locatello, S. Bauer, *et al.*, “Toward Causal Representation Learning,” *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, May 2021, ISSN: 0018-9219. DOI: [10.1109/JPROC.2021.3058954](https://doi.org/10.1109/JPROC.2021.3058954).
- [31] S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems,” May 2020.
- [32] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, Feb. 2019, ISSN: 15577317. DOI: [10.1145/3241036](https://doi.org/10.1145/3241036).

- [33] C. Fernández-Loría and F. Provost, “Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters,” *INFORMS Journal on Data Science*, Mar. 2022, ISSN: 2694-4022. DOI: [10.1287/ijds.2021.0006](https://doi.org/10.1287/ijds.2021.0006).
- [34] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ”right to explanation”,” *AI Magazine*, vol. 38, no. 3, p. 50, 2017, ISSN: 0738-4602. DOI: [10.1609/aimag.v38i3.2741](https://doi.org/10.1609/aimag.v38i3.2741).
- [35] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, ISSN: 2522-5839. DOI: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x).
- [36] Z. C. Lipton, “The Mythos of Model Interpretability,” *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, ISSN: 1542-7730. DOI: [10.1145/3236386.3241340](https://doi.org/10.1145/3236386.3241340).
- [37] C. Molnar, G. König, J. Herbinger, *et al.*, “General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models,” in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Jul. 2022, pp. 39–68. DOI: [10.1007/978-3-031-04083-2{-}4](https://doi.org/10.1007/978-3-031-04083-2{-}4).
- [38] J. Wilkinson, K. F. Arnold, E. J. Murray, *et al.*, “Time to reality check the promises of machine learning-powered precision medicine,” *The Lancet Digital Health*, vol. 2, no. 12, e677–e680, Dec. 2020, ISSN: 25897500. DOI: [10.1016/S2589-7500\(20\)30200-4](https://doi.org/10.1016/S2589-7500(20)30200-4).
- [39] F. E. Harrell, *Classification vs. Prediction*, 2017.
- [40] A. Marthe, V. D. Bles, S. V. D. Linden, and A. Lj, “The effects of communicating uncertainty on public trust in facts and numbers,” *Proceedings of the National Academy of Sciences*, pp. 1–43, Mar. 2020, ISSN: 0027-8424. DOI: [10.1073/pnas.1913678117](https://doi.org/10.1073/pnas.1913678117).
- [41] S. L. Franconeri, L. M. Padilla, P. Shah, J. M. Zacks, and J. Hullman, “The Science of Visual Data Communication: What Works,” *Psychological Science in the Public Interest*, vol. 22, no. 3, pp. 110–161, Dec. 2021, ISSN: 1529-1006. DOI: [10.1177/15291006211051956](https://doi.org/10.1177/15291006211051956).
- [42] R. J. Tanaka and M. Ono, “Skin disease modeling from a mathematical perspective,” *The Journal of investigative dermatology*, vol. 133, no. 6, pp. 1472–8, Jun. 2013, ISSN: 1523-1747. DOI: [10.1038/jid.2013.69](https://doi.org/10.1038/jid.2013.69).
- [43] R. E. Baker, J. M. Peña, J. Jayamohan, and A. Jérusalem, “Mechanistic models versus machine learning, a fight worth fighting for the biological community?” *Biology Letters*, vol. 14, no. 5, 2018, ISSN: 1744957X. DOI: [10.1098/rsbl.2017.0660](https://doi.org/10.1098/rsbl.2017.0660).
- [44] A. Saltelli, “A short comment on statistical versus mathematical modelling,” *Nature Communications*, vol. 10, no. 1, p. 3870, Dec. 2019, ISSN: 2041-1723. DOI: [10.1038/s41467-019-11865-8](https://doi.org/10.1038/s41467-019-11865-8).
- [45] L. Breiman, “Statistical Modeling: The Two Cultures,” *Statistical Science*, vol. 16, no. 3, pp. 199–231, Aug. 2002, ISSN: 0883-4237. DOI: [10.1214/ss/1009213726](https://doi.org/10.1214/ss/1009213726).

- [46] F. E. Harrell, *Road Map for Choosing Between Statistical Modeling and Machine Learning*, 2018.
- [47] S. Greenland, “The Causal Foundations of Applied Probability and Statistics,” in *Probabilistic and Causal Inference*, New York, NY, USA: ACM, Feb. 2022, pp. 605–624. DOI: [10.1145/3501714.3501747](https://doi.org/10.1145/3501714.3501747).
- [48] B. Efron, “Prediction, Estimation, and Attribution,” *Journal of the American Statistical Association*, vol. 115, no. 530, pp. 636–655, Apr. 2020, ISSN: 1537274X. DOI: [10.1080/01621459.2020.1762613](https://doi.org/10.1080/01621459.2020.1762613).
- [49] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, Apr. 2005, ISSN: 1369-7412. DOI: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x).
- [50] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 110, pp. 12–22, Jun. 2019, ISSN: 08954356. DOI: [10.1016/j.jclinepi.2019.02.004](https://doi.org/10.1016/j.jclinepi.2019.02.004).
- [51] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “Statistical and Machine Learning forecasting methods: Concerns and ways forward,” *PLoS ONE*, vol. 13, no. 3, e0194889, 2018, ISSN: 19326203. DOI: [10.1371/journal.pone.0194889](https://doi.org/10.1371/journal.pone.0194889).
- [52] E. Begoli, T. Bhattacharya, and D. Kusnezov, “The need for uncertainty quantification in machine-assisted medical decision making,” *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, Jan. 2019, ISSN: 2522-5839. DOI: [10.1038/s42256-018-0004-1](https://doi.org/10.1038/s42256-018-0004-1).
- [53] D. J. Hand, “Classifier technology and the illusion of progress,” *Statistical Science*, vol. 21, no. 1, pp. 1–14, Feb. 2006, ISSN: 08834237. DOI: [10.1214/088342306000000060](https://doi.org/10.1214/088342306000000060).
- [54] A. L. Beam, A. K. Manrai, and M. Ghassemi, “Challenges to the Reproducibility of Machine Learning Models in Health Care,” *JAMA*, vol. 323, no. 4, p. 305, Jan. 2020, ISSN: 0098-7484. DOI: [10.1001/jama.2019.20866](https://doi.org/10.1001/jama.2019.20866).
- [55] C. M. Bishop, “Model-based machine learning,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1984, p. 20 120 222, Feb. 2013, ISSN: 1364-503X. DOI: [10.1098/rsta.2012.0222](https://doi.org/10.1098/rsta.2012.0222).
- [56] B. Carpenter, A. Gelman, M. D. Hoffman, *et al.*, “Stan : A Probabilistic Programming Language,” *Journal of Statistical Software*, vol. 76, no. 1, pp. 1–32, 2017, ISSN: 1548-7660. DOI: [10.18637/jss.v076.i01](https://doi.org/10.18637/jss.v076.i01).
- [57] A. Gelman, A. Vehtari, D. Simpson, *et al.*, “Bayesian Workflow,” Nov. 2020.
- [58] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, ser. Springer Series in Statistics. New York, NY: Springer New York, 1985, ISBN: 978-1-4419-3074-3. DOI: [10.1007/978-1-4757-4286-2](https://doi.org/10.1007/978-1-4757-4286-2).

- [59] K. Pan, G. Hurault, K. Arulkumaran, H. C. Williams, and R. J. Tanaka, “EczemaNet: Automating Detection and Severity Assessment of Atopic Dermatitis,” in *Machine Learning in Medical Imaging*, Springer, Cham, Oct. 2020, pp. 220–230. DOI: [10.1007/978-3-030-59861-7_{-}23](https://doi.org/10.1007/978-3-030-59861-7_{-}23).
- [60] G. Hurault, E. Domínguez-Hüttinger, S. M. Langan, H. C. Williams, and R. J. Tanaka, “Personalized prediction of daily eczema severity scores using a mechanistic machine learning model,” *Clinical and Experimental Allergy*, vol. 50, no. 11, pp. 1258–1266, Aug. 2020, ISSN: 13652222. DOI: [10.1111/cea.13717](https://doi.org/10.1111/cea.13717).
- [61] G. Hurault, V. Delorieux, Y.-M. Kim, K. Ahn, H. C. Williams, and R. J. Tanaka, *Impact of environmental factors in predicting daily severity scores of atopic dermatitis*, Apr. 2021. DOI: [10.1002/clt2.12019](https://doi.org/10.1002/clt2.12019).
- [62] G. Hurault, E. Roekevisch, M. E. Schram, *et al.*, “Can serum biomarkers predict the outcome of systemic immunosuppressive therapy in adult atopic dermatitis patients?” *Skin Health and Disease*, vol. 2, no. 1, e77, Jan. 2022, ISSN: 2690-442X. DOI: [10.1002/ski2.77](https://doi.org/10.1002/ski2.77).
- [63] G. Hurault, J. F. Stalder, S. Mery, *et al.*, “EczemaPred: A computational framework for personalised prediction of eczema severity dynamics,” *Clinical and Translational Allergy*, vol. 12, no. 3, e12140, Mar. 2022, ISSN: 2045-7022. DOI: [10.1002/clt2.12140](https://doi.org/10.1002/clt2.12140).
- [64] T. Bieber, “Atopic Dermatitis,” *New England Journal of Medicine*, vol. 358, no. 14, pp. 1483–1494, Apr. 2008, ISSN: 0028-4793. DOI: [10.1056/NEJMra074081](https://doi.org/10.1056/NEJMra074081).
- [65] S. Weidinger, L. A. Beck, T. Bieber, K. Kabashima, and A. D. Irvine, “Atopic dermatitis,” *Nature Reviews Disease Primers*, vol. 4, no. 1, p. 1, Dec. 2018, ISSN: 2056-676X. DOI: [10.1038/s41572-018-0001-z](https://doi.org/10.1038/s41572-018-0001-z).
- [66] J. Hanifin and G. Rajka, “Diagnostic features of atopic eczema.,” *Acta Dermatol Venereol (Stockh)*, vol. 92, no. 92, pp. 44–47, 1980, ISSN: 0926-9959. DOI: [10.1111/j.1468-3083.2006.01664.x](https://doi.org/10.1111/j.1468-3083.2006.01664.x).
- [67] H. C. Williams, P. G. J. Burney, A. C. Pembrokef, R. J. Hay, and P. Burney, “The U.K. Working Party’s Diagnostic Criteria for Atopic Dermatitis. III. Independent hospital validation Composition of the Atopic Dermatitis Diagnostic Criteria Working Party,” *British Journal of Dermatology*, vol. 131, pp. 406–416, 1994.
- [68] M. Sullivan and N. B. Silverberg, “Current and emerging concepts in atopic dermatitis pathogenesis,” *Clinics in Dermatology*, vol. 35, no. 4, pp. 349–353, Jul. 2017, ISSN: 18791131. DOI: [10.1016/j.clindermatol.2017.03.006](https://doi.org/10.1016/j.clindermatol.2017.03.006).
- [69] S. M. Langan, A. D. Irvine, and S. Weidinger, “Atopic dermatitis,” *The Lancet*, vol. 396, no. 10247, pp. 345–360, Aug. 2020, ISSN: 1474547X. DOI: [10.1016/S0140-6736\(20\)31286-1](https://doi.org/10.1016/S0140-6736(20)31286-1).

- [70] C. Flohr, K. England, S. Radulovic, *et al.*, “Filaggrin loss-of-function mutations are associated with early-onset eczema, eczema severity and transepidermal water loss at 3 months of age,” *British Journal of Dermatology*, vol. 163, no. 6, pp. 1333–1336, Dec. 2010. DOI: [10.1111/j.1365-2133.2010.10068.x](https://doi.org/10.1111/j.1365-2133.2010.10068.x).
- [71] A. J. Hendricks, L. F. Eichenfield, and V. Y. Shi, “The impact of airborne pollution on atopic dermatitis: a literature review,” *British Journal of Dermatology*, vol. 183, no. 1, pp. 16–23, Jul. 2020, ISSN: 13652133. DOI: [10.1111/bjd.18781](https://doi.org/10.1111/bjd.18781).
- [72] L. M. Buys, “Treatment options for atopic dermatitis.,” *American family physician*, vol. 75, no. 4, pp. 523–8, Feb. 2007, ISSN: 0002-838X.
- [73] E. L. Simpson, “Current Medical Research and Opinion Atopic dermatitis: a review of topical treatment options Brief review Atopic dermatitis: a review of topical treatment options,” *Current Medical Research & Opinion Curr Med Res Opin*, vol. 26, no. 26, 2010. DOI: [10.1185/03007990903512156](https://doi.org/10.1185/03007990903512156).
- [74] E. L. Simpson, M. Bruin-Weller, C. Flohr, *et al.*, “When does atopic dermatitis warrant systemic therapy? Recommendations from an expert panel of the International Eczema Council,” *Journal of the American Academy of Dermatology*, vol. 77, no. 4, pp. 623–633, Oct. 2017, ISSN: 10976787. DOI: [10.1016/j.jaad.2017.06.042](https://doi.org/10.1016/j.jaad.2017.06.042).
- [75] A. Wollenberg and L. M. Ehmman, “Long term treatment concepts and proactive therapy for atopic eczema,” *Annals of Dermatology*, vol. 24, no. 3, pp. 253–260, Aug. 2012, ISSN: 10139087. DOI: [10.5021/ad.2012.24.3.253](https://doi.org/10.5021/ad.2012.24.3.253).
- [76] A. W. Li, E. S. Yin, and R. J. Antaya, “Topical Corticosteroid Phobia in Atopic Dermatitis,” *JAMA Dermatology*, vol. 153, no. 10, p. 1036, Oct. 2017, ISSN: 2168-6068. DOI: [10.1001/jamadermatol.2017.2437](https://doi.org/10.1001/jamadermatol.2017.2437).
- [77] S. Tofte, M. Graeber, R. Cherill, M. Omoto, M. Thurston, and J. M. Hanifin, “Eczema area and severity index (EASI): A new tool to evaluate atopic dermatitis,” *Journal of the European Academy of Dermatology and Venereology*, Abstracts of the 7th Congress of the European Academy of Dermatology and Venereology, vol. 11, S197, Sep. 1998, ISSN: 0926-9959. DOI: [10.1016/S0926-9959\(98\)95291-6](https://doi.org/10.1016/S0926-9959(98)95291-6).
- [78] J. M. Hanifin, M. Thurston, M. Omoto, R. Cherill, S. J. Tofte, and M. Graeber, “The eczema area and severity index (EASI): Assessment of reliability in atopic dermatitis,” *Experimental Dermatology*, vol. 10, no. 1, pp. 11–18, Feb. 2001, ISSN: 09066705. DOI: [10.1034/j.1600-0625.2001.100102.x](https://doi.org/10.1034/j.1600-0625.2001.100102.x).
- [79] J. Schmitt, P. I. Spuls, K. S. Thomas, *et al.*, “The Harmonising Outcome Measures for Eczema (HOME) statement to assess clinical signs of atopic eczema in trials.,” *The Journal of allergy and clinical immunology*, vol. 134, no. 4, pp. 800–7, Oct. 2014, ISSN: 1097-6825. DOI: [10.1016/j.jaci.2014.07.043](https://doi.org/10.1016/j.jaci.2014.07.043).

- [80] J. F. Stalder, A. Taïeb, D. J. Atherton, *et al.*, “Severity scoring of atopic dermatitis: the SCORAD index. Consensus Report of the European Task Force on Atopic Dermatitis,” *eng, Dermatology (Basel, Switzerland)*, vol. 186, no. 1, pp. 23–31, 1993, ISSN: 1018-8665. DOI: [10.1159/000247298](https://doi.org/10.1159/000247298).
- [81] J. Schmitt, S. Langan, H. C. Williams, and European Dermato-Epidemiology Network, “What are the best outcome measurements for atopic eczema? A systematic review,” *The Journal of allergy and clinical immunology*, vol. 120, no. 6, pp. 1389–98, Dec. 2007, ISSN: 1097-6825. DOI: [10.1016/j.jaci.2007.08.011](https://doi.org/10.1016/j.jaci.2007.08.011).
- [82] J. Schmitt, S. Langan, S. Deckert, *et al.*, “Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation,” *The Journal of allergy and clinical immunology*, vol. 132, no. 6, pp. 1337–47, Dec. 2013, ISSN: 1097-6825. DOI: [10.1016/j.jaci.2013.07.008](https://doi.org/10.1016/j.jaci.2013.07.008).
- [83] J. Berth-Jones, “Six Area, Six Sign Atopic Dermatitis (SASSAD) severity score: A simple system for monitoring disease activity in atopic dermatitis,” *British Journal of Dermatology, Supplement*, vol. 135, no. 48, pp. 25–30, Sep. 1996, ISSN: 0366077X. DOI: [10.1111/j.1365-2133.1996.tb00706.x](https://doi.org/10.1111/j.1365-2133.1996.tb00706.x).
- [84] A. Wolkerstorfer, F. B. De Waard van der Spek, E. J. Glazenburg, P. G. Mulder, and A. P. Oranje, “Scoring the severity of atopic dermatitis: Three item severity score as a rough system for daily practice and as a pre-screening tool for studies,” *Acta Dermato-Venereologica*, vol. 79, no. 5, pp. 356–359, Aug. 1999, ISSN: 00015555. DOI: [10.1080/000155599750010256](https://doi.org/10.1080/000155599750010256).
- [85] C. R. Charman, A. J. Venn, and H. C. Williams, “The patient-oriented eczema measure: Development and initial validation of a new tool for measuring atopic eczema severity from the patients’ perspective,” *Archives of Dermatology*, vol. 140, no. 12, pp. 1513–1519, 2004, ISSN: 0003987X. DOI: [10.1001/archderm.140.12.1513](https://doi.org/10.1001/archderm.140.12.1513).
- [86] P. I. Spuls, L. A. Gerbens, E. Simpson, *et al.*, “Patient-Oriented Eczema Measure (POEM), a core instrument to measure symptoms in clinical trials: a Harmonising Outcome Measures for Eczema (HOME) statement,” *British Journal of Dermatology*, vol. 176, no. 4, pp. 979–984, 2017, ISSN: 13652133. DOI: [10.1111/bjd.15179](https://doi.org/10.1111/bjd.15179).
- [87] M. Vourc’h-Jourdain, S. Barbarot, A. Taïeb, *et al.*, “Patient-oriented SCORAD: A self-assessment score in atopic dermatitis - A preliminary feasibility study,” *Dermatology*, vol. 218, no. 3, pp. 246–251, Feb. 2009, ISSN: 10188665. DOI: [10.1159/000193997](https://doi.org/10.1159/000193997).
- [88] Y. A. Leshem, J. R. Chalmers, C. Apfelbacher, *et al.*, “Measuring atopic eczema symptoms in clinical practice: The first consensus statement from the Harmonising Outcome Measures for Eczema in clinical practice initiative,” *Journal of the American Academy of Dermatology*, vol. 82, no. 5, pp. 1181–1186, May 2020, ISSN: 10976787. DOI: [10.1016/j.jaad.2019.12.055](https://doi.org/10.1016/j.jaad.2019.12.055).

- [89] R. Chopra, P. P. Vakharia, R. Sacotte, *et al.*, “Severity strata for Eczema Area and Severity Index (EASI), modified EASI, Scoring Atopic Dermatitis (SCORAD), objective SCORAD, Atopic Dermatitis Severity Index and body surface area in adolescents and adults with atopic dermatitis,” *eng. British Journal of Dermatology*, vol. 177, no. 5, pp. 1316–1321, May 2017, ISSN: 13652133. DOI: [10.1111/bjd.15641](https://doi.org/10.1111/bjd.15641).
- [90] G. Hurault, M. Schram, E. Roekevisch, P. Spuls, and R. Tanaka, “Relationship and probabilistic stratification of Eczema Area and Severity Index and objective Scoring Atopic Dermatitis severity scores for atopic dermatitis,” *British Journal of Dermatology*, vol. 179, no. 4, pp. 1003–1005, Oct. 2018, ISSN: 13652133. DOI: [10.1111/bjd.16916](https://doi.org/10.1111/bjd.16916).
- [91] J. F. Stalder, S. Barbarot, A. Wollenberg, *et al.*, “Patient-Oriented SCORAD (PO-SCORAD): A new self-assessment scale in atopic dermatitis validated in Europe,” *Allergy: European Journal of Allergy and Clinical Immunology*, vol. 66, no. 8, pp. 1114–1121, Aug. 2011, ISSN: 01054538. DOI: [10.1111/j.1398-9995.2011.02577.x](https://doi.org/10.1111/j.1398-9995.2011.02577.x).
- [92] K. L. Gwet, *Handbook of Inter-Rater Reliability: the definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2010, 197 p. ISBN: 9780970806222.
- [93] C. R. Charman, A. J. Venn, and H. C. Williams, “Measurement of body surface area involvement in atopic eczema: An impossible task?” *British Journal of Dermatology*, vol. 140, no. 1, pp. 109–111, 1999, ISSN: 00070963. DOI: [10.1046/j.1365-2133.1999.02617.x](https://doi.org/10.1046/j.1365-2133.1999.02617.x).
- [94] H. C. de Vet and C. B. Terwee, “The minimal detectable change should not replace the minimal important difference,” *Journal of Clinical Epidemiology*, vol. 63, no. 7, pp. 804–805, 2010, ISSN: 08954356. DOI: [10.1016/j.jclinepi.2009.12.015](https://doi.org/10.1016/j.jclinepi.2009.12.015).
- [95] R. Jaeschke, J. Singer, and G. H. Guyatt, “Measurement of health status. Ascertaining the minimal clinically important difference,” *Controlled Clinical Trials*, vol. 10, no. 4, pp. 407–415, Dec. 1989, ISSN: 01972456. DOI: [10.1016/0197-2456\(89\)90005-6](https://doi.org/10.1016/0197-2456(89)90005-6).
- [96] H. J. Schünemann and G. H. Guyatt, “Commentary—goodbye M(C)ID! Hello MID, where do you come from?” *Health services research*, vol. 40, no. 2, pp. 593–7, 2005, ISSN: 0017-9124. DOI: [10.1111/j.1475-6773.2005.00374.x](https://doi.org/10.1111/j.1475-6773.2005.00374.x).
- [97] P. Royston, D. G. Altman, and W. Sauerbrei, “Dichotomizing continuous predictors in multiple regression: a bad idea,” *Statistics in Medicine*, vol. 25, no. 1, pp. 127–141, Jan. 2006, ISSN: 0277-6715. DOI: [10.1002/sim.2331](https://doi.org/10.1002/sim.2331).
- [98] M. Futamura, Y. A. Leshem, K. S. Thomas, H. Nankervis, H. C. Williams, and E. L. Simpson, “A systematic review of Investigator Global Assessment (IGA) in atopic dermatitis (AD) trials: Many options, no standards.” *Journal of the American Academy of Dermatology*, vol. 74, no. 2, pp. 288–94, Feb. 2016, ISSN: 1097-6787. DOI: [10.1016/j.jaad.2015.09.062](https://doi.org/10.1016/j.jaad.2015.09.062).

- [99] A. G. Copay, B. R. Subach, S. D. Glassman, D. W. Polly, and T. C. Schuler, “Understanding the minimum clinically important difference: a review of concepts and methods,” *The Spine Journal*, vol. 7, no. 5, pp. 541–546, Sep. 2007, ISSN: 1529-9430. DOI: [10.1016/J.SPINEE.2007.01.008](https://doi.org/10.1016/J.SPINEE.2007.01.008).
- [100] D. Revicki, R. D. Hays, D. Cella, and J. Sloan, *Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes*, Feb. 2008. DOI: [10.1016/j.jclinepi.2007.03.012](https://doi.org/10.1016/j.jclinepi.2007.03.012).
- [101] S. K. Rai, J. Yazdany, P. R. Fortin, and J. A. Aviña-Zubieta, “Approaches for estimating minimal clinically important differences in systemic lupus erythematosus,” *Arthritis research & therapy*, vol. 17, no. 1, p. 143, Jun. 2015, ISSN: 1478-6362. DOI: [10.1186/s13075-015-0658-6](https://doi.org/10.1186/s13075-015-0658-6).
- [102] M. E. Schram, P. I. Spuls, M. M. G. Leeflang, R. Lindeboom, J. D. Bos, and J. Schmitt, “EASI, (objective) SCORAD and POEM for atopic eczema: responsiveness and minimal clinically important difference,” *Allergy*, vol. 67, no. 1, pp. 99–106, Jan. 2012, ISSN: 01054538. DOI: [10.1111/j.1398-9995.2011.02719.x](https://doi.org/10.1111/j.1398-9995.2011.02719.x).
- [103] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009, pp. 1–763, ISBN: 978-0-387-84857-0. DOI: [10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7).
- [104] R. van de Schoot, S. Depaoli, R. King, *et al.*, “Bayesian statistics and modelling,” *Nature Reviews Methods Primers*, vol. 1, no. 1, p. 1, Dec. 2021, ISSN: 2662-8449. DOI: [10.1038/s43586-020-00001-2](https://doi.org/10.1038/s43586-020-00001-2).
- [105] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis, third edition*. CRC Press, Jan. 2013, pp. 1–646, ISBN: 9781439898208. DOI: [10.1201/b16018](https://doi.org/10.1201/b16018).
- [106] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E. J. Wagenmakers, “The fallacy of placing confidence in confidence intervals,” *Psychonomic Bulletin and Review*, vol. 23, no. 1, pp. 103–123, 2016, ISSN: 15315320. DOI: [10.3758/s13423-015-0947-8](https://doi.org/10.3758/s13423-015-0947-8).
- [107] A. Gelman, “Objections to Bayesian statistics,” *Bayesian Analysis*, vol. 3, no. 3, pp. 445–449, Sep. 2008, ISSN: 1936-0975. DOI: [10.1214/08-BA318](https://doi.org/10.1214/08-BA318).
- [108] A. Gelman, “Rejoinder,” *Bayesian Analysis*, vol. 3, no. 3, pp. 467–478, Sep. 2008, ISSN: 1936-0975. DOI: [10.1214/08-BA318REJ](https://doi.org/10.1214/08-BA318REJ).
- [109] A. Gelman and Y. Yao, “Holes in Bayesian statistics,” *Journal of Physics G: Nuclear and Particle Physics*, vol. 48, no. 1, p. 014 002, Jan. 2021, ISSN: 0954-3899. DOI: [10.1088/1361-6471/abc3a5](https://doi.org/10.1088/1361-6471/abc3a5).

- [110] R. Silberzahn, E. L. Uhlmann, D. P. Martin, *et al.*, “Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results,” *Advances in Methods and Practices in Psychological Science*, p. 251 524 591 774 764, 2018, ISSN: 2515-2459. DOI: [10.1177/2515245917747646](https://doi.org/10.1177/2515245917747646).
- [111] A. Gelman and E. Loken, “The Statistical Crisis in Science,” *American Scientist*, vol. 102, no. 6, p. 460, 2014, ISSN: 0003-0996. DOI: [10.1511/2014.111.460](https://doi.org/10.1511/2014.111.460).
- [112] J. G. Ibrahim, M. H. Chen, Y. Gwon, and F. Chen, “The power prior: Theory and applications,” *Statistics in Medicine*, vol. 34, no. 28, pp. 3724–3749, 2015, ISSN: 10970258. DOI: [10.1002/sim.6728](https://doi.org/10.1002/sim.6728).
- [113] A. Gelman and C. Hennig, “Beyond subjective and objective in statistics,” *Journal of the Royal Statistical Society. Series A: Statistics in Society*, vol. 180, no. 4, pp. 967–1033, 2017, ISSN: 1467985X. DOI: [10.1111/rssa.12276](https://doi.org/10.1111/rssa.12276).
- [114] A. Gelman, D. Simpson, and M. Betancourt, “The prior can often only be understood in the context of the likelihood,” *Entropy*, vol. 19, no. 10, p. 555, Oct. 2017, ISSN: 10994300. DOI: [10.3390/e19100555](https://doi.org/10.3390/e19100555).
- [115] T. T. P. Minka, “Expectation Propagation for Approximate Bayesian Inference,” in *Conference on Uncertainty in Artificial Intelligence*, Jan. 2001, pp. 362–369, ISBN: 1-55860-800-1.
- [116] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC., 2011, vol. 20116022, p. 619, ISBN: 978-1-4200-7941-8. DOI: [10.1201/b10905](https://doi.org/10.1201/b10905).
- [117] M. D. Hoffman and A. Gelman, “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research*, vol. 15, pp. 1593–1623, Nov. 2011, ISSN: 15337928.
- [118] A. Gelman and D. B. Rubin, “Inference from Iterative Simulation Using Multiple Sequences,” *Statistical Science*, vol. 7, no. 4, pp. 457–472, Nov. 1992, ISSN: 0883-4237. DOI: [10.1214/ss/1177011136](https://doi.org/10.1214/ss/1177011136).
- [119] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner, “Rank-Normalization, Folding, and Localization: An Improved \hat{R} for Assessing Convergence of MCMC (with Discussion),” *Bayesian Analysis*, vol. 16, no. 2, Jun. 2021, ISSN: 1936-0975. DOI: [10.1214/20-BA1221](https://doi.org/10.1214/20-BA1221).
- [120] M. Betancourt, *Towards a Principled Bayesian Workflow*, 2018.
- [121] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman, “Visualization in Bayesian workflow,” *J. R. Statist. Soc. A*, vol. 182, pp. 1–14, 2017, ISSN: 08966273. DOI: [10.1109/NAFIPS.2004.1337440](https://doi.org/10.1109/NAFIPS.2004.1337440).

- [122] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman, “Validating Bayesian Inference Algorithms with Simulation-Based Calibration,” Apr. 2018.
- [123] T. Gneiting, F. Balabdaoui, and A. E. Raftery, “Probabilistic forecasts, calibration and sharpness,” *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 69, no. 2, pp. 243–268, Apr. 2007, ISSN: 13697412. DOI: [10.1111/j.1467-9868.2007.00587.x](https://doi.org/10.1111/j.1467-9868.2007.00587.x).
- [124] P. C. Austin and E. W. Steyerberg, “Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers,” *Statistics in Medicine*, vol. 33, no. 3, pp. 517–535, Feb. 2014, ISSN: 02776715. DOI: [10.1002/sim.5941](https://doi.org/10.1002/sim.5941).
- [125] A. Gelman, J. Hwang, and A. Vehtari, “Understanding predictive information criteria for Bayesian models,” *Statistics and Computing*, vol. 24, no. 6, pp. 997–1016, Nov. 2014, ISSN: 15731375. DOI: [10.1007/s11222-013-9416-2](https://doi.org/10.1007/s11222-013-9416-2).
- [126] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, Mar. 2007, ISSN: 01621459. DOI: [10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437).
- [127] F. E. Harrell, *Regression Modeling Strategies*, ser. Springer Series in Statistics. Springer International Publishing, 2015, ISBN: 978-3-319-19424-0. DOI: [10.1007/978-3-319-19425-7](https://doi.org/10.1007/978-3-319-19425-7).
- [128] A. Jordan, F. Krueger, and S. Lerch, *scoringRules: Scoring Rules for Parametric and Simulated Distribution Forecasts*, 2017.
- [129] R. J. Hyndman and G. Athanasopoulos, *Forecasting : principles and practice*. OTexts: Melbourne, Australia, 2018, p. 291, ISBN: 9780987507105.
- [130] P.-C. Bürkner, J. Gabry, and A. Vehtari, “Approximate leave-future-out cross-validation for Bayesian time series models,” *Journal of Statistical Computation and Simulation*, vol. 90, no. 14, pp. 2499–2523, Sep. 2020, ISSN: 0094-9655. DOI: [10.1080/00949655.2020.1783262](https://doi.org/10.1080/00949655.2020.1783262).
- [131] A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, ISSN: 14764687. DOI: [10.1038/nature21056](https://doi.org/10.1038/nature21056).
- [132] T. J. Brinker, A. Hekler, A. H. Enk, *et al.*, “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task,” *European Journal of Cancer*, vol. 113, pp. 47–54, May 2019, ISSN: 09598049. DOI: [10.1016/j.ejca.2019.04.001](https://doi.org/10.1016/j.ejca.2019.04.001).
- [133] N. Hameed, A. M. Shabut, and M. A. Hossain, “Multi-Class Skin Diseases Classification Using Deep Convolutional Neural Network and Support Vector Machine,” in *2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, vol. 2018-Decem, IEEE, Dec. 2018, pp. 1–7, ISBN: 978-1-5386-9141-0. DOI: [10.1109/SKIMA.2018.8631525](https://doi.org/10.1109/SKIMA.2018.8631525).

- [134] D. Padilla, A. Yumang, A. L. Diaz, and G. Inlong, "Differentiating Atopic Dermatitis and Psoriasis Chronic Plaque using Convolutional Neural Network MobileNet Architecture," in *2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, IEEE, Nov. 2019, pp. 1–6, ISBN: 978-1-7281-3044-6. DOI: [10.1109/HNICEM48295.2019.9073482](https://doi.org/10.1109/HNICEM48295.2019.9073482).
- [135] H. Wu, H. Yin, H. Chen, *et al.*, "A deep learning, image based approach for automated diagnosis for inflammatory skin diseases," *Annals of Translational Medicine*, vol. 8, no. 9, pp. 581–581, May 2020, ISSN: 23055839. DOI: [10.21037/atm.2020.04.39](https://doi.org/10.21037/atm.2020.04.39).
- [136] M. N. Alam, T. T. K. Munia, K. Tavakolian, F. Vasefi, N. Mackinnon, and R. Fazel-Rezai, "Automatic detection and severity measurement of eczema using image processing," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, IEEE, Aug. 2016, pp. 1365–1368, ISBN: 9781457702204. DOI: [10.1109/EMBC.2016.7590961](https://doi.org/10.1109/EMBC.2016.7590961).
- [137] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, p. 180 161, Dec. 2018, ISSN: 2052-4463. DOI: [10.1038/sdata.2018.161](https://doi.org/10.1038/sdata.2018.161).
- [138] A. Pal, A. Chaturvedi, U. Garain, A. Chandra, R. Chatterjee, and S. Senapati, "Severity assessment of psoriatic plaques using deep CNN based ordinal classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11041 LNCS, Springer Verlag, Sep. 2018, pp. 252–259, ISBN: 9783030012007. DOI: [10.1007/978-3-030-01201-4_{_}27](https://doi.org/10.1007/978-3-030-01201-4_{_}27).
- [139] K. Thomas, K. Koller, T. Dean, *et al.*, "A multicentre randomised controlled trial and economic evaluation of ion-exchange water softeners for the treatment of eczema in children: the Softened Water Eczema Trial (SWET)," *eng, Health Technology Assessment*, vol. 15, no. 08, pp. 5–156, Feb. 2011, ISSN: 1366-5278. DOI: [10.3310/hta15080](https://doi.org/10.3310/hta15080).
- [140] K. S. Thomas, T. Dean, C. O'Leary, *et al.*, "A randomised controlled trial of ion-exchange water softeners for the treatment of eczema in children," *PLoS Medicine*, vol. 8, no. 2, A. Sheikh, Ed., e1000395, Feb. 2011, ISSN: 15491277. DOI: [10.1371/journal.pmed.1000395](https://doi.org/10.1371/journal.pmed.1000395).
- [141] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems*, vol. 4, no. January, pp. 3320–3328, Nov. 2014, ISSN: 10495258.
- [142] M. Abadi, P. Barham, J. Chen, *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, USENIX Association, 2016, pp. 265–283, ISBN: 9781931971331. DOI: [10.5555/3026877.3026899](https://doi.org/10.5555/3026877.3026899).

- [143] J. Huang, V. Rathod, C. Sun, *et al.*, “Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017-Janua, IEEE, Jul. 2017, pp. 3296–3297, ISBN: 978-1-5386-0457-1. DOI: [10.1109/CVPR.2017.351](https://doi.org/10.1109/CVPR.2017.351).
- [144] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, ISSN: 0162-8828. DOI: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031).
- [145] A. G. Howard, M. Zhu, B. Chen, *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 2017.
- [146] E. Frank and M. Hall, “A simple approach to ordinal classification,” in *Machine Learning: ECML 2001. Lecture Notes in Computer Science*, vol. 2167, Springer Verlag, 2001, pp. 145–156, ISBN: 3540425365. DOI: [10.1007/3-540-44795-4{_}13](https://doi.org/10.1007/3-540-44795-4_{_}13).
- [147] J. S. Cardoso and J. F. Pinto Da Costa, “Learning to classify ordinal data: The data replication method,” *Journal of Machine Learning Research*, vol. 8, pp. 1393–1429, 2007, ISSN: 15324435.
- [148] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, “Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning,” in *ICLR 2020 : Eighth International Conference on Learning Representations, 2020*, pp. 1–29.
- [149] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014, ISSN: 15337928.
- [150] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the Inception Architecture for Computer Vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-Decem, IEEE, Jun. 2016, pp. 2818–2826, ISBN: 978-1-4673-8851-1. DOI: [10.1109/CVPR.2016.308](https://doi.org/10.1109/CVPR.2016.308).
- [151] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-Decem, IEEE, Jun. 2016, pp. 770–778, ISBN: 978-1-4673-8851-1. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [152] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, Sep. 2015.

- [153] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *34th International Conference on Machine Learning, ICML 2017*, vol. 3, 2017, pp. 2130–2143, ISBN: 9781510855144.
- [154] M. Groh, C. Harris, L. Soenksen, *et al.*, “Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1820–1828, Apr. 2021, ISSN: 21607516. DOI: [10.1109/CVPRW53098.2021.00201](https://doi.org/10.1109/CVPRW53098.2021.00201).
- [155] B. P. Kaufman, E. Guttman-Yassky, and A. F. Alexis, “Atopic dermatitis in diverse racial and ethnic groups—Variations in epidemiology, genetics, clinical presentation and treatment,” *Experimental Dermatology*, vol. 27, no. 4, pp. 340–357, Apr. 2018, ISSN: 1600-0625. DOI: [10.1111/EXD.13514](https://doi.org/10.1111/EXD.13514).
- [156] M. S. Junayed, A. N. M. Sakib, N. Anjum, M. B. Islam, and A. A. Jeny, “EczemaNet: A Deep CNN-based Eczema Diseases Classification,” in *4th International Conference on Image Processing, Applications and Systems, IPAS 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 174–179, ISBN: 9781728175744. DOI: [10.1109/IPAS50080.2020.9334929](https://doi.org/10.1109/IPAS50080.2020.9334929).
- [157] C. H. Bang, J. W. Yoon, J. Y. Ryu, *et al.*, “Automated severity scoring of atopic dermatitis patients by a deep neural network,” *Scientific Reports*, vol. 11, no. 1, p. 6049, Dec. 2021, ISSN: 20452322. DOI: [10.1038/s41598-021-85489-8](https://doi.org/10.1038/s41598-021-85489-8).
- [158] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, p. 78, 2012, ISSN: 00010782. DOI: [10.1145/2347736.2347755](https://doi.org/10.1145/2347736.2347755).
- [159] E. Domínguez-Hüttinger, P. Christodoulides, K. Miyauchi, *et al.*, “Mathematical modeling of atopic dermatitis reveals “double-switch” mechanisms underlying 4 common disease phenotypes,” *Journal of Allergy and Clinical Immunology*, vol. 139, no. 6, pp. 1861–1872, 2017, ISSN: 10976825. DOI: [10.1016/j.jaci.2016.10.026](https://doi.org/10.1016/j.jaci.2016.10.026).
- [160] A. Simpson, V. Y. F. Tan, J. Winn, *et al.*, “Beyond Atopy,” *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 11, pp. 1200–1206, Jun. 2010, ISSN: 1073-449X. DOI: [10.1164/rccm.200907-1101OC](https://doi.org/10.1164/rccm.200907-1101OC).
- [161] Y. Zhang and K. Berhane, “Bayesian Mixed Hidden Markov Models: A Multi-Level Approach to Modeling Categorical Outcomes with Differential Misclassification,” *Stat Med*, vol. 33, no. 8, pp. 1395–1408, 2014. DOI: [10.1002/sim.6039](https://doi.org/10.1002/sim.6039).
- [162] P. Christodoulides, Y. Hirata, E. Domínguez-Hüttinger, *et al.*, “Computational design of treatment strategies for proactive therapy on atopic dermatitis using optimal control theory,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 375, no. 2096, p. 20 160 285, Jun. 2017, ISSN: 1364-503X. DOI: [10.1098/rsta.2016.0285](https://doi.org/10.1098/rsta.2016.0285).

- [163] S. M. Langan, P. Silcocks, and H. C. Williams, “What causes flares of eczema in children?” *The British journal of dermatology*, vol. 161, no. 3, pp. 640–6, Sep. 2009, ISSN: 1365-2133. DOI: [10.1111/j.1365-2133.2009.09320.x](https://doi.org/10.1111/j.1365-2133.2009.09320.x).
- [164] N. Steen, A. Hutchinson, E. Mccoll, *et al.*, “Development of a symptom based outcome measure for asthma,” *BMJ*, vol. 309, no. 6961, p. 1065, Oct. 1994, ISSN: 14685833. DOI: [10.1136/bmj.309.6961.1065](https://doi.org/10.1136/bmj.309.6961.1065).
- [165] G. Tanaka, E. Domínguez-Hüttinger, P. Christodoulides, K. Aihara, and R. J. Tanaka, “Bifurcation analysis of a mathematical model of atopic dermatitis to determine patient-specific effects of treatments on dynamic phenotypes,” *Journal of Theoretical Biology*, vol. 448, pp. 66–79, 2018, ISSN: 10958541. DOI: [10.1016/j.jtbi.2018.04.002](https://doi.org/10.1016/j.jtbi.2018.04.002).
- [166] Y. M. Kim, J. Kim, Y. Han, B. H. Jeon, H. K. Cheong, and K. Ahn, “Short-term effects of weather and air pollution on atopic dermatitis symptoms in children: A panel study in Korea,” *PLoS ONE*, vol. 12, no. 4, Y. R. Kou, Ed., e0175229, Apr. 2017, ISSN: 19326203. DOI: [10.1371/journal.pone.0175229](https://doi.org/10.1371/journal.pone.0175229).
- [167] H. Williams, A. Stewart, E. von Mutius, W. Cookson, and H. R. Anderson, “Is eczema really on the increase worldwide?” *Journal of Allergy and Clinical Immunology*, vol. 121, no. 4, pp. 947–954, Apr. 2008, ISSN: 00916749. DOI: [10.1016/j.jaci.2007.11.004](https://doi.org/10.1016/j.jaci.2007.11.004).
- [168] T. Passeron, J. Krutmann, M. L. Andersen, R. Katta, and C. C. Zouboulis, “Clinical and biological impact of the exposome on the skin,” *Journal of the European Academy of Dermatology and Venereology*, vol. 34, no. S4, pp. 4–25, Jul. 2020, ISSN: 14683083. DOI: [10.1111/jdv.16614](https://doi.org/10.1111/jdv.16614).
- [169] K. Ahn, B. E. Kim, J. Kim, and D. Y. Leung, “Recent advances in atopic dermatitis,” *Current Opinion in Immunology*, vol. 66, pp. 14–21, Oct. 2020, ISSN: 18790372. DOI: [10.1016/j.coi.2020.02.007](https://doi.org/10.1016/j.coi.2020.02.007).
- [170] K. Ahn, “The role of air pollutants in atopic dermatitis,” *Journal of Allergy and Clinical Immunology*, vol. 134, no. 5, pp. 993–999, Nov. 2014, ISSN: 10976825. DOI: [10.1016/j.jaci.2014.09.023](https://doi.org/10.1016/j.jaci.2014.09.023).
- [171] P. Kathuria and J. I. Silverberg, “Association of pollution and climate with atopic eczema in US children,” *Pediatric Allergy and Immunology*, vol. 27, no. 5, pp. 478–485, Aug. 2016, ISSN: 09056157. DOI: [10.1111/pai.12543](https://doi.org/10.1111/pai.12543).
- [172] J.-O. Baek, J. Cho, and J.-Y. Roh, “Associations between ambient air pollution and medical care visits for atopic dermatitis,” *Environmental Research*, vol. 195, p. 110 153, Apr. 2021, ISSN: 00139351. DOI: [10.1016/j.envres.2020.110153](https://doi.org/10.1016/j.envres.2020.110153).

- [173] Y.-M. Kim, J. Kim, K. Jung, S. Eo, and K. Ahn, "The effects of particulate matter on atopic dermatitis symptoms are influenced by weather type: Application of spatial synoptic classification (SSC).," *International journal of hygiene and environmental health*, vol. 221, no. 5, pp. 823–829, Jun. 2018, ISSN: 1618-131X. DOI: [10.1016/j.ijheh.2018.05.006](https://doi.org/10.1016/j.ijheh.2018.05.006).
- [174] S. R. Noh, J. S. Kim, E. H. Kim, *et al.*, "Spectrum of susceptibility to air quality and weather in individual children with atopic dermatitis," *Pediatric Allergy and Immunology*, vol. 30, no. 2, J. Genuneit, Ed., pp. 179–187, Mar. 2019, ISSN: 13993038. DOI: [10.1111/pai.13005](https://doi.org/10.1111/pai.13005).
- [175] V. Patella, G. Florio, M. Palmieri, *et al.*, "Atopic dermatitis severity during exposure to air pollutants and weather changes with an Artificial Neural Network (ANN) analysis," *Pediatric Allergy and Immunology*, vol. 31, no. 8, J. Genuneit, Ed., pp. 938–945, Nov. 2020, ISSN: 0905-6157. DOI: [10.1111/pai.13314](https://doi.org/10.1111/pai.13314).
- [176] J. Y. Lee, M. Kim, H. K. Yang, *et al.*, "Reliability and validity of the Atopic Dermatitis Symptom Score (ADSS)," *Pediatric Allergy and Immunology*, vol. 29, no. 3, pp. 290–295, May 2018, ISSN: 13993038. DOI: [10.1111/pai.12865](https://doi.org/10.1111/pai.12865).
- [177] T. S. Tang, T. Bieber, and H. C. Williams, "Does "autoreactivity" play a role in atopic dermatitis?" *Journal of Allergy and Clinical Immunology*, vol. 129, no. 5, pp. 1209–1215, May 2012, ISSN: 00916749. DOI: [10.1016/j.jaci.2012.02.002](https://doi.org/10.1016/j.jaci.2012.02.002).
- [178] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, no. JUN, p. 524, Jun. 2019, ISSN: 16648021. DOI: [10.3389/fgene.2019.00524](https://doi.org/10.3389/fgene.2019.00524).
- [179] J. Thijs, T. Krastev, S. Weidinger, *et al.*, "Biomarkers for atopic dermatitis," *Current Opinion in Allergy & Clinical Immunology*, vol. 15, no. 5, pp. 453–460, Oct. 2015, ISSN: 1528-4050. DOI: [10.1097/ACI.000000000000198](https://doi.org/10.1097/ACI.000000000000198).
- [180] J. L. Thijs, R. Fiechter, C. A. Bruijnzeel-Koomen, *et al.*, "EASI p-EASI: Utilizing a combination of serum biomarkers offers an objective measurement tool for disease severity in atopic dermatitis patients," *Journal of Allergy and Clinical Immunology*, vol. 140, no. 6, pp. 1703–1705, Aug. 2017, ISSN: 10976825. DOI: [10.1016/j.jaci.2017.06.046](https://doi.org/10.1016/j.jaci.2017.06.046).
- [181] L. Krause, V. Mourantchian, K. Brockow, *et al.*, "A computational model to predict severity of atopic eczema from 30 serum proteins," *The Journal of Allergy and Clinical Immunology*, vol. 138, pp. 1207–1210, 2016. DOI: [10.1016/j.jaci.2016.04.017](https://doi.org/10.1016/j.jaci.2016.04.017).
- [182] J. G. Holm, G. Hurault, T. Agner, *et al.*, "Immunoinflammatory Biomarkers in Serum Are Associated with Disease Severity in Atopic Dermatitis," *Dermatology*, vol. 237, no. 4, pp. 513–520, Mar. 2021, ISSN: 1018-8665. DOI: [10.1159/000514503](https://doi.org/10.1159/000514503).

- [183] R. Jurakic Tonic, I. Jakasa, Y. Sun, *et al.*, “Stratum corneum markers of innate and T helper cell-related immunity and their relation to the disease severity in Croatian patients with atopic dermatitis,” *Journal of the European Academy of Dermatology and Venereology*, vol. 35, no. 5, pp. 1186–1196, May 2021, ISSN: 0926-9959. DOI: [10.1111/jdv.17132](https://doi.org/10.1111/jdv.17132).
- [184] E. Roekevisch, K. Szegedi, D. P. Hack, *et al.*, “Effect of immunosuppressive treatment on biomarkers in adult atopic dermatitis patients,” *Journal of the European Academy of Dermatology and Venereology*, vol. 34, no. 7, pp. 1545–1554, Jul. 2020, ISSN: 14683083. DOI: [10.1111/jdv.16164](https://doi.org/10.1111/jdv.16164).
- [185] V. Kiiski, O. Karlsson, A. Remitz, and S. Reitamo, “High Serum Total IgE Predicts Poor Long-term Outcome in Atopic Dermatitis,” *Acta Dermato Venereologica*, vol. 95, no. 8, pp. 943–947, Nov. 2015, ISSN: 0001-5555. DOI: [10.2340/00015555-2126](https://doi.org/10.2340/00015555-2126).
- [186] T. Nakahara, K. Izuhara, D. Onozuka, *et al.*, “Exploration of biomarkers to predict clinical improvement of atopic dermatitis in patients treated with dupilumab,” *Medicine*, vol. 99, no. 38, e22043, Sep. 2020, ISSN: 0025-7974. DOI: [10.1097/MD.00000000000022043](https://doi.org/10.1097/MD.00000000000022043).
- [187] J. Piironen and A. Vehtari, “Sparsity information and regularization in the horseshoe and other shrinkage priors,” *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5018–5051, Jan. 2017, ISSN: 1935-7524. DOI: [10.1214/17-EJS1337SI](https://doi.org/10.1214/17-EJS1337SI).
- [188] C. M. Carvalho, N. G. Polson, and J. G. Scott, “Handling Sparsity via the Horseshoe,” in *International Conference on Artificial Intelligence and Statistics*, 2009, pp. 73–80.
- [189] G. S. Tiplica, F. Boralevi, P. Konno, *et al.*, “The regular use of an emollient improves symptoms of atopic dermatitis in children: a randomized controlled study,” *Journal of the European Academy of Dermatology and Venereology*, vol. 32, no. 7, pp. 1180–1187, Jul. 2018, ISSN: 14683083. DOI: [10.1111/jdv.14849](https://doi.org/10.1111/jdv.14849).
- [190] K. S. Thomas, L. E. Bradshaw, T. H. Sach, *et al.*, “Silk garments plus standard care compared with standard care for treating eczema in children: A randomised, controlled, observer-blind, pragmatic trial (CLOTHES Trial),” *PLOS Medicine*, vol. 14, no. 4, J. K. Tumwine, Ed., e1002280, Apr. 2017, ISSN: 1549-1676. DOI: [10.1371/journal.pmed.1002280](https://doi.org/10.1371/journal.pmed.1002280).
- [191] A. E. Ades and A. J. Sutton, “Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 169, no. 1, pp. 5–35, Jan. 2006, ISSN: 0964-1998. DOI: [10.1111/j.1467-985X.2005.00377.x](https://doi.org/10.1111/j.1467-985X.2005.00377.x).
- [192] D. C. M. Belgrave, R. Granell, A. Simpson, *et al.*, “Developmental Profiles of Eczema, Wheeze, and Rhinitis: Two Population-Based Birth Cohort Studies,” *PLoS Medicine*, vol. 11, no. 10, B. P. Lanphear, Ed., e1001748, Oct. 2014, ISSN: 1549-1676. DOI: [10.1371/journal.pmed.1001748](https://doi.org/10.1371/journal.pmed.1001748).

- [193] R. McElreath, *Statistical Rethinking*. Chapman and Hall/CRC, Mar. 2020, ISBN: 9780429029608. DOI: [10.1201/9780429029608](https://doi.org/10.1201/9780429029608).
- [194] C. Cinelli, A. Forney, and J. Pearl, “A Crash Course in Good and Bad Controls,” *SSRN Electronic Journal*, Sep. 2020. DOI: [10.2139/ssrn.3689437](https://doi.org/10.2139/ssrn.3689437).
- [195] T. J. VanderWeele, “Principles of confounder selection,” *European Journal of Epidemiology*, vol. 34, no. 3, pp. 211–219, Mar. 2019, ISSN: 0393-2990. DOI: [10.1007/s10654-019-00494-6](https://doi.org/10.1007/s10654-019-00494-6).
- [196] J. L. Hill, “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, Mar. 2011, ISSN: 10618600. DOI: [10.1198/jcgs.2010.08162](https://doi.org/10.1198/jcgs.2010.08162).
- [197] F. Lattimore and D. Rohde, “Replacing the do-calculus with Bayes rule,” Jun. 2019.
- [198] L. Paternoster, O. E. Savenije, J. Heron, *et al.*, “Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts,” *Journal of Allergy and Clinical Immunology*, vol. 141, no. 3, pp. 964–971, Mar. 2018, ISSN: 00916749. DOI: [10.1016/j.jaci.2017.09.044](https://doi.org/10.1016/j.jaci.2017.09.044).
- [199] J. Hensman, M. Rattray, and N. D. Lawrence, “Fast nonparametric clustering of structured time-series,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 2, pp. 383–393, Feb. 2015, ISSN: 01628828. DOI: [10.1109/TPAMI.2014.2318711](https://doi.org/10.1109/TPAMI.2014.2318711).
- [200] I. C. McDowell, D. Manandhar, C. M. Vockley, A. K. Schmid, T. E. Reddy, and B. E. Engelhardt, “Clustering gene expression time series data using an infinite Gaussian process mixture model,” *PLoS Computational Biology*, vol. 14, no. 1, 2018, ISSN: 15537358. DOI: [10.1371/journal.pcbi.1005896](https://doi.org/10.1371/journal.pcbi.1005896).
- [201] S. Aminikhanghahi and D. J. Cook, “A survey of methods for time series change point detection,” *Knowledge and Information Systems*, vol. 51, no. 2, pp. 339–367, May 2017, ISSN: 02193116. DOI: [10.1007/s10115-016-0987-z](https://doi.org/10.1007/s10115-016-0987-z).
- [202] J. D. Hamilton, “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle,” *Econometrica*, vol. 57, no. 2, p. 357, Mar. 1989, ISSN: 00129682. DOI: [10.2307/1912559](https://doi.org/10.2307/1912559).
- [203] A. E. Raftery, M. Kárný, and P. Ettler, “Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill,” *Technometrics*, vol. 52, no. 1, pp. 52–66, Feb. 2010, ISSN: 00401706. DOI: [10.1198/TECH.2009.08104](https://doi.org/10.1198/TECH.2009.08104).

Appendix A

Supplementary figures to Chapter 3

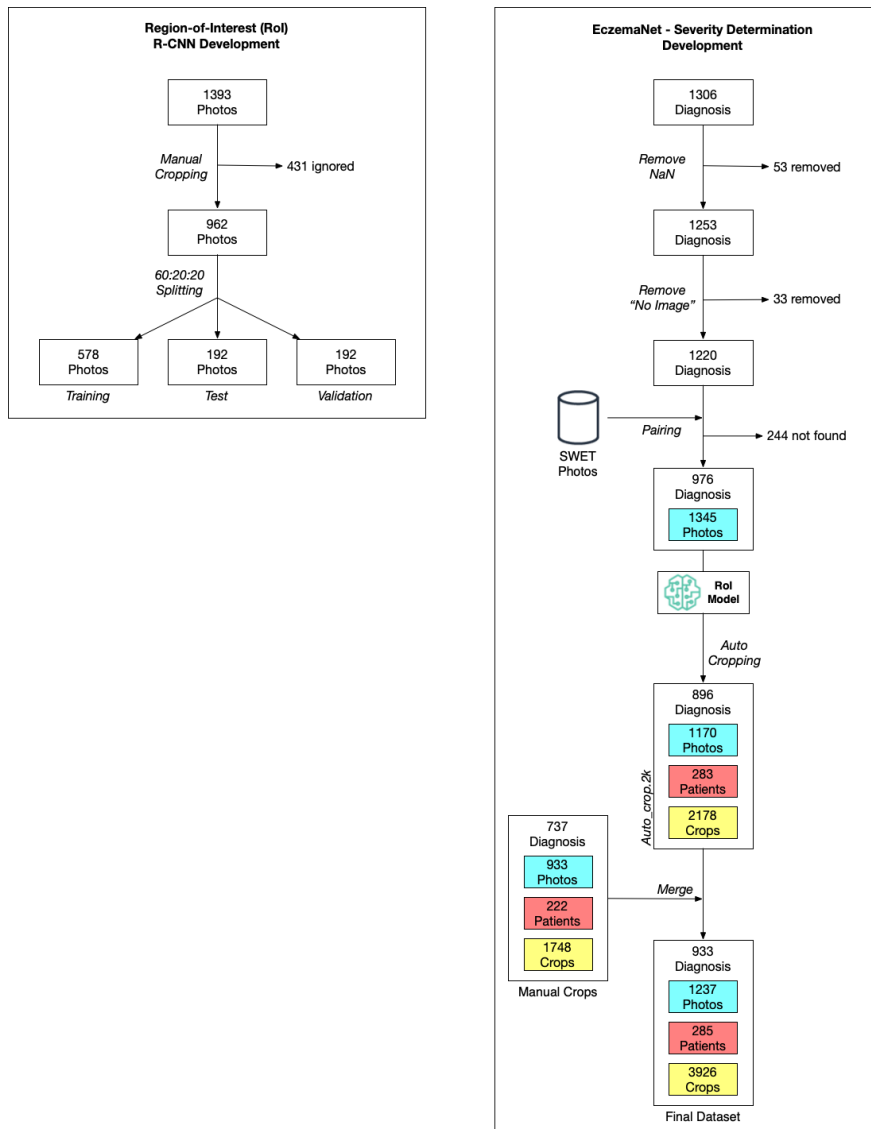


Figure A.1: Data inclusion, exclusion and validation splits

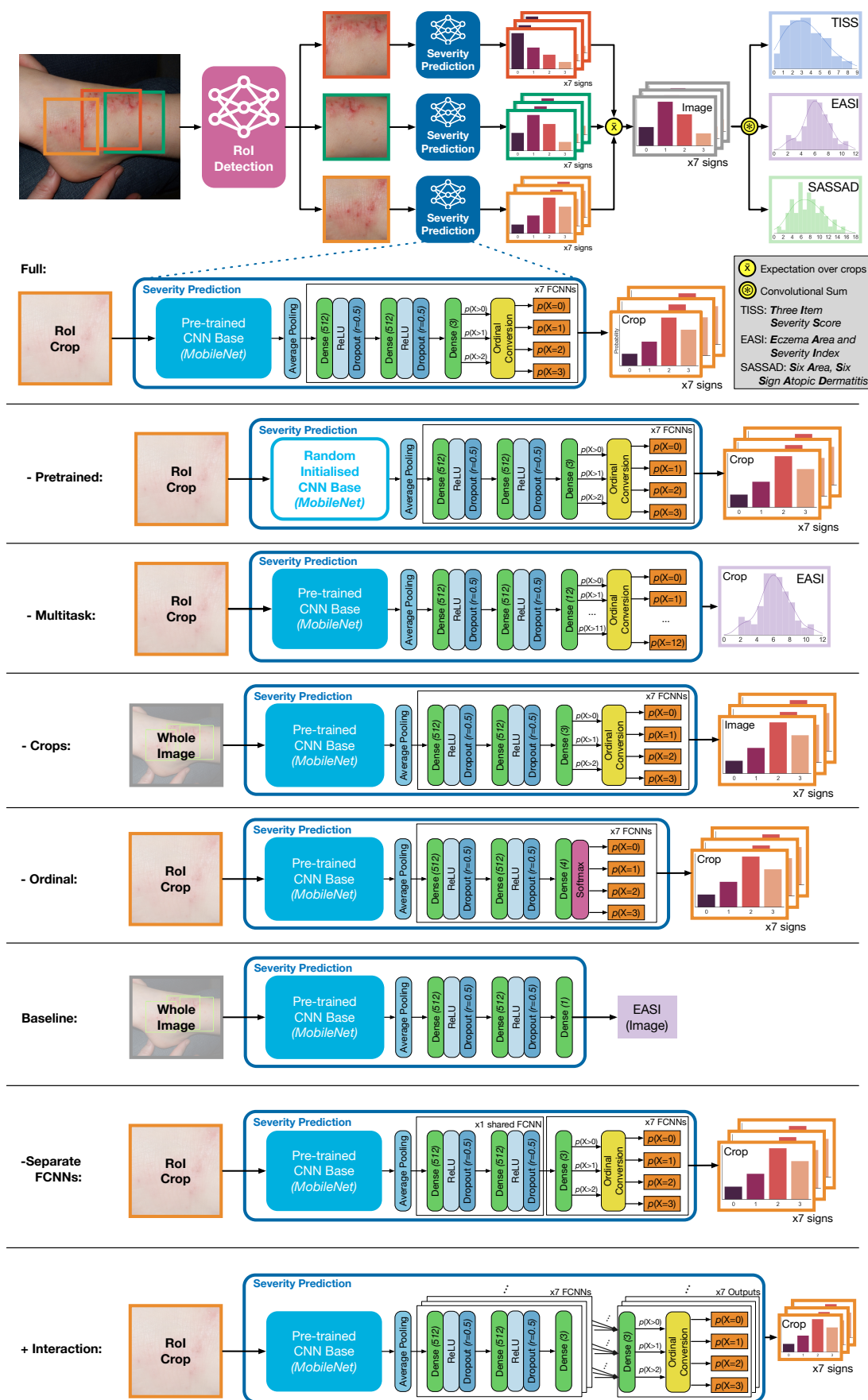


Figure A.2: Architectures of EczemaNet, baselines and ablations

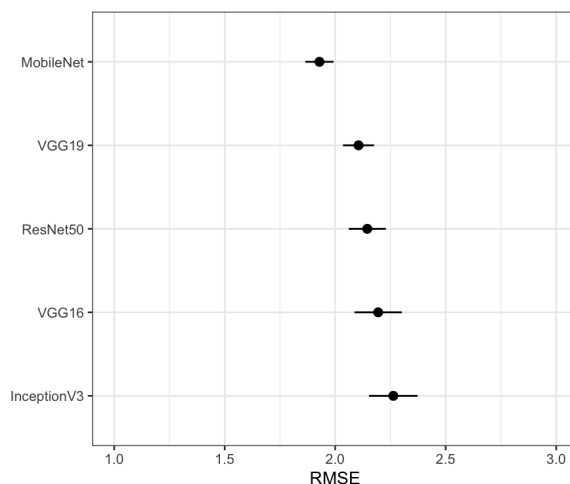


Figure A.3: Base architecture comparisons. RMSE (mean \pm 1 standard error over cross-validation) on EASI across models.

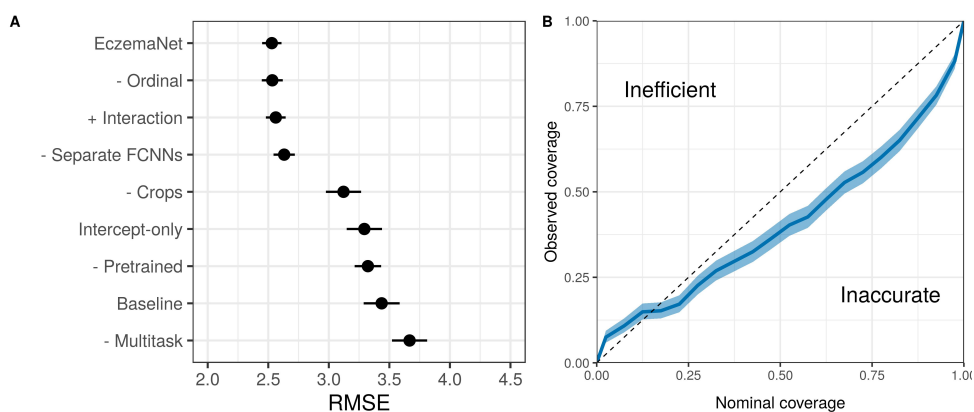


Figure A.4: SASSAD prediction. A) RMSE (mean \pm 1 standard error over cross-validation across models). B) Calibration of highest density prediction intervals (coverage).

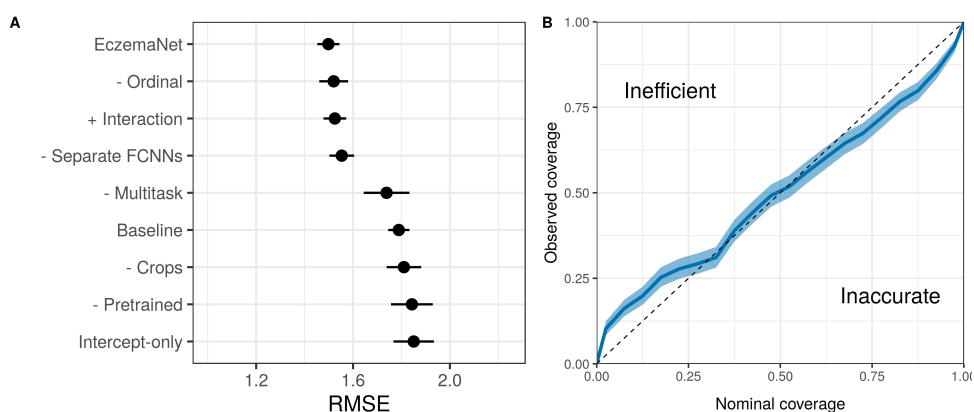


Figure A.5: TISS prediction. A) RMSE (mean \pm 1 standard error over cross-validation) across models. B) Calibration of highest density prediction intervals (coverage).

Appendix B

Appendix to Chapter 4

B.1 Clinical data

B.1.1 Flares dataset

The data comes from a study of 60 children (53% male) with moderate to severe eczema, fulfilling the U.K. refinement of the Hanifin and Rajka diagnostic criteria. The patients were recruited consecutively from the Queen's Medical Centre paediatric dermatology outpatient department, Nottingham, England. The patients had a median age at enrolment of 6.5 years (IQR 8.5 years) and 63% of them were from white ethnicity.

B.1.2 SWET dataset

The data comes from a study of 336 children (57% male) who were identified through secondary care referral centres, primary care, or in response to publicity. All had a diagnosis of eczema according to the UK working party's diagnostic criteria and a minimum eczema severity score of ten points using the Six Area Six Sign Atopic Dermatitis severity score (SASSAD). Patients were aged from 6 months to 16 years and 77% of them were from white ethnicity.

B.2 Description of the extended model

We extended our mechanistic Bayesian model to consider information present in the SWET dataset that was not available in the Flares dataset, notably to investigate the effects of potential

risk factors on severity scores and heterogeneity in treatment responsiveness.

The extended model consists of the measurement process (Eq. (4.1)), and an exponentially modified Gaussian distribution:

$$\hat{y}^{(k)}(t+1) \sim \mathcal{N}_{[0,10]}(\alpha^{(k)}\hat{y}^{(k)}(t) + (\boldsymbol{\theta}^{(k)})^T \mathbf{u}^{(k)}(t) + (\mathbf{x}^{(k)})^T \boldsymbol{\beta} + R^{(k)}(t) + \beta_0, \sigma^2) \quad (\text{B.1})$$

$$R^{(k)}(t) \sim \text{Exp}(\beta = P^{(k)}) \quad (\text{B.2})$$

- $(\boldsymbol{\theta}^{(k)})^T \mathbf{u}^{(k)}(t)$ represents the contribution of exogenous factors, including treatments and whether the patient “slept at home” $u_{Home}^{(k)}(t)$, with:
 $\boldsymbol{\theta}^{(k)} = (\theta_{SU}^{(k)}, \theta_{CS}^{(k)}, \theta_{CI}^{(k)}, \theta_{Home}^{(k)})^T$ and $\mathbf{u}^{(k)}(t) = (u_{SU}^{(k)}(t), u_{CS}^{(k)}(t), u_{CI}^{(k)}(t), u_{Home}^{(k)}(t))^T$.
 The contribution of treatments corresponds to the linear combination of treatment usage (Fig. B.5) for step-up ($u_{SU}^{(k)}(t)$), topical steroid ($u_{CS}^{(k)}(t)$) and calcineurin inhibitors ($u_{CI}^{(k)}(t)$).
- $(\mathbf{x}^{(k)})^T \boldsymbol{\beta}$ represents the contribution of demographics factors, including the presence of filaggrin mutation ($x_{FLG}^{(k)}$), sex ($x_{Sex}^{(k)}$), age ($x_{Age}^{(k)}$), and “white” ethnicity ($x_{White}^{(k)}$), such as:
 $\mathbf{x}^{(k)} = (x_{FLG}^{(k)}, x_{Sex}^{(k)}, x_{Age}^{(k)}, x_{White}^{(k)})^T$ and $\boldsymbol{\beta} = (\beta_{FLG}, \beta_{Sex}, \beta_{Age}, \beta_{White})^T$.

We assumed a hierarchical prior for $\theta_{SU}^{(k)}$:

$$\theta_{SU}^{(k)} \sim \mathcal{N}(\mu_{SU}, \sigma_{SU}^2) \quad (\text{B.3})$$

And we expressed $\theta_{CS}^{(k)}$ and $\theta_{CI}^{(k)}$ as a function of the daily quantity of treatment of different potencies used. For $T \in \{CS, CI\}$ and $P \in \{\text{Mild, Moderate, Potent, Very Potent}\}$:

$$\theta_T^{(k)} = \gamma_T^{(k)} + \sum_P \theta_{T,P} \hat{q}_{T,P}^{(k)} \quad (\text{B.4})$$

- $\hat{q}_{T,P}^{(k)}$ is the estimated daily quantity of treatment T of potency P used (detailed below).
- $\theta_{T,P}$ is the relative contribution of treatment T of potency P to the severity score.
- $\gamma_T^{(k)}$ is the intrinsic responsiveness of the k -th patient to treatment T . We assumed a hierarchical prior for $\gamma_T^{(k)}$:

$$\gamma_T^{(k)} \sim \mathcal{N}(\mu_T, \sigma_T^2) \quad (\text{B.5})$$

Finally, 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, $Q_{T,P}^{(k)}$:

$$\hat{q}_{T,P}^{(k)} = \begin{cases} 0 & \text{if } Q_{T,P}^{(k)} = 0 \\ \frac{\hat{Q}_{T,P}^{(k)}}{N_{T,P}^{(k)}} & \text{otherwise} \end{cases} \quad (\text{B.6})$$

Where $N_{T,P}^{(k)}$ is the number of treatment applications and $\hat{Q}_{T,P}^{(k)}$ is the total quantity of treatment used and estimated by a multiplicative error model:

$$\log(\hat{Q}_{T,P}^{(k)}) \sim \mathcal{N}\left(\log(Q_{T,P}^{(k)}), \frac{\sigma_Q^2}{Conf^{(k)}}\right) \quad (\text{B.7})$$

Where the quantity $Q_{T,P}^{(k)}$ is reported by the k -th patient with a confidence level, $Conf^{(k)} \in \{1 = \text{“not all sure”}, 2 = \text{“not sure”}, 3 = \text{“sure”}, 4 = \text{“very sure”}\}$.

B.3 Missing value imputation

As is often the case with real-world data, especially clinical data, the Flares and SWET datasets contain missing values in the bother score $y^{(k)}(t)$ (38.8% and 1.9%, respectively). Ignoring missing values, for example by removing them entirely, could result in a drastic reduction of the available data and information, especially when we deal with time-series data where the observations are related to each other. Our model explicitly accepts the absence of measurements without requiring prior imputation of $y^{(k)}(t)$, as we model the measurement process of $y^{(k)}(t)$ and the dynamics of the latent score $\hat{y}^{(k)}(t)$ separately. In other words, $\hat{y}^{(k)}(t)$ is a parameter to be inferred regardless of whether $y^{(k)}(t)$ is observed or not.

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. If the model was imputing the missing values for treatment data, it may 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 treatment $u^{(k)}(t)$ (including $u_{CS}^{(k)}(t)$, $u_{CI}^{(k)}(t)$, $u_{SU}^{(k)}(t)$) 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 $Conf^{(k)}$ (2/327 patients) were imputed by “not at all sure”, missing $x_{FLG}^{(k)}$ (22/327 patients) were imputed by 0 (absence of mutation), missing $x_{White}^{(k)}$ (2/327 patients) were imputed by 0 (non-white or do not wish to declare) and missing $u_{Home}^{(k)}(t)$ were imputed by 1 (“slept at home”, the most common answer).

B.4 Choice of priors

We chose the following weakly informative priors for the population parameters:

- $\beta_0, \beta_0 + \theta^{(k)} \sim \mathcal{N}(0, 2^2)$, for the constant term (intercept, intercept + responsiveness)

- when $u^{(k)}(t) = 1$ (a similar prior is given to the constant term in the extended model).
- $\mu_\theta, \mu_{SU}, \mu_{CS}, \mu_{CI} \sim \mathcal{N}(0, 1)$ for the population mean of responsiveness to treatment. We expect the effect to be less than two points of bother in absolute value.
 - $\sigma_\theta, \sigma_{SU}, \sigma_{CS}, \sigma_{CI} \sim \mathcal{N}(0, 0.5^2)$ for the population-level standard deviation of responsiveness to treatment. We expect the standard deviation of responsiveness to be less than one point of bother, which could imply a difference of 4 points between the most responsive and the least responsive patients.
 - $\mu_\alpha \sim \mathcal{N}(0, 1)$ and $\sigma_\alpha \sim \mathcal{N}(0, 1.5^2)$ for the population mean and standard deviation of the autocorrelation (in logit scale). These priors resulted in an approximately uniform distribution for $\alpha^{(k)}$.
 - $\sigma_P \sim \mathcal{N}^+(0, 1)$. As a rule of thumb, this prior indicates that the most plausible values are ≤ 2 , which would imply that the most plausible values for $P^{(k)}$ would be ≤ 4 , and $R^{(k)}(t)$ could take values from 0 to 10 (maximum value of the score).
 - $\sigma \sim \mathcal{N}^+(0, 1.5^2)$. As a rule of thumb, this prior indicates the most plausible values for σ are ≤ 3 , which implies that the width of the distribution could be equal to 12 (compared to a score that ranges from 0 to 10).
 - $\beta_{FLG}, \beta_{Sex}, \beta_{White}, \theta_{Home} \sim \mathcal{N}(0, 0.5^2)$ for the binary risk factors in the extended model. We expect the absolute effect to be less than 1 point.
 - $\beta_{Age} \sim \mathcal{N}(0, 0.1^2)$ for age (in years) in the extended model. We expect that a 10-year difference would have an effect of less than 2 points.
 - $\theta_{CS,Mild}, \theta_{CS,Moderate}, \theta_{CS,Potent}, \theta_{CS,Very Potent}, \theta_{CI,Mild}, \theta_{CI,Moderate} \sim \mathcal{N}(0, 0.5^2)$ for the contribution of quantity in the responsiveness to treatment. This implies that we expect the change in bother per daily gram of treatment is less than 1.

The priors for the parameters of the reference models $(\alpha, \beta_0, \sigma)$ are the same as the priors for the mechanistic Bayesian model.

We originally set a prior for σ_Q , the standard deviation that controls how accurate the reported quantity of treatment is, but the parameter was not identifiable with our data. As a result and to ease the computational burden of the inference algorithm, we chose to set $\sigma_Q = 0.25$, a value that we considered conservative. With $\sigma_Q = 0.25$, the 95% CI of the true quantity of treatment used is between 60% and 165% of the reported quantity, when patients reported that they were “not at all sure” in their reported quantity of treatment $Q_{T,P}^{(k)}$.

Computational problems can arise if most of the mass of the non-truncated distribution predictive distribution of $\hat{y}^{(k)}(t + 1)$ is outside the support of the score ($[0, 10]$). For example, when $\mu = 20$ and $\sigma^2 = 1$, the normalisation constant, $Z = \int_0^{10} f_{\mathcal{N}}(x; \mu, \sigma^2) dx$, is infinitesimal and identifiability issues may arise because the shape of $\mathcal{N}_{[0,10]}(20, 1)$ is very similar to the shape of $\mathcal{N}_{[0,10]}(15, 1)$, for example. To overcome these potential problems, we implemented

a “soft-uniform” prior on the linear predictor term, $\alpha^{(k)}\hat{y}^{(k)}(t) + \theta^{(k)}u^{(k)}(t) + R^{(k)}(t) + \beta_0$, to reflect our expectation that the predicted score, $\hat{y}^{(k)}(t+1)$, should not be “too much” outside $[0, 10]$. The “soft-uniform” prior is defined by the probability density function $f(x) = \frac{\text{logit}^{-1}(x+1) - \text{logit}^{-1}(x-11)}{12}$, resulting in an approximately constant density between 0 and 10 (therefore not prioritising any values in this range) with a slow convergence to a density of 0 (i.e. penalising values) outside this range.

B.5 Learning curves

To investigate whether the model learns/improves its performance as more data comes in, we plotted the evolution of the RPS and lpd as a function of the training iterations of the forward chaining (Fig. 4.3). However, these metrics are not computed on the same population at each iteration due to missing observations (especially in the Flares dataset). Specifically, the sub-populations at a later time were not representative of the entire population, as patients with controlled AD tended to drop out earlier in clinical trials than patients with uncontrolled AD¹. The fact that patients with controlled AD are easier to predict (low noise, low RPS, high lpd) than patients with uncontrolled AD (high noise, high RPS, low lpd) resulted in Simpson’s paradox, where the RPS averaged over available subpopulations hit a minimum then increase, although individual RPS may decrease (Fig. B.1, Section 2.3.2).

We therefore decided to control for this patient-dependency by computing the performance metric $m \in \{RPS, lpd\}$ at the observation-level, proposing a model for m and taking the mean fit as an unbiased estimate of m . We used a Generative Additive Model (GAM) with cubic splines to achieve a flexible fit to the evolution of m while avoiding overfitting. The model was fitted using the `gamm4` package in R to the formula:

$$m \sim [i = 0] + [i > 0] : t + [i > 0] : s(i) + (1|Patient) \quad (\text{B.8})$$

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

¹In other words, missing values due to patients dropping the study early are not missing at random.

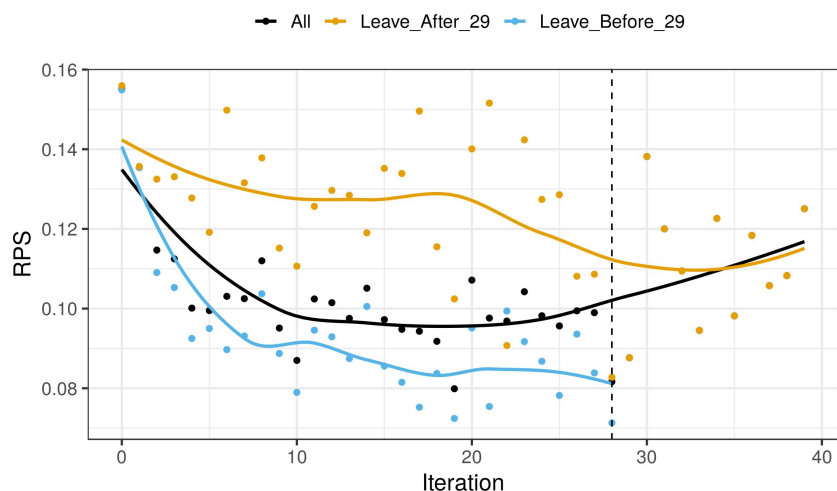


Figure B.1: Learning curves of RPS (smoothed by LOWESS) for the model trained on the Flares dataset. The orange and blue circles correspond to the patients who dropped out of the study after and before the 29th iteration, respectively. The total RPS (black) is the average of the two curves weighted by the proportion of the patients in each group at a given time. The orange and blue curves can both decrease (i.e. the performance improves), while the average increases (Simpson's paradox).

The performance for the first training iteration ($i = 0$) was estimated separately because the model has not started learning yet (the predictive distribution is thus the prior predictive distribution). We also controlled for the prediction horizon ($[i > 0] : t$), since predictions at each iteration were made for the entire week, but the performance was expected to decrease as the prediction horizon increased. We further introduced a mixed effect on the intercept ($(1 | Patient)$) to adjust for patient-dependence.

B.6 Supplementary tables

Table B.1: Posterior summary statistics for the population-level parameters of the model trained on the Flares dataset

Parameter	\hat{R}	N_{eff}	mean	SD	SE	2.5%	50%	97.5%
β_0	1.000	1080	-0.194	0.030	0.001	-0.253	-0.193	-0.136
σ	1.007	191	0.654	0.025	0.002	0.605	0.654	0.702
μ_α	1.005	956	0.211	0.251	0.008	-0.286	0.216	0.698
σ_α	1.000	3637	1.841	0.217	0.004	1.465	1.826	2.314
μ_θ	1.001	6480	-0.235	0.069	0.001	-0.371	-0.234	-0.101
σ_θ	1.002	3269	0.367	0.063	0.001	0.253	0.364	0.503
σ_P	1.009	937	1.669	0.156	0.005	1.389	1.658	2.003

Table B.2: Posterior summary statistics for the population-level parameters of the model trained on the SWET dataset

Parameter	\hat{R}	N_{eff}	mean	SD	SE	2.5%	50%	97.5%
β_0	1.001	2905	0.087	0.013	0.000	0.062	0.087	0.113
σ	1.008	840	0.575	0.005	0.000	0.565	0.575	0.585
μ_α	1.015	581	1.813	0.071	0.003	1.675	1.813	1.948
σ_α	1.005	1181	1.108	0.052	0.002	1.011	1.106	1.213
μ_θ	1.001	1896	-0.207	0.024	0.001	-0.253	-0.207	-0.161
σ_θ	1.004	1468	0.341	0.024	0.001	0.296	0.340	0.389
σ_P	1.010	448	0.904	0.037	0.002	0.834	0.903	0.981

Table B.3: 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%
β_0	Intercept of the evolution of \hat{y}	1.001	2686	0.179	0.035	0.001	0.110	0.179	0.247
σ	Standard deviation of the evolution of \hat{y}	1.005	722	0.572	0.005	0.000	0.562	0.572	0.583
μ_α	Population mean of the persistence (logit scale)	1.007	644	1.809	0.067	0.003	1.681	1.808	1.939
σ_α	Population standard deviation of the persistence (logit scale)	1.001	1195	1.071	0.053	0.002	0.973	1.068	1.181
σ_P	Standard deviation of the relative evolution of P	1.009	667	0.900	0.037	0.001	0.832	0.900	0.975
β_{FLG}	Expected change in the presence of flaggrin mutation	1.001	2311	0.088	0.026	0.001	0.036	0.088	0.140
β_{Sex}	Expected change if the patient is male (compared to female)	1.001	2750	0.041	0.022	0.000	-0.003	0.041	0.083
β_{Age}	Expected change for one extra year of age	1.001	3370	-0.013	0.003	0.000	-0.018	-0.013	-0.007
β_{White}	Expected change when the patient is of white ethnicity versus other/unknown ethnicity	1.001	2361	-0.031	0.027	0.001	-0.084	-0.031	0.023
θ_{Home}	Expected change when the patient is sleeping at home	1.000	5275	-0.040	0.017	0.000	-0.073	-0.040	-0.007
μ_{SU}	Population mean of the responsiveness to step-up	1.000	3880	-0.104	0.021	0.000	-0.146	-0.104	-0.063
σ_{SU}	Population standard deviation of the responsiveness to step-up	1.002	1575	0.200	0.026	0.001	0.149	0.199	0.250
μ_{CS}	Population mean of the responsiveness to corticosteroids	1.002	2787	-0.176	0.027	0.001	-0.229	-0.176	-0.124
σ_{CS}	Population standard deviation of the responsiveness to corticosteroids	1.002	1813	0.248	0.021	0.001	0.207	0.248	0.291
$\theta_{CS,Mild}$	Change in corticosteroids responsiveness due to one daily additional gram of mild corticosteroids	1.003	3146	0.024	0.026	0.000	-0.026	0.024	0.076
$\theta_{CS,Moderate}$	Change in corticosteroids responsiveness due to one daily additional gram of moderate corticosteroids	1.002	3011	0.008	0.018	0.000	-0.026	0.008	0.044
$\theta_{CS,Potent}$	Change in corticosteroids responsiveness due to one daily additional gram of potent corticosteroids	1.001	2915	0.027	0.022	0.000	-0.016	0.028	0.070
$\theta_{CS,VeryPotent}$	Change in corticosteroids responsiveness due to one daily additional gram of very potent corticosteroids	1.002	3242	-0.065	0.076	0.001	-0.222	-0.062	0.079
μ_{CI}	Population mean of the responsiveness to calcineurin inhibitors	1.001	4221	0.041	0.032	0.000	-0.022	0.041	0.103
σ_{CI}	Population standard deviation of the responsiveness to calcineurin inhibitors	1.005	849	0.081	0.044	0.002	0.005	0.082	0.166
$\theta_{CI,Mild}$	Change in calcineurin inhibitors responsiveness due to one daily additional gram of mild calcineurin inhibitors	1.000	3817	0.035	0.049	0.001	-0.059	0.035	0.133
$\theta_{CI,Moderate}$	Change in calcineurin inhibitors responsiveness due to one daily additional gram of moderate calcineurin inhibitors	1.001	1814	-0.476	0.089	0.002	-0.650	-0.476	-0.297

B.7 Supplementary figures

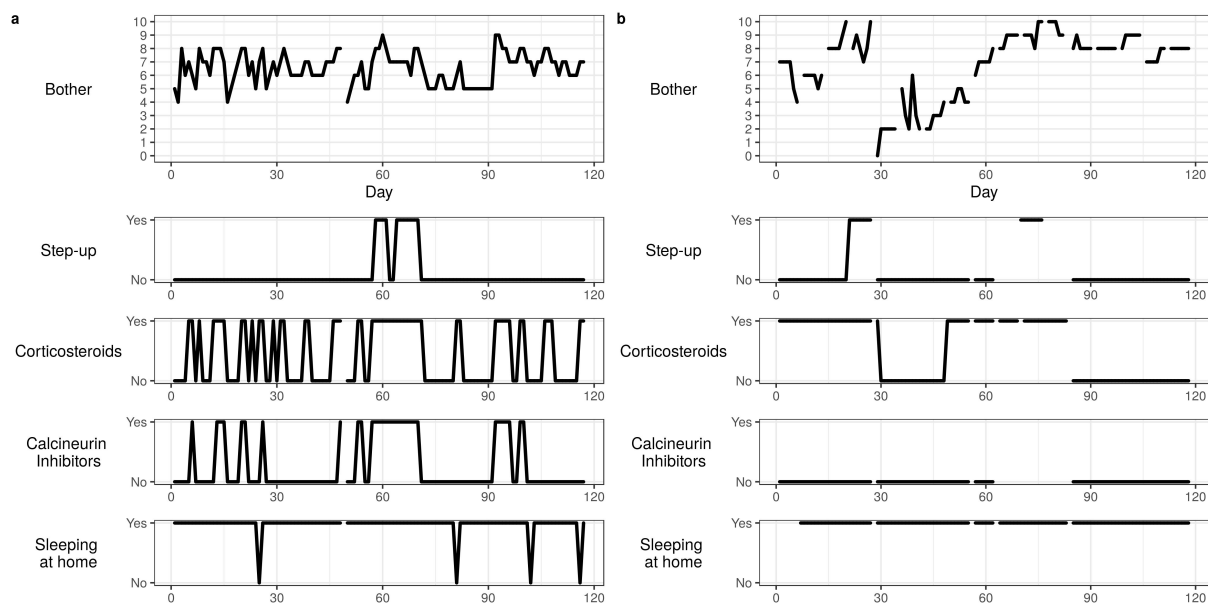


Figure B.2: Example data from SWET dataset. Discontinuities represent missing values. Data from the Flares dataset is similar but with only bother score and step-up, and with more missing values.

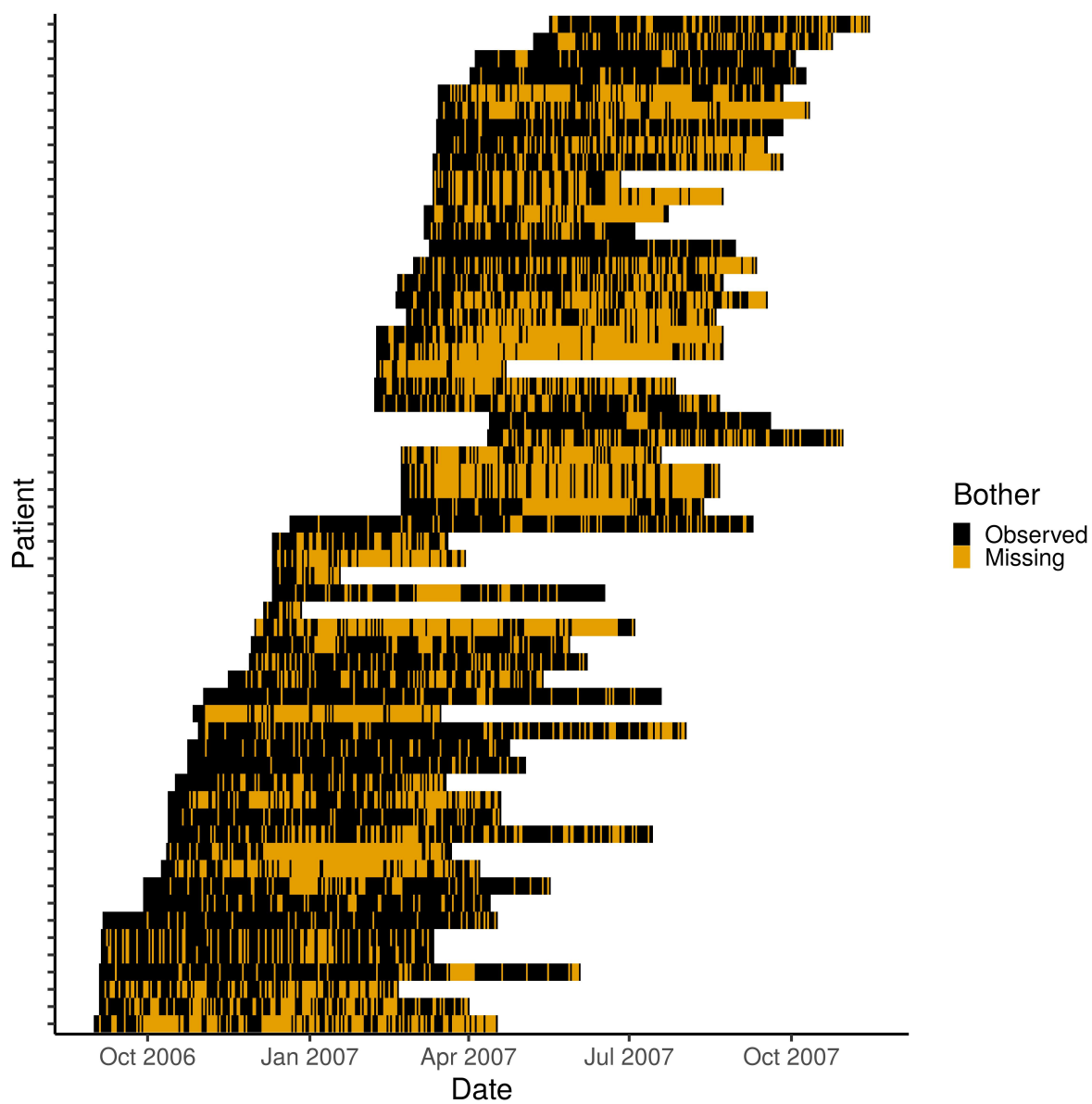


Figure B.3: Missing bother scores in Flares dataset. Black and orange indicate observed and missing scores, respectively. The x-axis indicates the date of the measurement.

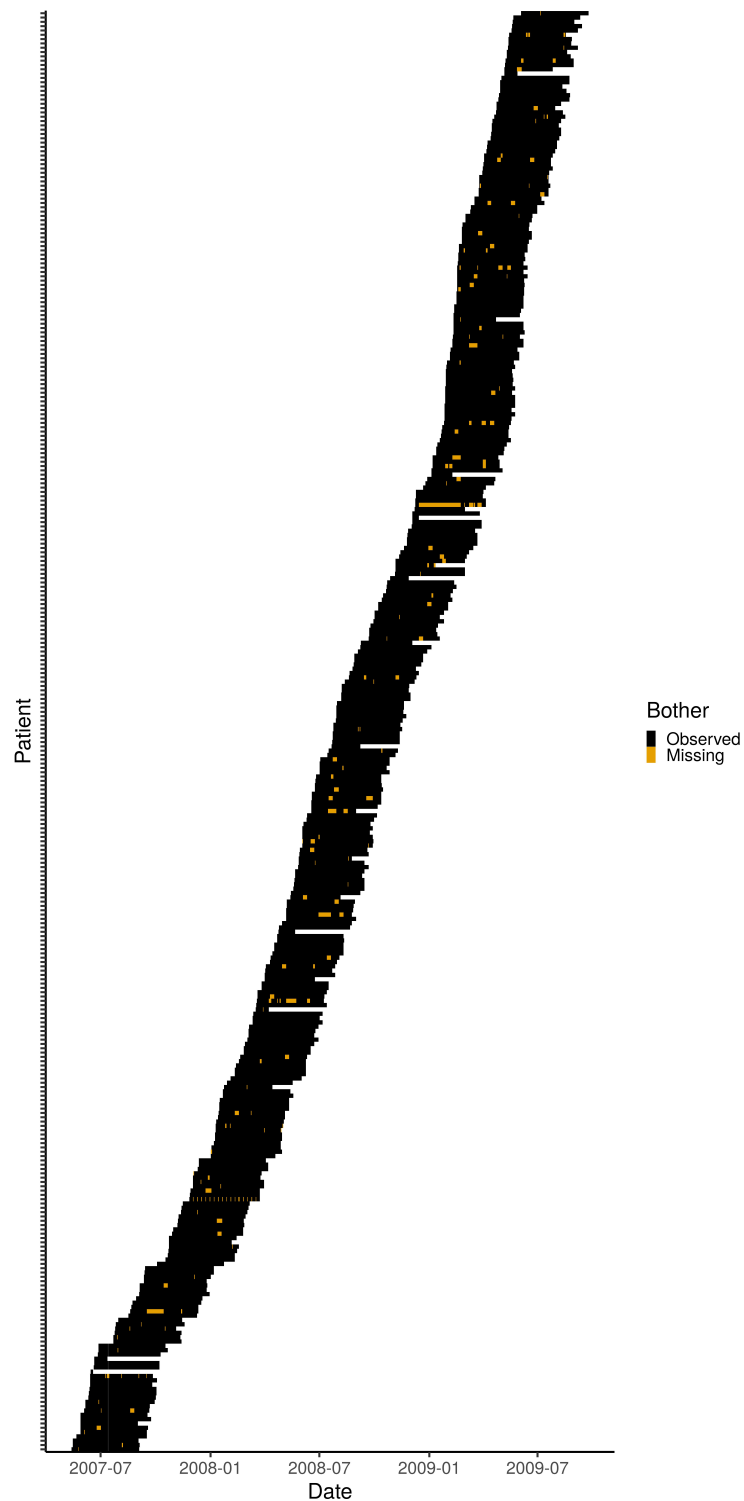


Figure B.4: Missing bother scores in SWET dataset. Black and orange indicate observed and missing scores, respectively. The x-axis indicates the date of the measurement.

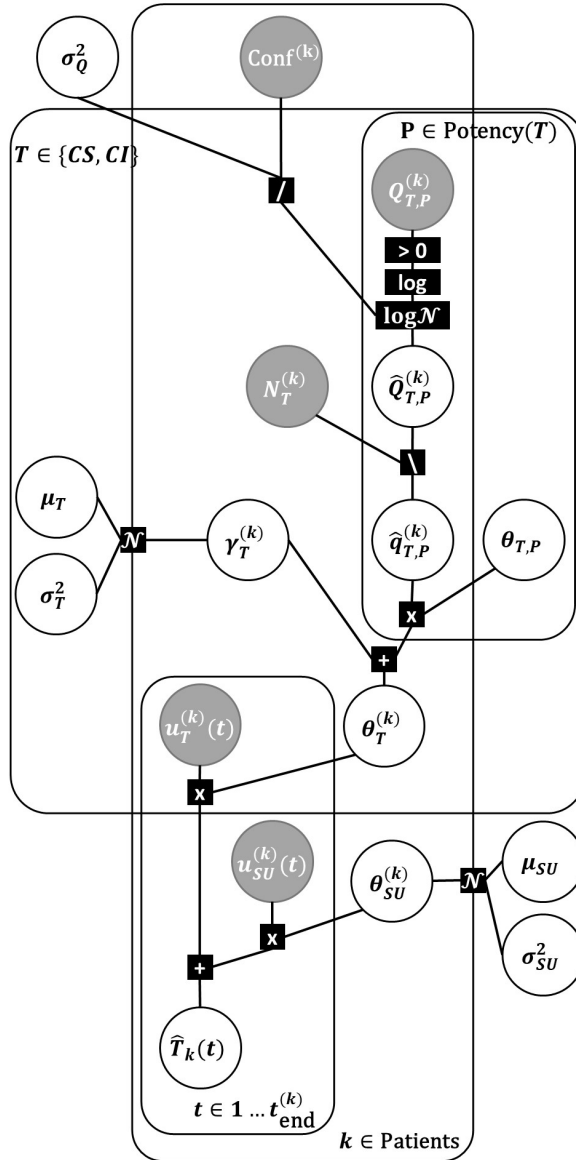


Figure B.5: Factor graph of the treatment term from the extended model (Appendix B.2). The grey and white circles represent the observed and latent variables, respectively. The variables are connected to factors (square nodes) that represent the operations or conditional probability distributions. For 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).

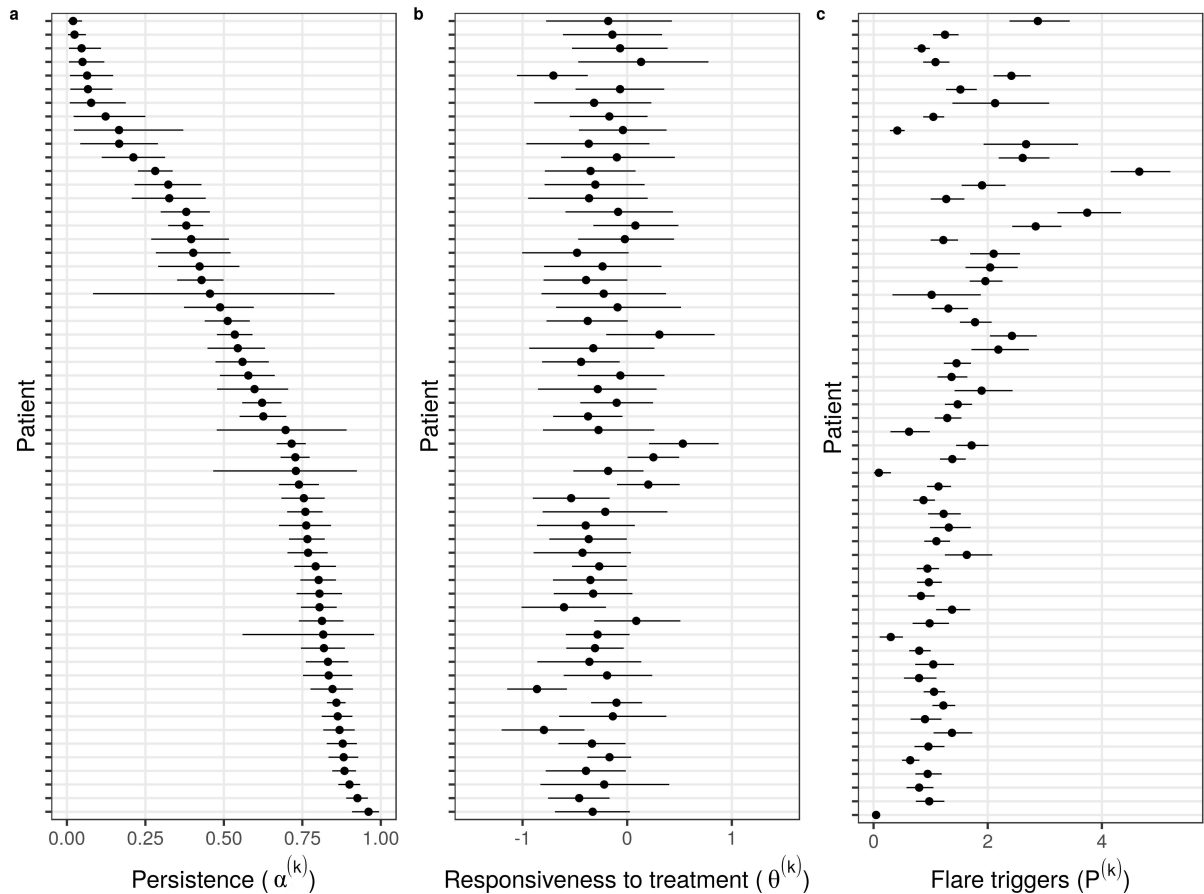


Figure B.6: Estimates of the patient-dependent model parameters ($\alpha^{(k)}, \theta^{(k)}, P^{(k)}$) fitted to Flares dataset. Black circles and the line segments represent the mean posterior and the 90% credible interval, respectively. Estimates greatly vary from one patient to another, confirming their patient-dependence. A: $\alpha^{(k)}$ (persistence of the severity score). The closer $\alpha^{(k)}$ is to 1, the more persistent the severity score is. B: $\theta^{(k)}$ (responsiveness to treatment). The value of $\theta^{(k)}$ quantifies the expected change in the severity score by the treatment. C: $P^{(k)}$ (flare triggers). Higher $P^{(k)}$ is associated with more severe and frequent flares.

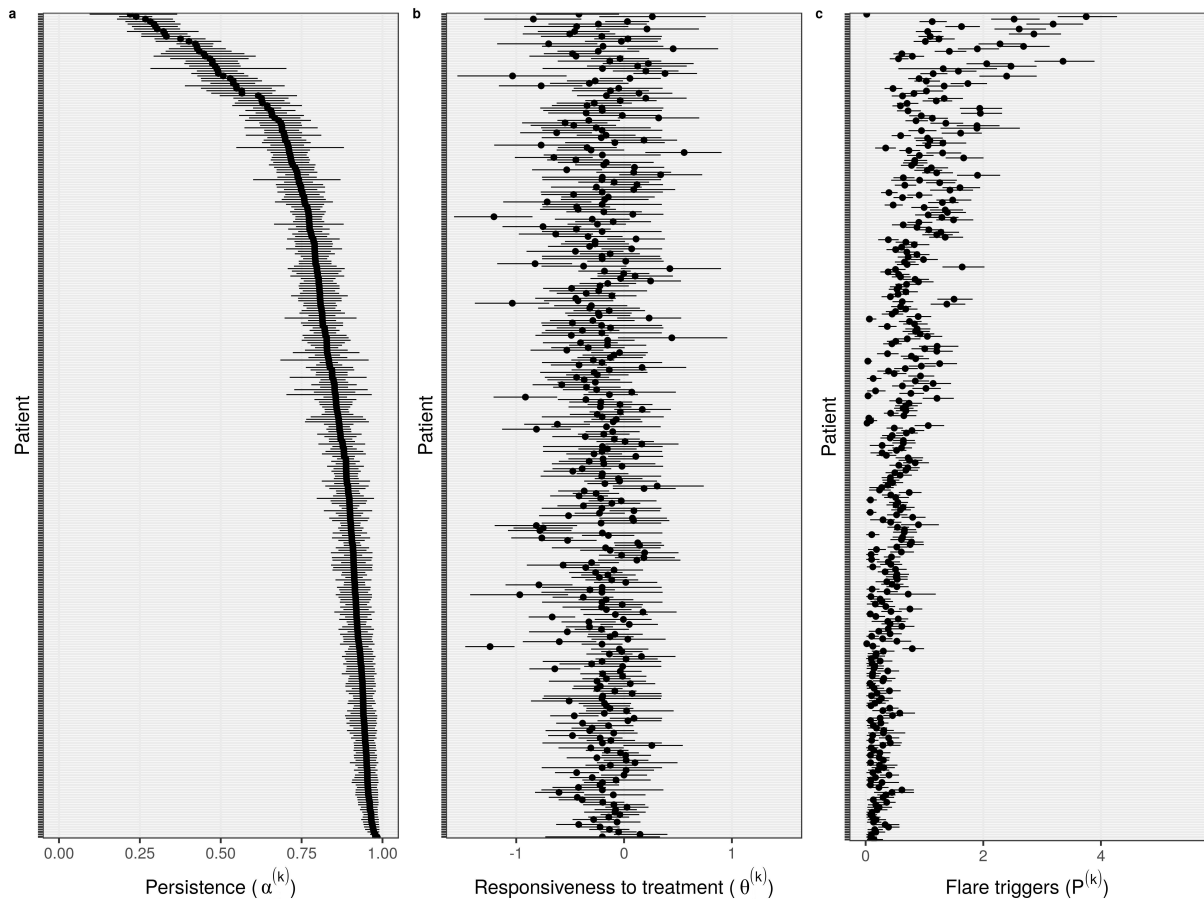


Figure B.7: Estimates of the patient-dependent model parameters $(\alpha^{(k)}, \theta^{(k)}, P^{(k)})$ fitted to SWET dataset. Black circles and the line segments represent the mean posterior and the 90% credible interval, respectively. Estimates greatly vary from one patient to another, confirming their patient-dependence. A: $\alpha^{(k)}$ (persistence of the severity score). The closer $\alpha^{(k)}$ is to 1, the more persistent the severity score is. B: $\theta^{(k)}$ (responsiveness to treatment). The value of $\theta^{(k)}$ quantifies the expected change in the severity score by the treatment. C: $P^{(k)}$ (flare triggers). Higher $P^{(k)}$ is associated with more severe and frequent flares.

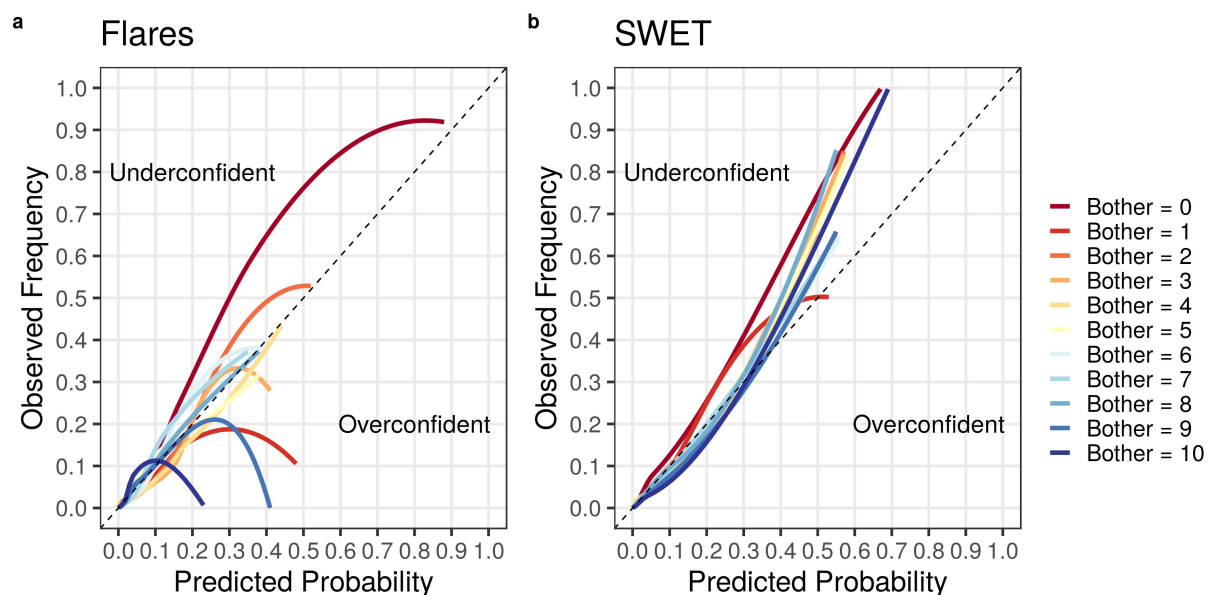


Figure B.8: Calibration curves for the models trained by Flares dataset (a) and SWET dataset (b), obtained using locally weighted scatterplot smoother (LOWESS).

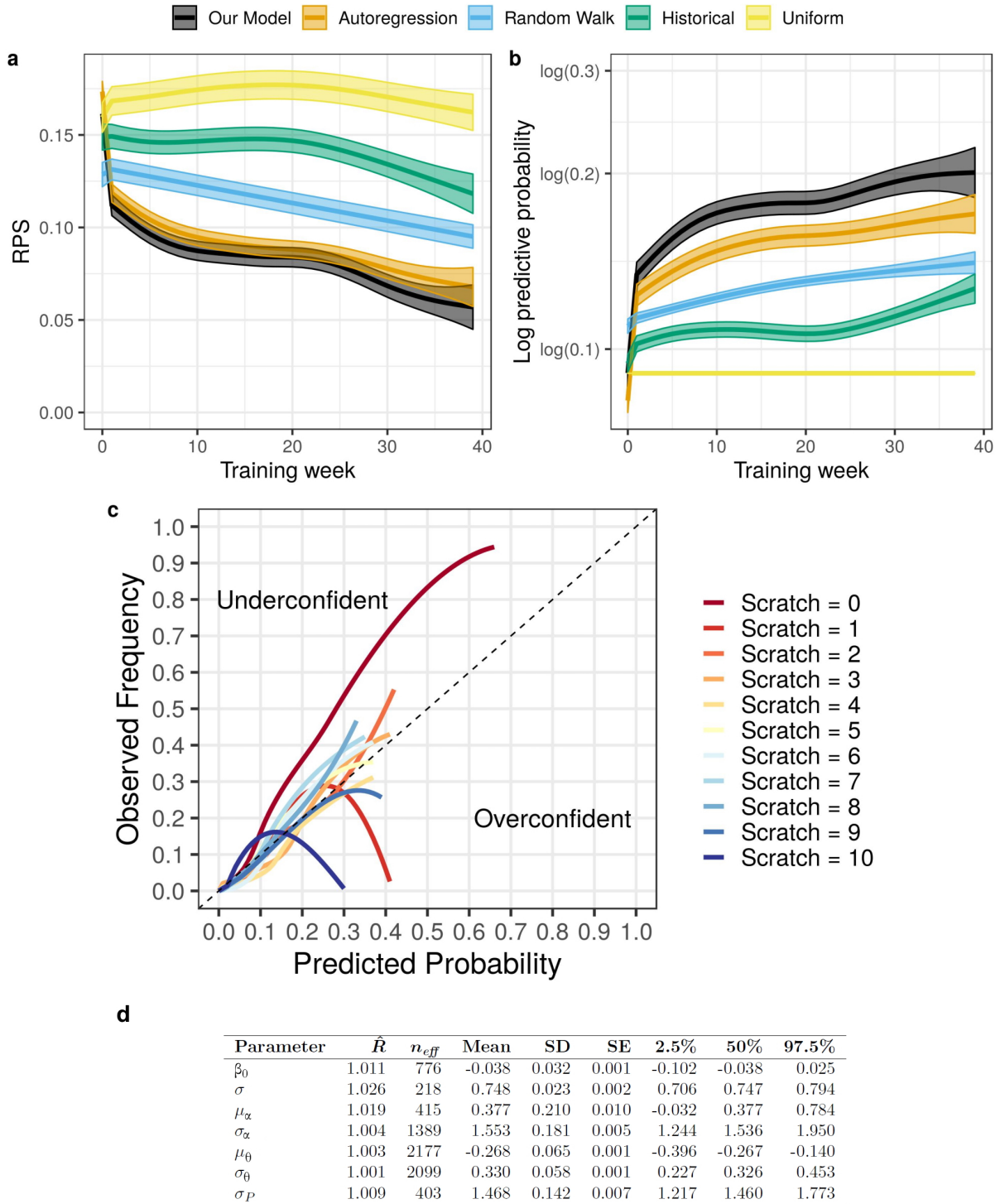


Figure B.9: Performance (A-C) and fit (D) of the model predicting the “scratch” severity score that was only available in Flares dataset. A-B: Learning curves for RPS (A) and lpd (B) for our model (black) compared to the benchmark models. C: Calibration curves. D: Posterior summary statistics of the main parameters.

Appendix C

Supplementary figure to Chapter 5

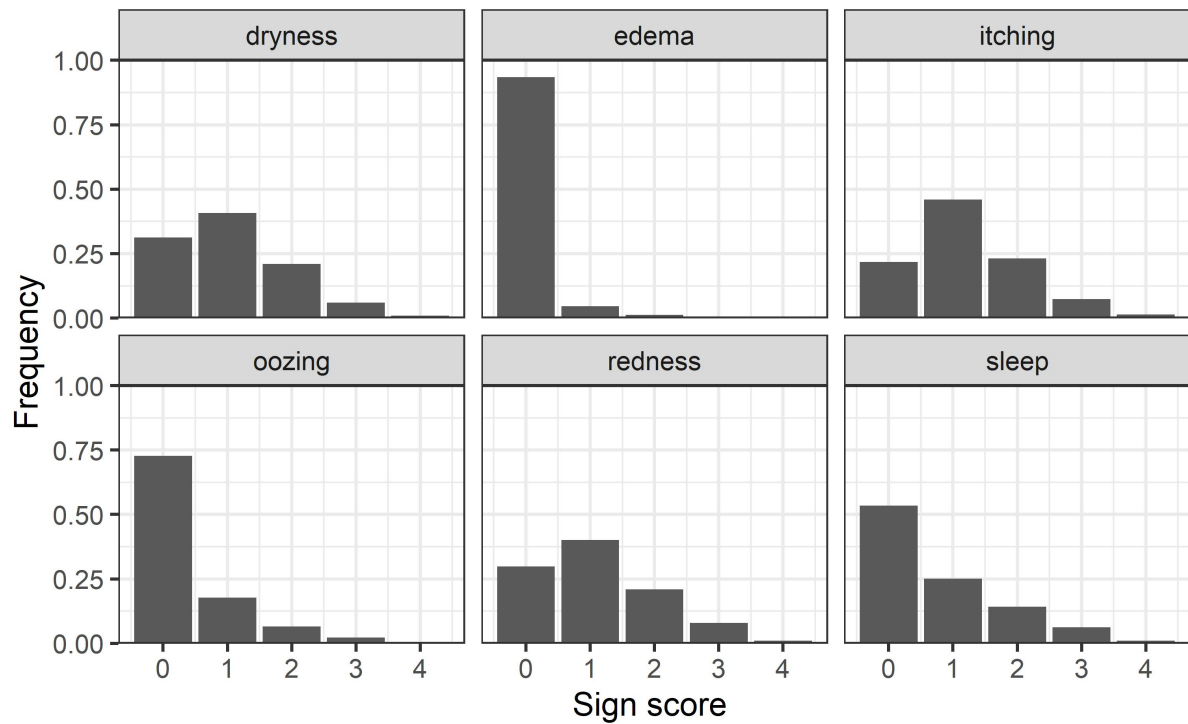


Figure C.1: Distribution of the AD signs scores across time and patients.

Appendix D

Appendix to Chapter 6

D.1 Choice of priors

D.1.1 Priors for the baseline model

We chose to parametrise the model with $\sigma_t = \sqrt{\sigma_m^2 + \sigma_1^2}$, the standard deviation for two-weeks-ahead predictions, and $\rho^2 = \frac{\sigma_m^2}{\sigma_t^2}$, the ratio of the measurement variance to the total variance. ρ^2 can be interpreted similarly to an R-squared (coefficient of determination), as the proportion of the explained variance (the variance of the measurements) in the total variance. We found that a parametrisation in terms (σ_t, ρ^2) resulted in more robust computation and increased effective sample sizes, as opposed to a parametrisation in terms of (σ_m, σ_1) . The priors for σ_t and ρ^2 are given by:

$$\frac{\sigma_t}{M} \sim \log \mathcal{N}\left(-\log(20), (0.5 \log(5))^2\right) \quad (\text{D.1})$$

$$\rho^2 \sim \text{Beta}(4, 2) \quad (\text{D.2})$$

The prior for σ_t is a lognormal distribution with a 95% confidence interval that corresponds to $[0.01M, 0.25M]$. The prior for ρ^2 is Beta distribution to reflect our expectation that future severity scores are predictable (i.e. $\sigma_1 < \sigma_m$).

We assumed a hierarchical prior for the autocorrelation parameter $\alpha^{(k)}$:

$$\alpha^{(k)} \sim \text{Beta}(\mu_\alpha \phi_\alpha, (1 - \mu_\alpha) \phi_\alpha) \quad (\text{D.3})$$

$$\mu_\alpha \sim \text{Beta}(2, 2) \quad (\text{D.4})$$

$$\phi_\alpha \sim \log \mathcal{N}(\log(10), (0.5 \log(10))^2) \quad (\text{D.5})$$

μ_α is the population mean of the Beta distribution, and is given a symmetric prior which slightly favours values around 0.5, as opposed to 0 or 1. ϕ_α is the population pseudo-sample size of the Beta distribution, and is given a lognormal prior with a 95% confidence interval being approximately $[1, 100]$, allowing a wide variety of distributions for $\alpha^{(k)}$, from well spread to concentrated. The resulting marginal prior for $\alpha^{(k)}$ is approximately uniform.

We defined the prior for the intercept $\beta_0^{(k)}$ by introducing the expected value of the autoregressive process $y_\infty^{(k)}$, such that $\beta_0^{(k)} = (1 - \alpha^{(k)})y_\infty^{(k)}$. We assumed a Gaussian hierarchical prior on $y_\infty^{(k)}$:

$$y_\infty^{(k)} \sim \mathcal{N}(\mu_\infty, \sigma_\infty^2) \quad (\text{D.6})$$

$$\frac{\mu_\infty}{M} \sim \mathcal{N}(0.5, 0.25^2) \quad (\text{D.7})$$

$$\frac{\sigma_\infty}{M} \sim \mathcal{N}^+(0, 0.125^2) \quad (\text{D.8})$$

μ_∞ is the population mean of $y_\infty^{(k)}$ and is given a normal prior that covers the range of the score $([0, M])$. σ_∞ is the population standard deviation, and is given a half-normal prior that reflects our assumption that σ_∞ is at most $0.25M$, which would result in the width of the distribution for $y_\infty^{(k)}$ to be at most M . The resulting marginal prior for $y_\infty^{(k)}$ is approximately uniform in $[0, M]$.

Finally, we implemented a soft-uniform prior on the latent score $\hat{y}^{(k)}(t)$ (similarly to Chapter 4) to avoid the situation where most of the mass of the non-truncated distribution for $\hat{y}^{(k)}(t)$ is outside of the range $[0, M]$, which can cause computational problems for the (truncated) measurement distribution. The soft-uniform prior is defined by the probability density function, $f(x) = \frac{\text{logit}^{-1}(x+0.01M) - \text{logit}^{-1}(x-1.01M)}{1.02M}$, resulting in an approximately constant density between 0 and M (i.e. not prioritising any values in this range) with a slow convergence to 0 (i.e. penalising values) outside this range.

D.1.2 Regularised horseshoe prior

We assumed a regularised horseshoe prior [187] for the coefficients β_i ($i = 1, \dots, 30 = D$), defined by:

$$\beta_i \sim \mathcal{N}(0, \tilde{\lambda}_i^2 \tau^2) \quad (\text{D.9})$$

$$\tilde{\lambda}_i = \frac{c^2 \lambda_i^2}{c^2 + \tau^2 \lambda_i^2} \quad (\text{D.10})$$

$$\lambda_i \sim \mathcal{C}^+(0, 1) \quad (\text{D.11})$$

$$\tau \sim \mathcal{C}^+\left(0, \frac{p_0}{D - p_0} \frac{\sigma_1}{\sqrt{N}}\right) \quad (\text{D.12})$$

$$c^2 \sim \text{Inv-}\chi^2(\nu, \sigma_\chi^2) \quad (\text{D.13})$$

Where:

- $\tilde{\lambda}_i$ is the local shrinkage parameter.
- τ is the global shrinkage parameter.
- c represents the scale of the signal. It is given a scaled-inverse chi-squared prior, where we assume the degree of freedom, $\nu = 5$, and the scale, $\sigma_\chi = 1$. This scale-inverse chi-squared prior translates to a Student-t slab with ν degrees of freedom and scale σ_χ for coefficients far from 0. This prior reflects the assumption that the order of magnitude of non-zero coefficients is around 1 but could be higher.
- $p_0 = 5$ is the expected number of covariates with non-zero coefficients.
- $D = 30$ is the number of covariates.
- $N = 42$ is the number of patients.

D.1.3 Reference model priors

- For the random walk model, the prior for σ is the same as that for σ_t in our SSM.
- The autoregressive is an extension of the random walk model, with a uniform prior for $\alpha \sim \mathcal{U}(0, 1)$ and a prior for y_∞ that is the same as the prior for μ_∞ in the SSM.
- The mixed effect autoregressive model extends the autoregressive model, and priors for $\alpha^{(k)}$ and $\beta_0^{(k)}$ are the same as those for the SSM.

D.2 Supplementary tables

Table D.1: Posterior summary statistics of the population-level parameters for the model predicting EASI without covariates.

Parameter	Interpretation	\hat{R}	N_{eff}	Mean	SD	2.5%	50%	97.5%	PS*
σ_t	Square root of the total variance	1.000	1760	4.499	0.283	3.976	4.488	5.094	0.988
ρ^2	Proportion of the σ_m^2 in the total variance	1.002	1055	0.947	0.037	0.853	0.955	0.993	0.955
σ_1	Standard deviation of the latent dynamic	1.003	1089	0.968	0.340	0.371	0.946	1.671	0.939
σ_m	Standard deviation of the measurement process	1.001	1453	4.379	0.310	3.794	4.372	5.013	0.982
MDC	Minimum Detectable Change for a 95% confidence level	1.001	1453	8.582	0.607	7.436	8.568	9.826	0.982
μ_α	Population mean of the autocorrelation parameter α	1.002	1572	0.688	0.042	0.600	0.691	0.765	0.961
ϕ_α	Population “pseudo-sample size” of the autocorrelation α	1.004	1174	3.664	1.094	2.026	3.504	6.187	0.999
μ_∞	Population mean of the autoregressive process	1.002	1597	3.574	0.588	2.446	3.561	4.789	0.997
σ_∞	Population standard deviation of the autoregressive process	1.002	2017	0.507	0.405	0.017	0.421	1.489	0.986

* The Posterior Shrinkage (PS) of the parameter θ is defined as $1 - \frac{\text{Var}(\theta_{\text{post}})}{\text{Var}(\theta_{\text{prior}})}$. PS near 0 indicates that the data provides little information beyond the prior and PS near 1 indicates that the data is much more informative than the prior.

Table D.2: MCID and MDC comparison

	EASI	SCORAD	oSCORAD	POEM
MCID [102]	6.6	8.7	8.2	3.4
MDC	8.6	11.4	9.1	7.7
(mean and 90% CI)	[7.6, 9.6]	[9.1, 13.5]	[7.4, 10.7]	[6.7, 8.9]
E(MDC) / M	8.6/72 = 12%	8.7/103 = 8.4%	9.1/83 = 11%	7.7/28 = 28%

D.3 Supplementary figures

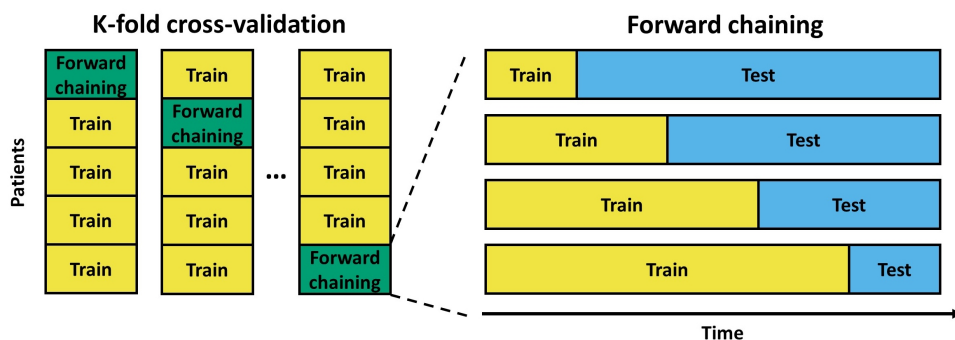


Figure D.1: K-fold cross-validation ($K = 5$ in this example) in a forward chaining setting, which reflects how the model would be used in a clinical setting. For each fold, the model was pre-trained with $(K - 1)$ subsets of patients and validated on the remaining subset of patients in a forward chaining setting.

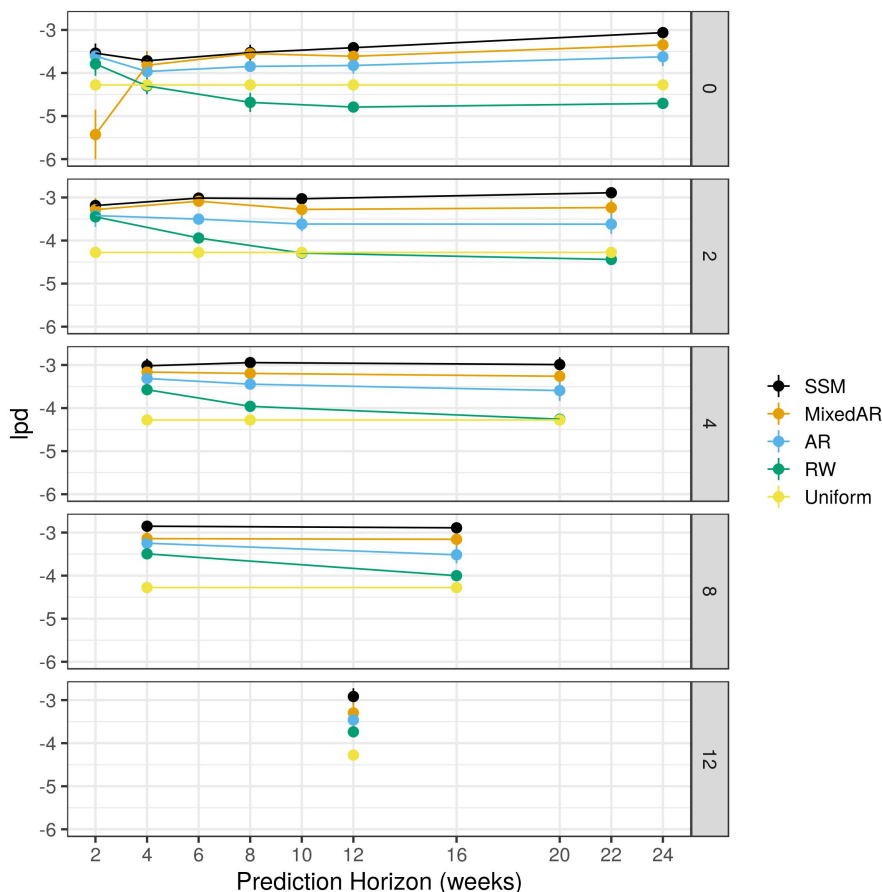


Figure D.2: Performance of our model (SSM) and reference models (MixedAR, AR, RW and Uniform) to predict EASI, evaluated by the lpd (the higher the better). Estimates of lpd are displayed as a function of the prediction horizon for various training weeks (panels) and models (colours).

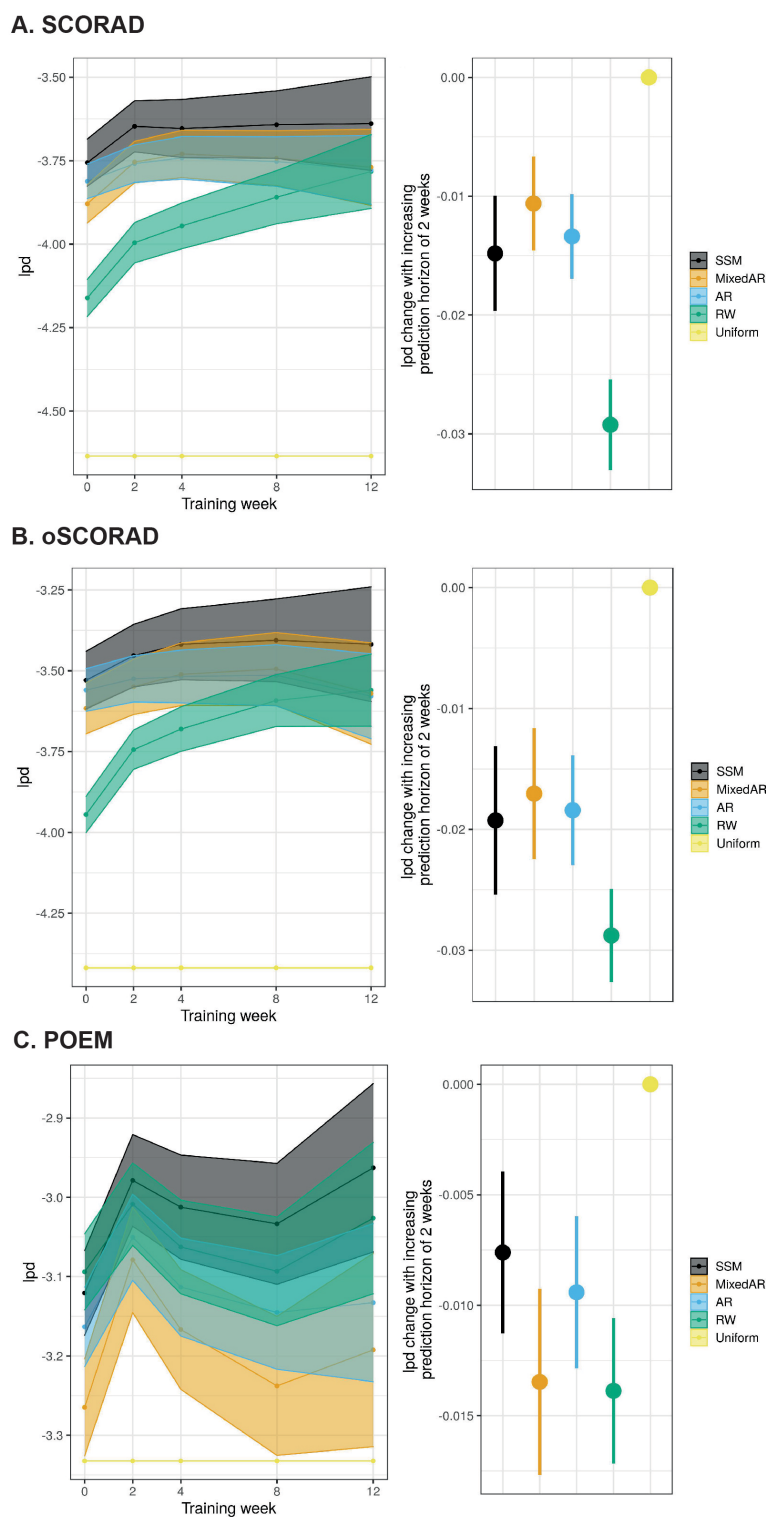


Figure D.3: Predictive performance of our model (SSM) and reference models (MixedAR, AR, RW and Uniform) for oSCORAD (A), SCORAD (B) and POEM (C). The performance was evaluated by lpd (higher the better). Left: Learning curves (mean \pm SE) for two-weeks-ahead predictions, after adjusting for different prediction horizons in a linear model. Right: Change in lpd as the prediction horizon is increased by two weeks.

Appendix E

Appendix to Chapter 7

E.1 EczemaPred models

E.1.1 Binomial Markov chain

The extent model assumes that we can subdivide the body area into 100 patches, each with a probability, \hat{y} , of being classified as lesional, and that each patch has fixed transition probabilities between lesional and non-lesional states. The measurement is specified as a binomial distribution to count the number of lesional patches to produce the extent score, $y = A \sim \mathcal{B}(100, \hat{y})$.

Two-state Markov chain models for latent dynamics

We model the transition from non-lesional to lesional or from lesional to nonlesional for any given patch of the skin with a two-state Markov chain (Fig. E.1) characterised by the transition matrix:

$$T = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1 - p_{01} & p_{01} \\ p_{10} & 1 - p_{10} \end{pmatrix} \quad (\text{E.1})$$

Where:

- p_{10} is the transition probability from lesional to non-lesional states, i.e. the probability that a lesional patch is classified as non-lesional at the next step.
- $p_{11} = 1 - p_{10}$ is the probability that a lesional patch stays lesional and can be interpreted as a measure of eczema persistence.
- p_{01} is the transition probability from non-lesional to lesional states, i.e. the probability that a non-lesional patch is classified as lesional at the next step, and can be interpreted

as a measure of the sensitivity to develop eczema symptoms.

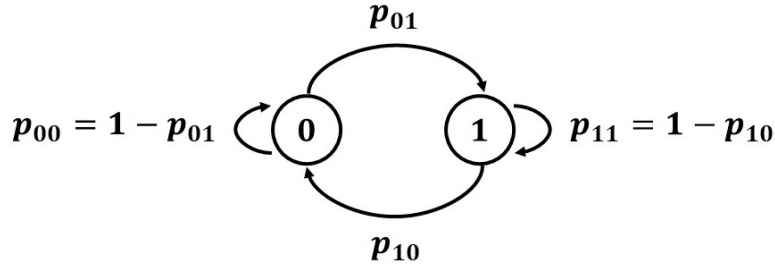


Figure E.1: Two state Markov Chain (0 = “healthy skin”, 1 = “lesional skin”)

The probability, $\hat{y}(t)$, of a given patch being lesional is computed by:

$$\hat{y}(t + 1) = p_{11} \hat{y}(t) + p_{01} (1 - \hat{y}(t)) \quad (\text{E.2})$$

If the state of a skin patch at time t is denoted by $x_t = \begin{pmatrix} 1 & 0 \end{pmatrix}$ when it is non-lesional, and $x_t = \begin{pmatrix} 0 & 1 \end{pmatrix}$ when it is lesional, then the predictions of the state at time $t + h$ is given by $x_{t+h} = x_t T^h$ with:

$$T^h = \begin{pmatrix} 1 - \pi & \pi \\ 1 - \pi & \pi \end{pmatrix} + \lambda^h \begin{pmatrix} \pi & -\pi \\ -(1 - \pi) & 1 - \pi \end{pmatrix} \quad (\text{E.3})$$

$\lambda = 1 - p_{01} - p_{10}$ is one of the eigenvalues of T (the other one is 1) and $\pi = \frac{p_{01}}{p_{01} + p_{10}}$ characterises the steady state (limiting) distribution, $x_\infty = \begin{pmatrix} 1 - \pi & \pi \end{pmatrix}$, to which the Markov chain converges if the prediction horizon h is long enough. The value of λ indicates the mobility of the Markov chain, i.e. how fast it converges to the steady-state distribution, with $|\lambda| \rightarrow 0$ indicating faster convergence.

We assume that the probabilities p_{01} and p_{10} are patient-dependent, and that either one or both probabilities are time-dependent, given that the Markov chain converges to the steady state distribution, $\pi = \frac{p_{01}}{p_{01} + p_{10}}$, which is likely to evolve over a long enough time. We further assume that p_{10} ($= 1 - p_{11}$, where p_{11} is a measure of eczema persistence) does not exhibit a strong time-dependence and that p_{01} (a measure of the sensitivity to develop eczema symptoms) dynamically changes due to endogenous and exogenous factors such as the skin barrier integrity and environmental stressors, respectively. We therefore parametrise the Markov chain by a time (t)- and patient (k)-dependent $\pi^{(k)}(t)$ and a patient-dependent $p_{10}^{(k)}$, from which we can derive:

$$p_{01}^{(k)}(t) = p_{10}^{(k)} \frac{\pi^{(k)}(t)}{1 - \pi^{(k)}(t)}$$

Summary and priors

The evolution of the extent, $y^{(k)}(t)$, for the k -th patient at time t is modelled with a binomial measurement model and latent Markov chain model described above¹:

$$y^{(k)}(t) \sim \mathcal{B}(100, \hat{y}^{(k)}(t)) \quad (\text{E.4})$$

$$\hat{y}^{(k)}(t+1) = p_{11}^{(k)} \hat{y}^{(k)}(t) + p_{01}^{(k)}(t) (1 - \hat{y}^{(k)}(t)) \quad (\text{E.5})$$

$$p_{11}^{(k)} = 1 - p_{10}^{(k)} \quad (\text{E.6})$$

$$p_{01}^{(k)}(t) = p_{10}^{(k)} \frac{\pi^{(k)}(t)}{1 - \pi^{(k)}(t)} \quad (\text{E.7})$$

$$\pi^{(k)}(t) = \frac{\tilde{\pi}^{(k)}(t)}{1 + p_{10}^{(k)}} \quad (\text{E.8})$$

$$\text{logit}(\tilde{\pi}^{(k)}(t+1)) \sim \mathcal{N}(\text{logit}(\tilde{\pi}^{(k)}(t)), \sigma^2) \quad (\text{E.9})$$

With the priors:

$$\sigma \sim \mathcal{N}^+(0, (0.25 \log(5))^2) \quad (\text{E.10})$$

$$p_{10}^{(k)} \sim \text{logit } \mathcal{N}(\mu_{10}, \sigma_{10}^2) \quad (\text{E.11})$$

$$\tilde{\pi}^{(k)}(t_0) \sim \text{logit } \mathcal{N}(-1, 1) \quad (\text{E.12})$$

$$\mu_{10} \sim \mathcal{N}(0, 1) \quad (\text{E.13})$$

$$\sigma_{10} \sim \mathcal{N}^+(0, 1.5^2) \quad (\text{E.14})$$

The priors are chosen to be weakly informative and translate to reasonable prior predictive distributions:

- The prior on σ translates to an odd ratio increment for $\tilde{\pi}$ of at most 5. For example, if $\tilde{\pi}(t) = 0.1$ ($OR(t) = 1/9$), it can evolve up to $\tilde{\pi}(t+1) = 0.36$ ($OR(t+1) = 5OR(t) \approx 0.56$), which could be considered as an unusually important change.
- The prior for the initial condition $\tilde{\pi}^{(k)}(t_0)$ of the first data point (at t_0) is set to be slightly skewed toward 0, as high values of extent are very unlikely.
- The priors on μ_{10} and σ_{10} translate to an approximately uniform prior on $p_{10}^{(k)}$.

Compared to the binomial random walk described below, the Markov chain latent dynamic implies a delay (quantified by $|\lambda|$) between changes in the steady state-distribution π and

¹The model equations include $\tilde{\pi}^{(k)}(t)$ since the upper bound of $\pi^{(k)}(t)$ conditioned on $p_{10}^{(k)}$ is $\frac{1}{1+p_{10}^{(k)}}$, because p_{01} is between 0 and 1.

changes in the latent scores \hat{y} . In particular, if we assume $\lambda = 0$ (no delay), then $\hat{y} = \pi = p_{01} = 1 - p_{01}$, and the model is equivalent to the Binomial random walk if $\tilde{y} = \text{logit}(\hat{y})$ follows a random walk latent dynamic as in Eq. (7.4). In addition, the binomial Markov chain introduces an upper bound for the steady-state distribution π , resulting in a more realistic prior predictive distribution for extent, since high values (e.g. 90% of the body covered by eczema) are very unlikely.

E.1.2 Binomial random walk

The ‘‘Binomial random walk’’ state-space model is characterised by a binomial measurement distribution $\mathcal{D}(\hat{y}^{(k)}(t)) = \mathcal{B}(M, \hat{y}^{(k)}(t))$, with success parameter $\hat{y}^{(k)}(t)$ that follows a latent random walk on the logit scale ($g = \text{logit}^{-1}$). The priors are given by:

$$\sigma \sim \mathcal{N}^+\left(0, (0.25 \log(5))^2\right) \quad (\text{E.15})$$

$$\mu_0 \sim \mathcal{N}(0, 1) \quad (\text{E.16})$$

$$\sigma_0 \sim \mathcal{N}^+(0, 1.5^2) \quad (\text{E.17})$$

The priors are chosen to be weakly informative and translate to reasonable prior predictive distributions:

- The prior on σ translates to an odd ratio increment for \hat{y} of at most 5. For instance, if $\hat{y}(t) = 0.1$ ($OR(t) = 1/9 \approx 0.11$), then the prior assumes it is unusual, but not impossible, that the next value is $\hat{y}(t+1) = 0.56$ ($OR(t+1) = 5OR(t) = 5/9 \approx 0.56$).
- The priors on μ_0 and σ_0 have a reasonable range and translate to a marginal prior for $\hat{y}^{(k)}(t_0)$ that is approximately uniform.

Binomial measurement on the number of days for POEM symptoms

POEM symptoms are graded on a discrete scale from 0 to 4 (0 = ‘‘no days’’, 1 = ‘‘1 or 2 days’’, 2 = ‘‘3 or 4 days’’, 3 = ‘‘5 or 6 days’’, 4 = ‘‘7 days’’). Rather than predict this 0-4 categorisation, we follow the data-generating process by modelling the number of days a symptom occurred, then aggregating the probabilities to obtain predictions in the 0-4 scale.

Let p be the probability of experiencing symptoms on a given day. We assume this probability remains constant during a week. Then, the number of days D a patient experienced symptoms during a week follows a binomial distribution with probability mass function: $P(D = d) = \mathcal{B}(d|7, p)$ for $d \in \{0, \dots, 7\}$. We define the distribution $\mathcal{B}_{\text{Day}}(y|p)$ for $y \in \{0, 1, 2, 3, 4\}$, which

corresponds to the measurement process of POEM symptoms:

$$\mathcal{B}_{\text{Day}}(y|p) = \begin{cases} \mathcal{B}(0|7, p) & \text{if } y = 0 \\ \mathcal{B}(1|7, p) + \mathcal{B}(2|7, p) & \text{if } y = 1 \\ \mathcal{B}(3|7, p) + \mathcal{B}(4|7, p) & \text{if } y = 2 \\ \mathcal{B}(5|7, p) + \mathcal{B}(6|7, p) & \text{if } y = 3 \\ \mathcal{B}(7|7, p) & \text{if } y = 4 \end{cases} \quad (\text{E.18})$$

This distribution replaces the binomial distribution in the ‘‘Binomial random walk’’ model, the rest of the model (latent dynamic) being unchanged.

E.1.3 Ordered logistic random walk (v1)

The ordered logistic random walk state-space model is characterised by an ordered logistic measurement distribution (Eq. (2.17)), parametrised by a location parameter (latent score) $\hat{y}^{(k)}(t)$ and cut-offs \mathbf{c} ($c_1 = 0^2 < c_2 < \dots < c_M$):

$$y^{(k)}(t) \sim \text{OrderedLogistic}(\hat{y}^{(k)}(t), \mathbf{c}) \quad (\text{E.19})$$

The scale \hat{y} depends on \mathbf{c} , and is approximately $c_M - c_1 = c_M$, so we choose to express priors and the latent dynamic with $\tilde{y}^{(k)}(t) = \hat{y}^{(k)}(t)c_M$, a normalised version of \hat{y} ($g(x) = \frac{1}{c_M}x$), following a latent random walk (Eq. (7.4)).

Rather than setting priors on \mathbf{c} , we parametrise the model with δ , corresponding to the differences between consecutive cutpoints, i.e. $\delta_i = c_{i+1} - c_i$ for $i \in [1, M - 1]$, and assume the priors:

$$\delta \sim \mathcal{N}^+(0, (2\pi/\sqrt{3})^2) \quad (\text{E.20})$$

$$\mu_0 \sim \mathcal{N}(0.5, 0.25^2) \quad (\text{E.21})$$

$$\sigma_0 \sim \mathcal{N}^+(0, 0.125^2) \quad (\text{E.22})$$

$$\sigma \sim \mathcal{N}^+(0, 0.1^2) \quad (\text{E.23})$$

The priors are chosen to be weakly informative and translate to reasonable prior predictive

²We set $c_1 = 0$ without loss of generality since the ordered logistic distribution model is invariant by translation (adding λ to \hat{y} and \mathbf{c} does not change the distribution).

distributions:

- The prior on δ translates to values less than the width of the standard logistic distribution³, allowing the possibility of $P(y^{(k)}(t) = i) \approx 1$.
- The priors on μ_0 and σ_0 translate to an approximately uniform (within the range of \hat{y}) prior for the initial latent score $\hat{y}^{(k)}(t_0)$.
- The prior on σ assumes it is possible to go from a state where $y = 0$ is the most likely outcome, to a state where $y = M$ is the most likely outcome, in two transitions.

E.1.4 Ordered logistic random walk (v2)

When developing models for POEM, we proposed a different parametrisation of the ordered logistic distribution compared to the one above. As noted above, the cutpoints are unknown, which implies that the range of the latent score ($c_M - c_1 = c_M$) varies depending on the measurement error of the symptom. As a result, it was difficult to interpret the latent score, and compare and quantify the amount of measurement error for different severity items. The distribution was also parametrised by the difference δ between two consecutive cutpoints, which implies that the variance of the marginal distribution of cutpoints increased with the cutpoint index, even though there is no reason to expect this a priori.

We propose to parametrise the ordered logistic distribution using a logistic distribution with an unknown standard deviation σ (which is related to the scale s of the logistic distribution by $\sigma = \frac{\pi}{\sqrt{3}}s$) and unknown cutpoints but for which the range (difference between first and last cutpoint) is fixed (cf. scale invariance as noted in Section 2.4). For $y \in \{0, \dots, M\}$ ($M + 1$ categories):

$$\text{OrderedLogistic}(y|\eta, \sigma, \boldsymbol{\delta}) = \begin{cases} 1 - \text{logit}^{-1}\left(\frac{\eta - c_1}{s}\right) & \text{if } y = 0 \\ \text{logit}^{-1}\left(\frac{\eta - c_y}{s}\right) - \text{logit}^{-1}\left(\frac{\eta - c_{y+1}}{s}\right) & \text{if } 0 < y < M \\ \text{logit}^{-1}\left(\frac{\eta - c_M}{s}\right) & \text{if } y = M \end{cases} \quad (\text{E.24})$$

Where:

- $\boldsymbol{\delta}$ is a simplex⁴ vector of size $M - 1$.
- \boldsymbol{c} is a vector of size M such as $c_1 = 0.5$ and $c_{i+1} = c_i + (M - 1)\delta_i$ for $i > 0$. This implies $c_M = M - 0.5$.

Defining \boldsymbol{c} as above implies that the range of the latent score is approximately the same as

³The standard deviation of the standard logistic distribution is $\pi/\sqrt{3}$.

⁴i.e. $\forall i, \delta_i > 0$ and $\sum_{i=1}^{M-1} \delta_i = 1$.

the range of the score $[0, M]$, and that if cutpoints are equally spaced, the expected value of the distribution when $\eta \in \{1, \dots, M - 1\}$ is equal to η . In addition, we assumed a symmetric Dirichlet prior on δ , which specifies a joint prior on c to regularise the cutpoints, making the model scale well to a high number of categories:

$$\delta \sim \text{Dirichlet}(\mathbf{2}) \quad (\text{E.25})$$

And we assumed a lognormal prior on σ , which translates to a 95% CI that is approximately $[0.02M, 0.40M]$, thus allowing very precise or very imprecise measurement that covers the entire range of the score:

$$\frac{\sigma}{M} \sim \log \mathcal{N}\left(-\log(10), (0.5 \log(4))^2\right) \quad (\text{E.26})$$

In the context of the ordered logistic random walk model, we now refer to the standard deviation of the measurement distribution as σ_m and the standard deviation of the latent dynamic as σ_1 . The link function is now the identity, such that $\hat{y} = \tilde{y}$, i.e. \hat{y} follows a Gaussian random walk. For consistency, we change the prior σ_1 so that it follows the same distribution as the prior for σ_m , still allowing fast or slow transitions from a state where $y = 0$ is the most likely outcome to a state where $y = M$ is the most likely outcome. We otherwise keep the same priors for μ_0 and σ_0 .

E.1.5 Multivariate dynamics

For the binomial random walk and ordered logistic random walk models, we assumed that $\tilde{\mathbf{y}}^{(k)}(t)$ followed a Gaussian random walk (Eq. (7.4)). These dynamics are univariate, i.e. the latent dynamics of different symptoms are independent.

For $D > 1$ severity items, we can consider the vector $\tilde{\mathbf{y}}^{(k)}(t)$ containing the transformed latent scores for the D severity items for the k -th patient at time t and generalise the latent dynamic of Eq. (7.4) by making it multivariate, using a multivariate Gaussian random walk with covariance matrix Σ :

$$\tilde{\mathbf{y}}^{(k)}(t+1) \sim \mathcal{N}(\tilde{\mathbf{y}}^{(k)}(t), \Sigma) \quad (\text{E.27})$$

We can express Σ as a function of the correlation matrix Ω and $\text{diag}(\boldsymbol{\sigma})$, the diagonal matrix with values $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_D)^T$ containing the marginal standard deviation of the latent

dynamics of each severity item:

$$\Sigma = \text{diag}(\boldsymbol{\sigma}) \Omega \text{diag}(\boldsymbol{\sigma}) \quad (\text{E.28})$$

We assume a LKJ prior on the correlation matrix Ω ⁵, with shape parameter equal to 1, which corresponds to a “uniform” distribution over correlation matrices:

$$\Omega \sim \text{LKJ}(1) \quad (\text{E.29})$$

We assume independent priors for all $\sigma_i, i \in \{1, \dots, D\}$, where the priors for each σ_i is the same as the prior given in the univariate models.

We also generalise the distribution of the initial condition $\tilde{\mathbf{y}}^{(k)}(t_0)$ (Eq. (7.5)) by making it multivariate:

$$\tilde{\mathbf{y}}^{(k)}(t_0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0 = \text{diag}(\boldsymbol{\sigma}_0) \Omega_0 \text{diag}(\boldsymbol{\sigma}_0)) \quad (\text{E.30})$$

With:

$$\Omega_0 \sim \text{LKJ}(1) \quad (\text{E.31})$$

And each element of $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0$ following independent priors that are the same as the priors for the univariate models.

E.2 Performance metrics

This study used probabilistic models whose predictions are distributions. Severity items are considered to be discrete with their predictions described by probability mass functions, and PO-(o)SCORAD is treated as a continuous variable⁶ with its predictions described by probability density functions.

E.2.1 Accuracy

In addition to the log predictive density introduced in Section 2.2.5, we compute an accuracy metric to facilitate the interpretation of the performance of PO-(o)SCORAD predictions. Since the accuracy is not a proper scoring rule, it is used only for model interpretation but not for

⁵For computational efficiency, the model is parametrised by the Cholesky factor decomposition L of the correlation matrix $\Omega = LL^T$, where L is a lower triangular matrix.

⁶Otherwise, PO-SCORAD is a discrete ordinal variable with 1031 categories.

model selection.

We define the accuracy Acc_n^d for a single data-point y_n and a fixed threshold d by the probability that the absolute error $|\epsilon_n| = |\hat{y}_n - y_n|$ is less than d , where \hat{y}_n denotes the random variable whose distribution is the predictive distribution with density $p(y)$,

$$\begin{aligned} Acc_n^d &= P(|\hat{y}_n - y_n| < d) \\ &= \int_{y_n-d}^{y_n+d} p(y) dy. \end{aligned} \tag{E.32}$$

Here, the threshold d represents a maximum acceptable error, and should ideally be a Minimal Important Difference (MID) or a Minimal Detectable Change (MDC). In this study, we arbitrarily chose $d = 5$, which is smaller than published estimates of the MID of 8.7 for SCORAD and 8.2 for oSCORAD [102]. We preferred to take a smaller value than the published estimate of MID, especially because the uncertainty around these estimates were not reported in [102], and because of the limitations of the MID described in Section 2.1.4.

The accuracy values of single data points are then averaged to produce the accuracy for a set of predictions. The accuracy is defined in $[0, 1]$. However, it should be noted that the maximum accuracy could be less than 100% if $d < MDC$. Conversely, if a 100% accuracy is achieved for a certain value \tilde{d} , it suggests that $MDC \leq \tilde{d}$.

E.2.2 Learning curves

To address the potential confounding of the average lpd (resp. the accuracy) by patient IDs and prediction horizon (cf. Section 2.3.2), we propose a meta-model to estimate the mean lpd (resp. the mean accuracy) of the test dataset. We fitted a model to explain the lpd as a function of the number of observations in the training set, the prediction horizon and the patient IDs, and used the mean fit as the lpd (resp. accuracy) estimate. We used a Generative Additive Model (GAM) with cubic splines to achieve a flexible fit to the evolution of the lpd, while limiting overfitting. The model was fitted using the `gamm4` package in R, with the formula:

$$lpd \sim s(N(i)) + t + (1 | Patient), \tag{E.33}$$

Where:

- $s(N(i))$ corresponds to a cubic spline on the number of observations $N(i)$ in the training set, at the i -th iteration. $N(i)$ is proportional to i except for late iterations, when a significant fraction of patients have dropped out of the study in Dataset 1.

- t corresponds to the prediction horizon. For simplicity, we assume that the decrease in performance is linear and does not interact with i .
- $(1 | Patient)$ represents a mixed effect on the intercept for different patients.

E.3 Reference models

E.3.1 Markov chain model

When the number of discrete categories is small (e.g. for intensity signs), a Markov chain model may be more appropriate than a random walk model to provide a somewhat flat forecast.

This model assumes that the evolution of y is described by a Markov chain with $M + 1$ states and $P(y(t+1) = j | y(t) = i) = p_{i,j}$. More generally, $P(y(t+h) = j | y(t) = i) = (T^h)_{i,j}$, for a h -steps-ahead transition, where T is the transition matrix, $(T)_{i,j} = p_{i,j}$. As a baseline model, the transition probabilities, $p_{i,j}$, are assumed to be patient- and time-independent. For the vector, $\mathbf{p}_i = T_{i,\cdot}$, representing the transition probability distribution from state i , we assume an uninformative uniform prior over \mathbf{p}_i using a symmetric Dirichlet distribution,

$$\mathbf{p}_i \sim \text{Dirichlet}(\mathbf{1}). \tag{E.34}$$

E.3.2 Priors for the other reference models

The likelihood of the other reference models is given in Section 2.3.3.

Random walk model

$$\sigma \sim \mathcal{N}^+(0, (0.1M)^2). \tag{E.35}$$

The scale of the prior was set to be 10% of the range, M , of the score. That is, we expect σ to be approximately at most $0.2M$, which further translates in a width of the 95% prediction interval to be $0.8M$ (i.e. almost the range of the score, considering the approximations).

Exponential smoothing model

We used the same prior for σ as the random walk model, and assume a log-normal prior for the time constant τ , which spans several orders of magnitude between 10^{-1} and 10^2 days:

$$\tau \sim \log \mathcal{N}\left(0.5 \log(10), (0.75 \log(10))^2\right) \quad (\text{E.36})$$

Autoregressive model

We used the same prior for σ as the random walk model, and assume the following priors for the autocorrelation coefficient α and the expected value of the series y_∞ :

$$\alpha \sim \mathcal{U}(0, 1) \quad (\text{E.37})$$

$$y_\infty \sim \mathcal{N}(0.5M, (0.25M)^2) \quad (\text{E.38})$$

The prior for α is uniform and the prior on y_∞ covers the range of the score.

Mixed autoregressive model

We used the same prior for σ as the random walk model, and assume hierarchical priors for the patient-dependent parameters $\alpha^{(k)}$ and $y_\infty^{(k)}$:

$$\alpha^{(k)} \sim \text{logit } \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad (\text{E.39})$$

$$\mu_\alpha \sim \mathcal{N}(0, 1) \quad (\text{E.40})$$

$$\sigma_\alpha \sim \mathcal{N}^+(0, 1.5^2) \quad (\text{E.41})$$

$$y_\infty^{(k)} \sim \mathcal{N}(\mu_\infty, \sigma_\infty^2) \quad (\text{E.42})$$

$$\mu_\infty \sim \mathcal{N}(0.5M, (0.25M)^2) \quad (\text{E.43})$$

$$\sigma_\infty \sim \mathcal{N}^+(0, (0.125M)^2) \quad (\text{E.44})$$

The priors are chosen to be weakly informative and translate to reasonable prior predictive distributions:

- The priors on μ_α and σ_α have reasonable ranges and translate to a marginal prior for $\alpha^{(k)}$ that is approximately uniform.
- The prior for μ_∞ spans the range $[0, M]$ in which y is defined.
- The prior for σ_∞ implies that the range of the distribution of y_∞ is at most M .

E.4 Supplementary PO-SCORAD figures

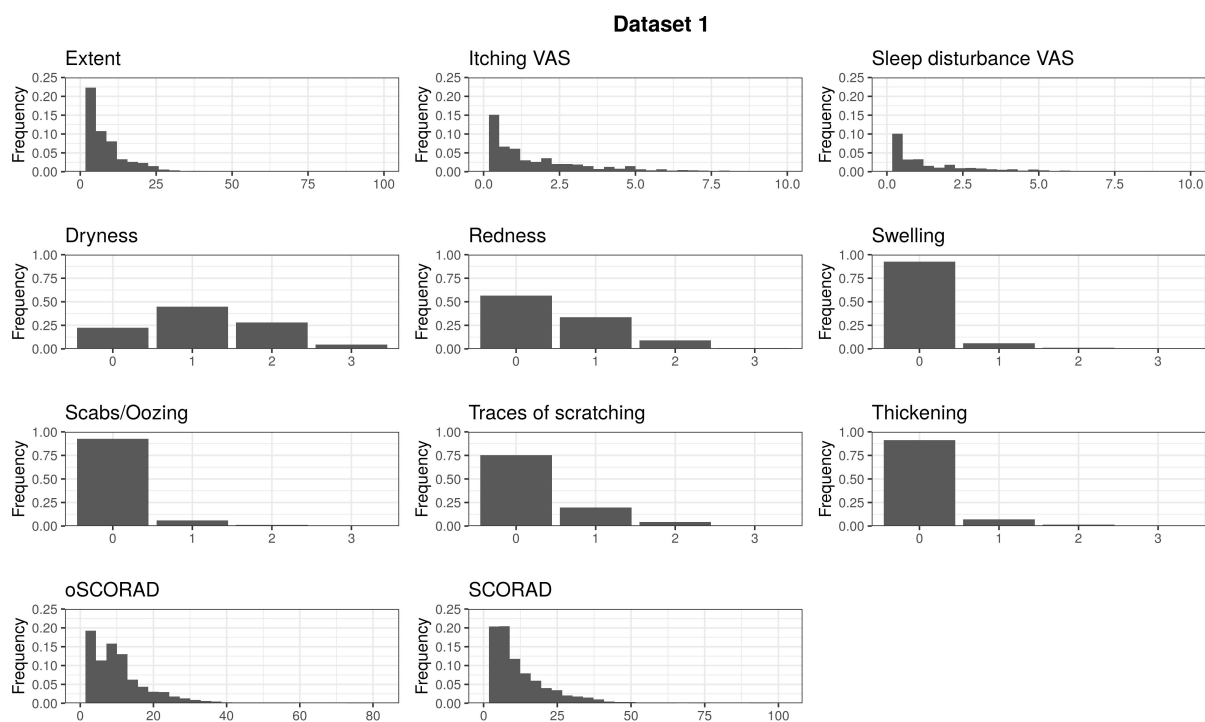


Figure E.2: Distribution of the nine severity items and PO-(o)SCORAD in dataset 1.

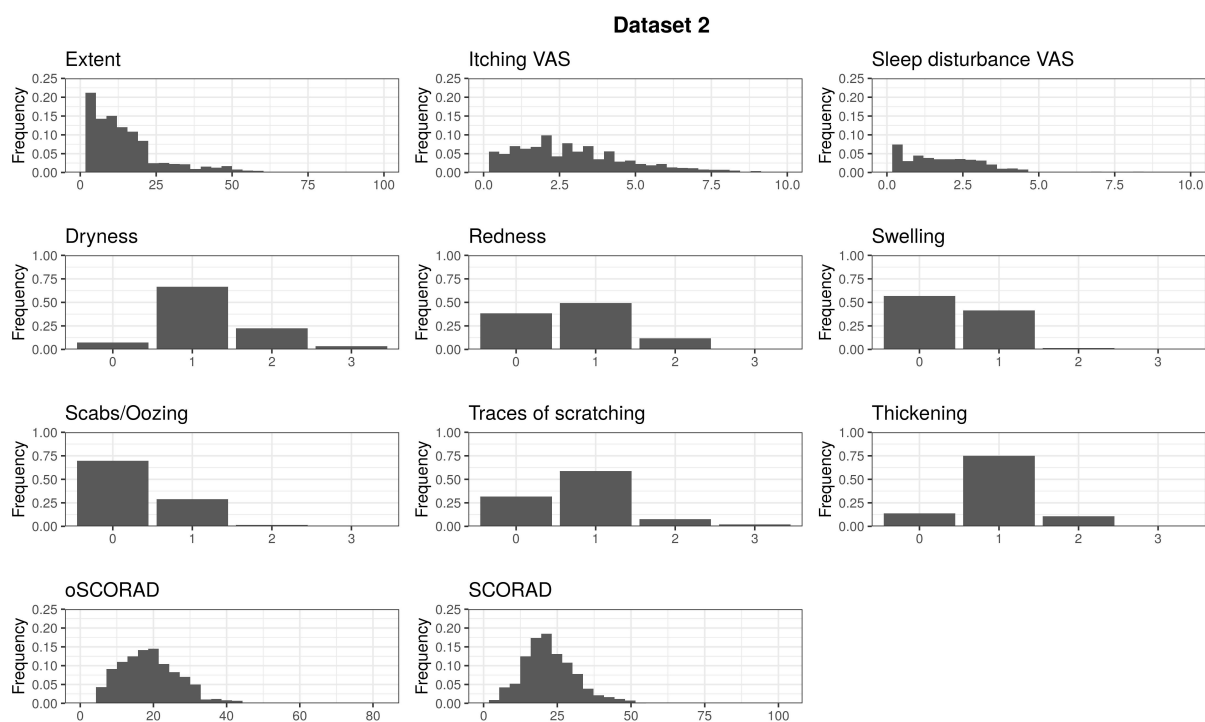


Figure E.3: Distribution of the nine severity items and PO-(o)SCORAD in dataset 2.

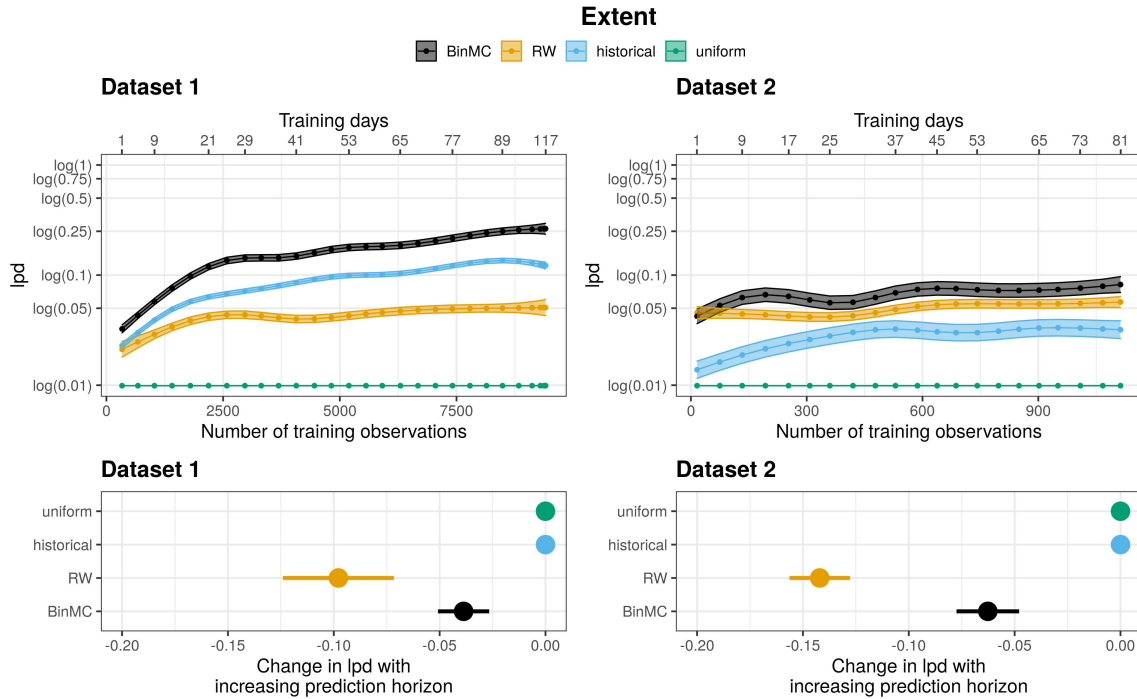


Figure E.4: Predictive performance of the Extent model with datasets 1 (left) and 2 (right), evaluated by lpd (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

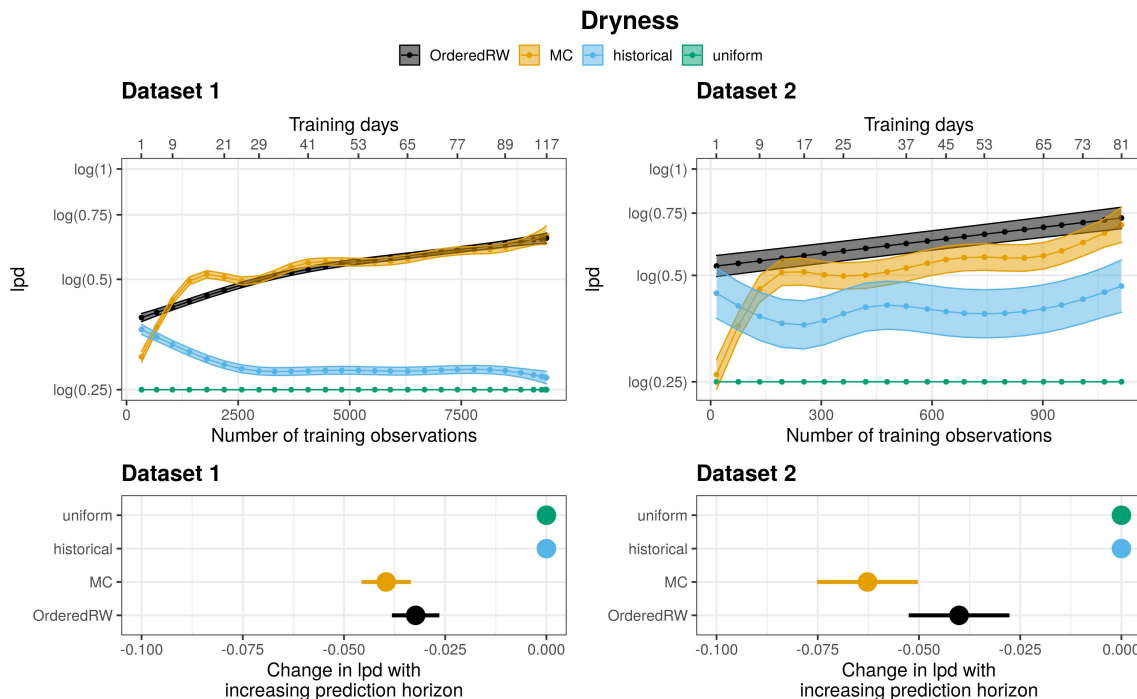


Figure E.5: Predictive performance of the Dryness model with datasets 1 (left) and 2 (right), evaluated by lpd (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

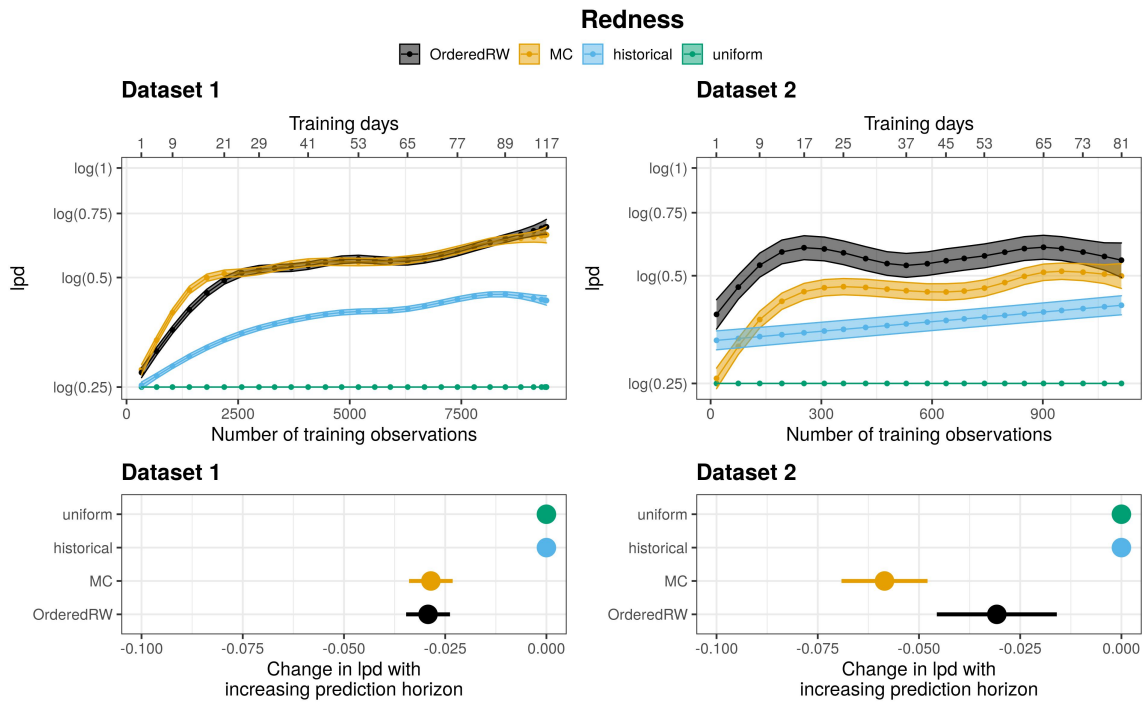


Figure E.6: Predictive performance of the Redness model with datasets 1 (left) and 2 (right), evaluated by l_{pd} (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in l_{pd} as the prediction horizon is increased by a day.

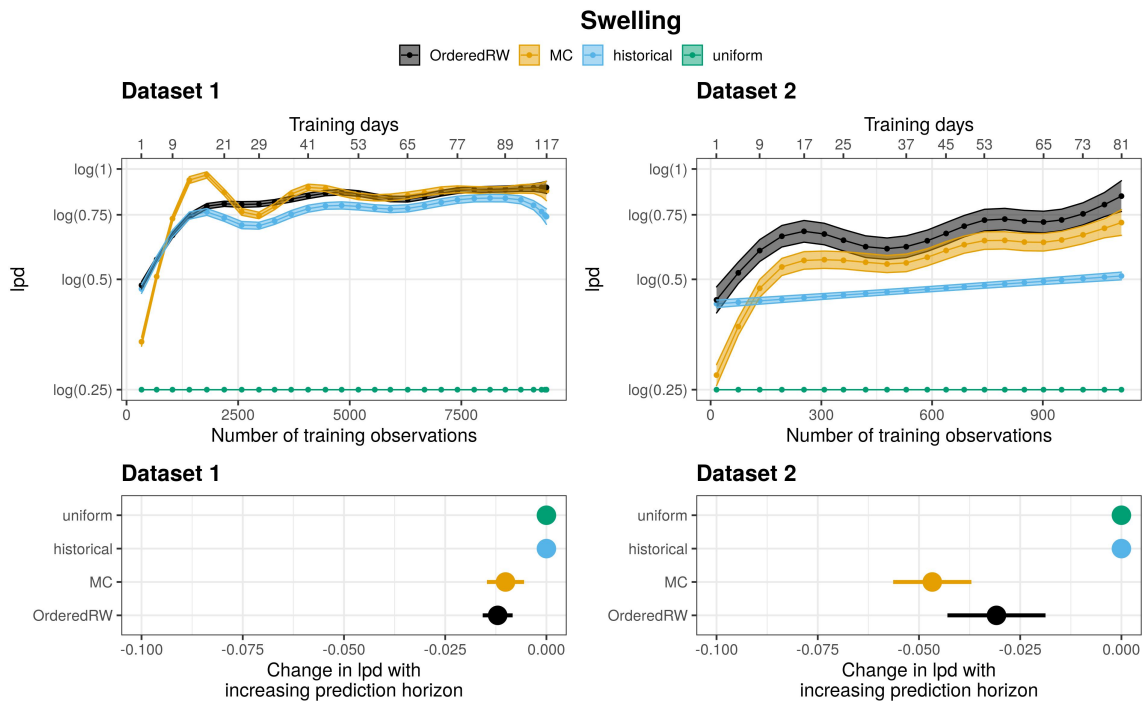


Figure E.7: Predictive performance of the Swelling model with datasets 1 (left) and 2 (right), evaluated by l_{pd} (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in l_{pd} as the prediction horizon is increased by a day.

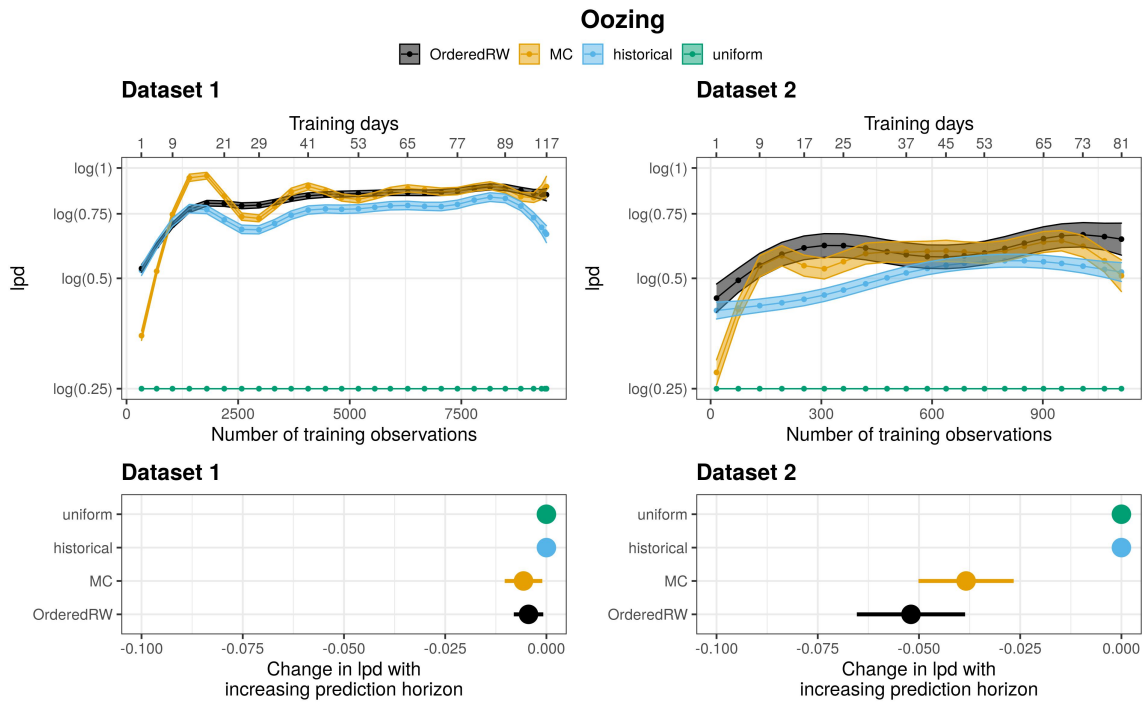


Figure E.8: Predictive performance of the Oozing model with datasets 1 (left) and 2 (right), evaluated by $\text{lpd} (\pm \text{SE})$, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

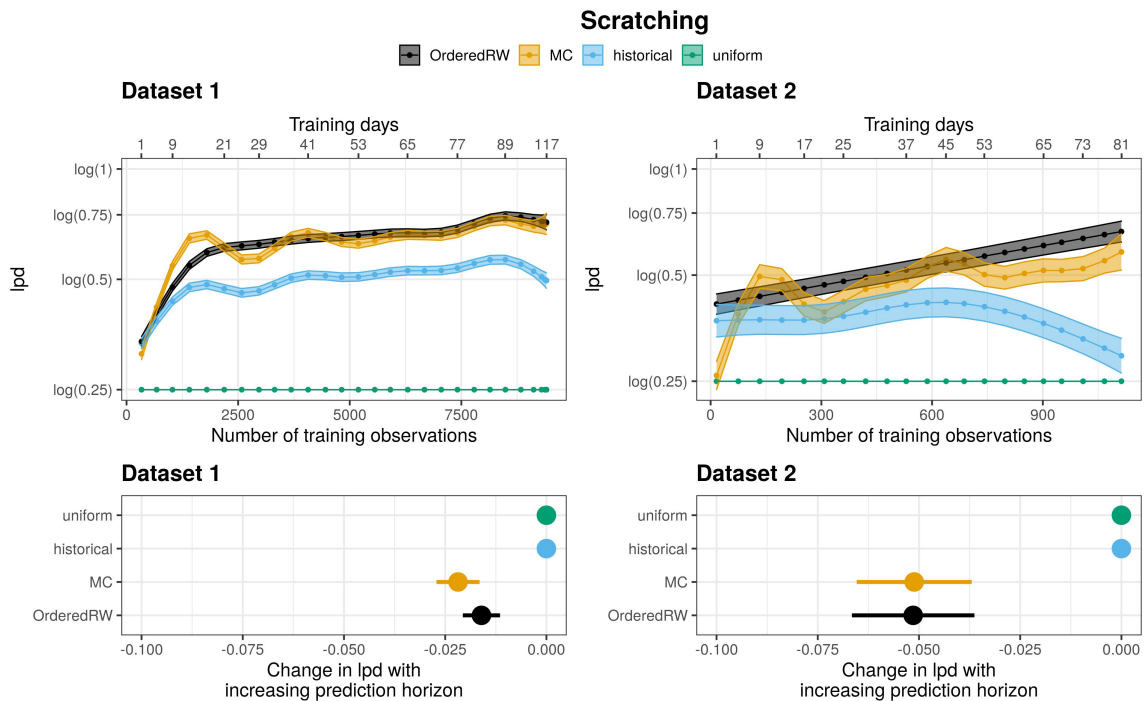


Figure E.9: Predictive performance of the Scratching model with datasets 1 (left) and 2 (right), evaluated by $\text{lpd} (\pm \text{SE})$, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

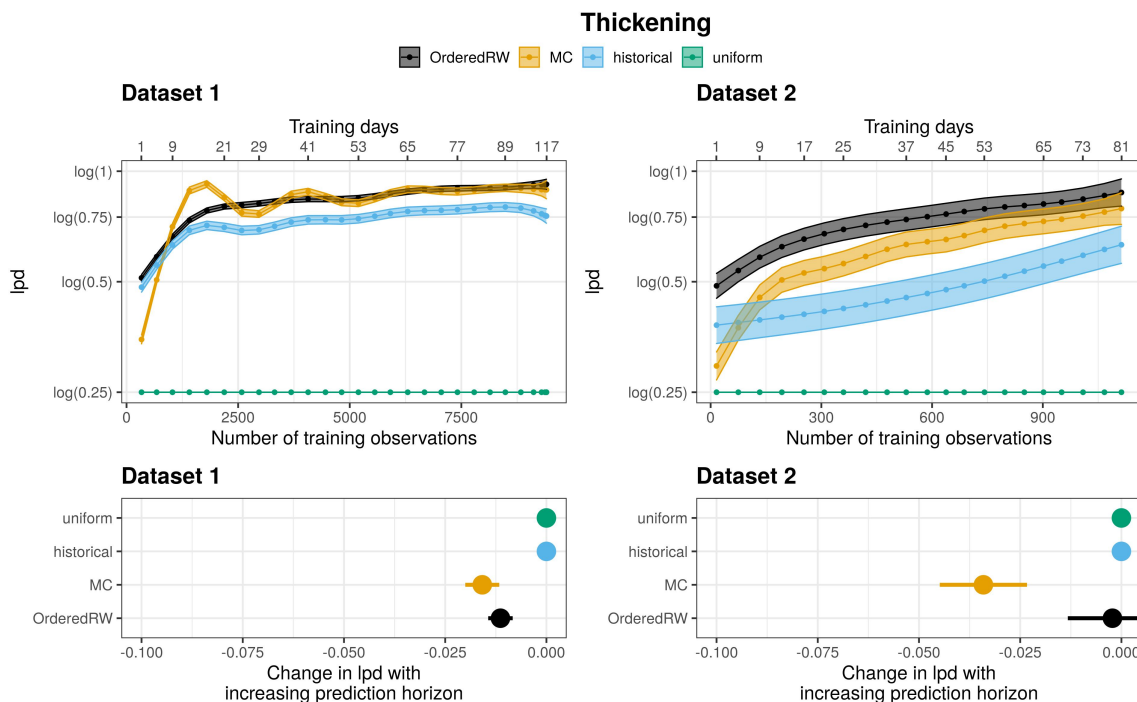


Figure E.10: Predictive performance of the Thickening model with datasets 1 (left) and 2 (right), evaluated by lpd (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

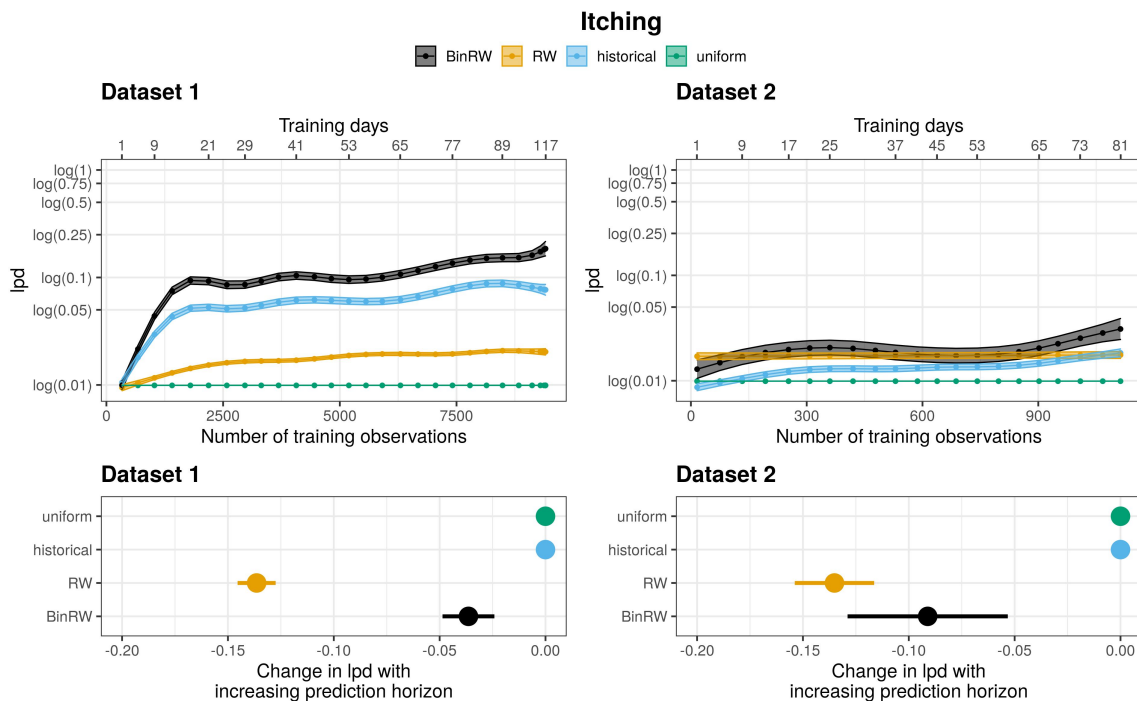


Figure E.11: Predictive performance of the Itching model with datasets 1 (left) and 2 (right), evaluated by lpd (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

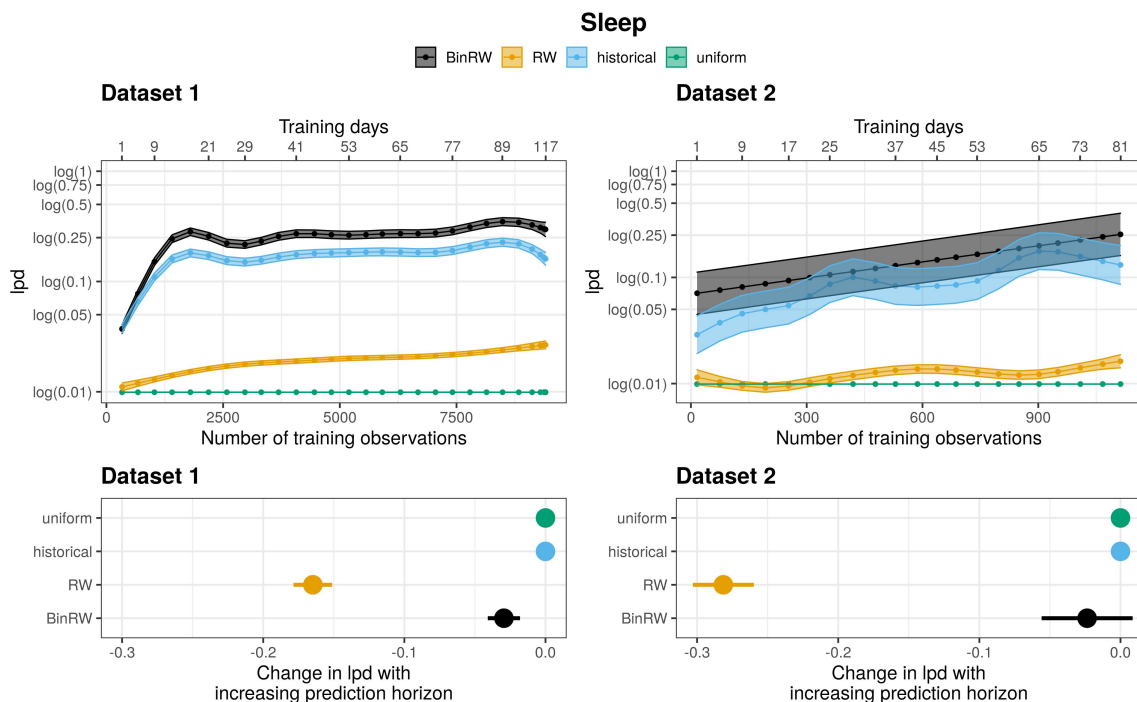


Figure E.12: Predictive performance of the Sleep loss model with datasets 1 (left) and 2 (right), evaluated by lpd (\pm SE, the higher the better). Top: Learning curves for 4-days-ahead predictions as a function of the number of training days (top axis) and the number of training observations (bottom axis). Bottom: Change in lpd as the prediction horizon is increased by a day.

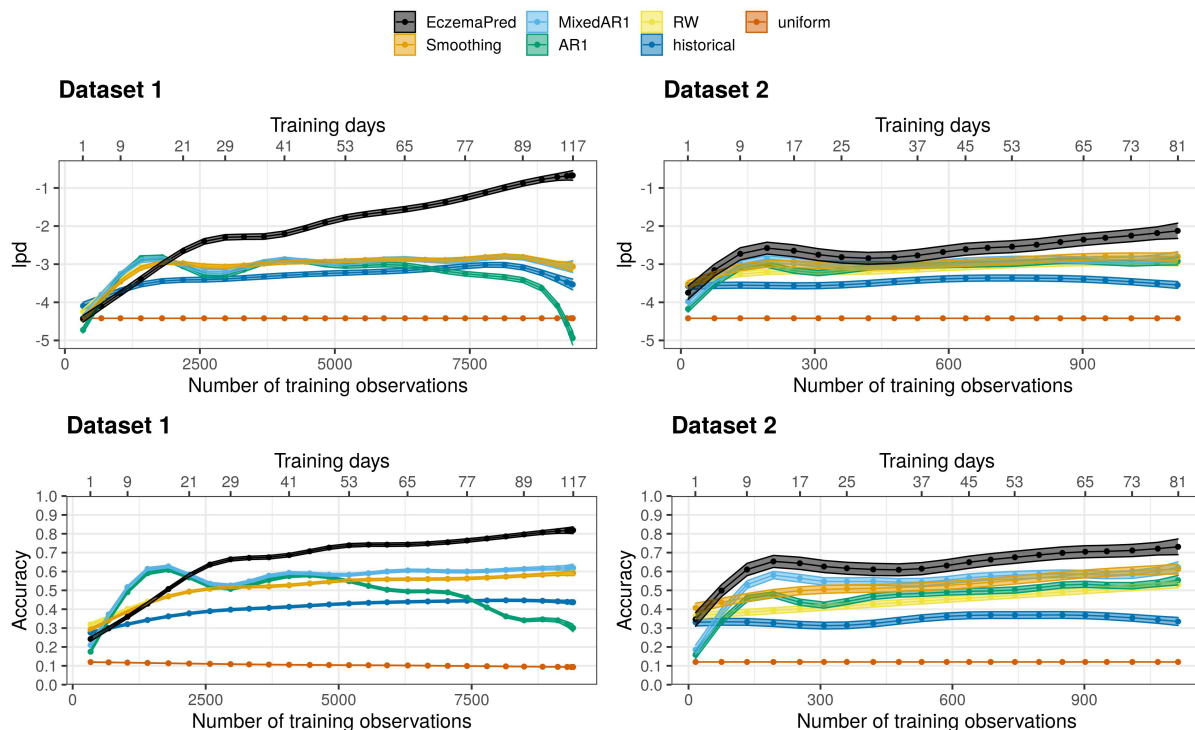


Figure E.13: Learning curves of models predicting PO-oSCORAD, measured by lpd (top) and accuracy (bottom), as a function of the number of training observations (and the number of training days), for datasets 1 (left) and 2 (right). EczemaPred model performs better than the reference models.

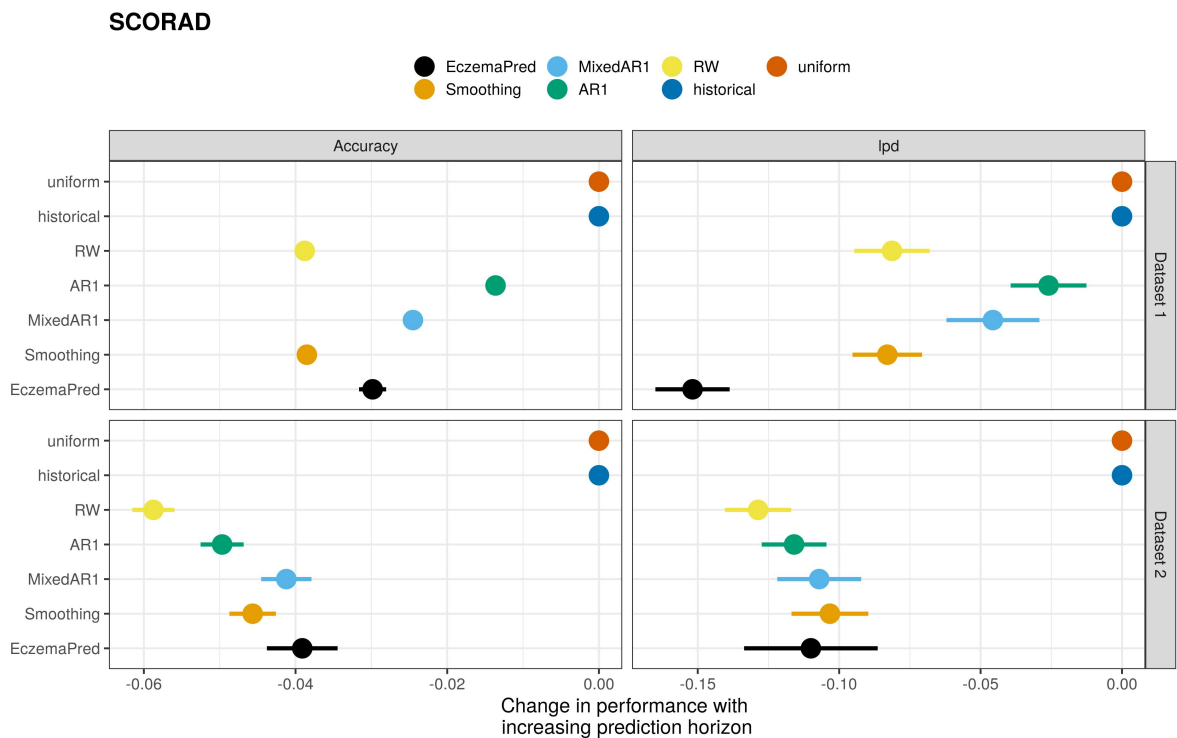


Figure E.14: PO-SCORAD predictive performance changes as the prediction horizon is increased by one day, measured by Accuracy (left) and lpd (right), for datasets 1 (top) and 2 (bottom). The predictive performance for PO-SCORAD decreases with an increase in prediction horizon, for all models.

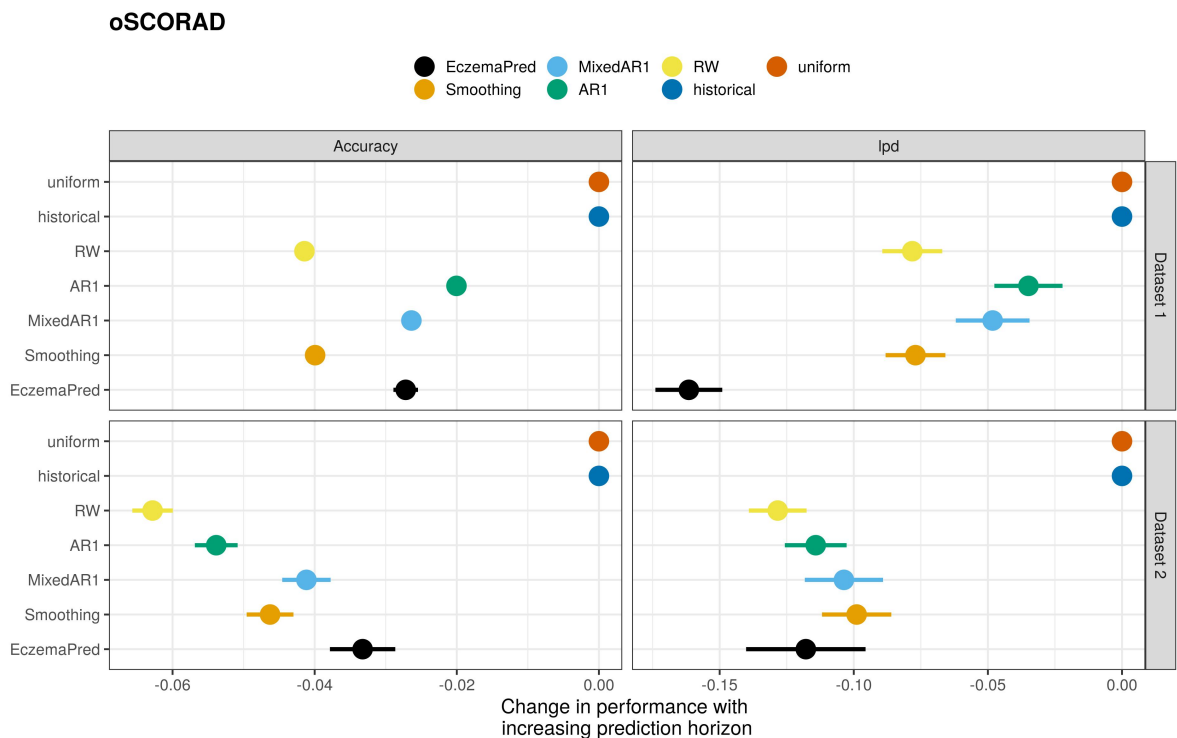


Figure E.15: PO-oSCORAD predictive performance changes as the prediction horizon is increased by one day, measured by Accuracy (left) and lpd (right), for datasets 1 (top) and 2 (bottom).

E.5 Supplementary POEM Figures

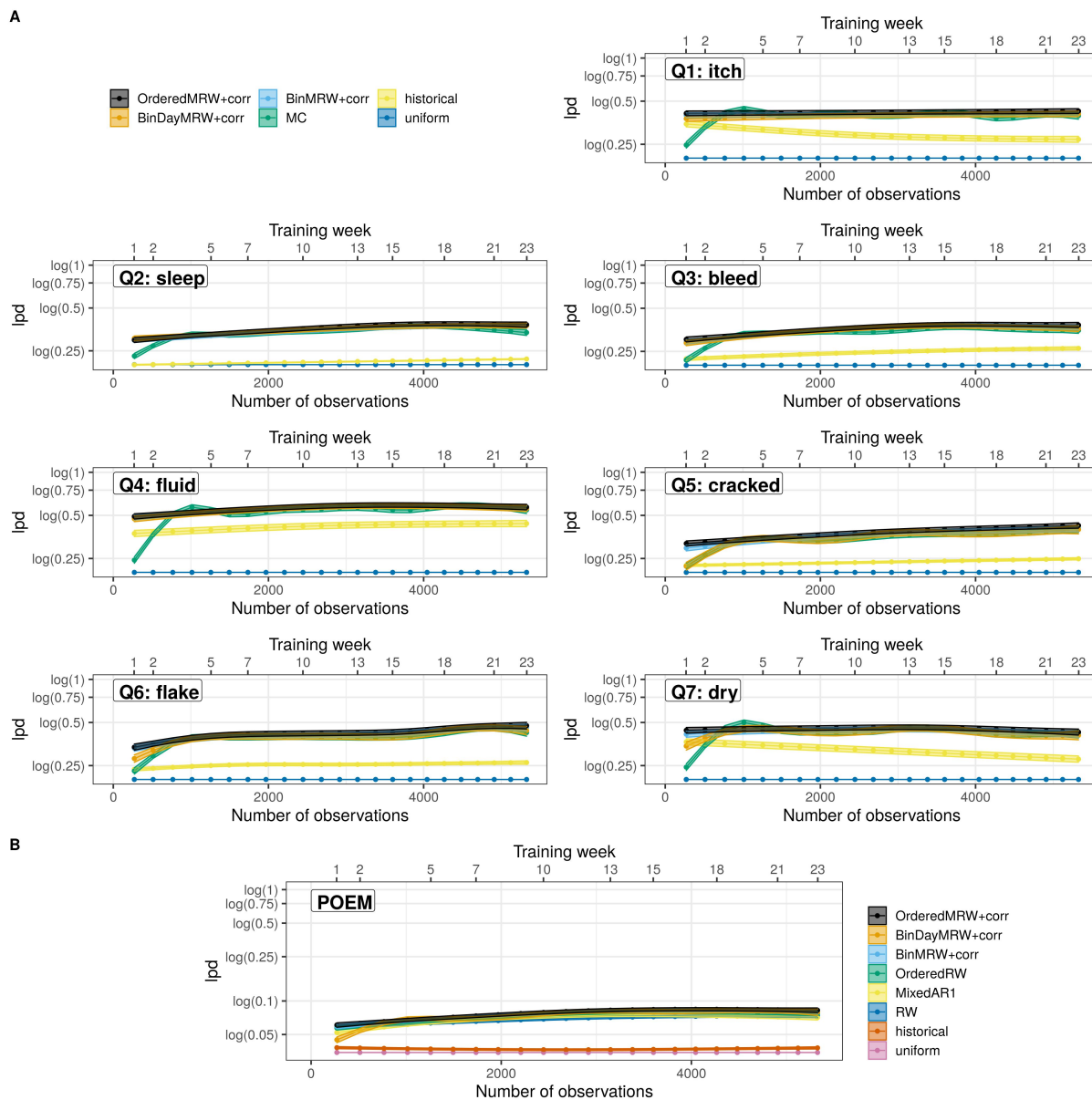


Figure E.16: $lpd (\pm SE)$, the higher the better) learning curves for one-week-ahead forecasts as a function of the number of training observations (bottom x-axis), or equivalently training week (top x-axis). A) Symptom prediction. B) POEM prediction.

Appendix F

Appendix to Chapter 8

F.1 Model

We extended the Bayesian state-space model with an ordered logistic measurement distribution and multivariate latent random walk dynamic developed for POEM prediction in Chapter 7. We call “ScoradPred” the model with independent latent random walks¹.

We define the vector \mathbf{M} such that the i -th severity item takes values in $[0, M_i]$ where $M_i = 3$ for intensity signs and $M_i = 100$ for extent and subjective symptoms (subjective symptoms have a resolution of 0.1 so are scaled to take integer values in $[0, 100]$ and scaled back to $[0, 10]$ for predictions).

We will use the following notation:

- Latent score: $\hat{\mathbf{y}}^{(k)}(t)$ is the vector containing the latent score for each severity item for the k -th patient at time t (dimension 9×1).
- Treatment: $\hat{\mathbf{u}}^{(k)}(t)$ is the vector containing the probability that treatment was used (0 or 1 when inferred deterministically, otherwise unknown), for each treatment, for the k -th patient at time t (dimension 2×1).
- Trend²: $\mathbf{b}^{(k)}(t)$ is the vector containing the trend for each latent score for the k -th patient at time t (dimension 9×1).
- PO-SCORAD measurements: $\mathbf{y}^{(k)}(t)$ is the vector of measured PO-SCORAD items for the k -th patient at time t (dimension 9×1).

¹Which is the same as fitting independent univariate models with an ordered logistic measurement distribution and latent random walk dynamic.

²We deviate from our convention here, as $\mathbf{b}^{(k)}(t)$ are transformed parameters rather than data. This is to be consistent with the time-series forecasting literature.

- SCORAD measurements: $\mathbf{q}^{(k)}(t)$ is the vector of measured SCORAD items for the k -th patient at time t (dimension 9×1).

F.1.1 Latent dynamics

We assume that the evolution of $\hat{\mathbf{y}}$ follows a multivariate normal distribution, defining a vector autoregressive model:

$$\hat{\mathbf{y}}^{(k)}(t) \sim \mathcal{N}(\hat{\mathbf{y}}^{(k)}(t-1) + \mathbf{b}^{(k)}(t-1) + \Theta \hat{\mathbf{u}}^{(k)}(t-1), \Sigma) \quad (\text{F.1})$$

Where:

- $\Sigma = \text{diag}(\boldsymbol{\sigma}_1)\Omega \text{diag}(\boldsymbol{\sigma}_1)$ is the 9×9 covariance matrix of the multivariate normal distribution.
- $\boldsymbol{\sigma}_1$ is the vector of marginal standard deviation (for each severity item) of the multivariate normal distribution.
- Ω is the 9×9 correlation matrix of the multivariate normal distribution.
- Θ is the 9×2 matrix containing the treatment effects of each treatment (columns) for each severity item (rows).

Initial conditions

We also assume that the initial conditions (at $t = t_0$) follow a multivariate normal distribution:

$$\hat{\mathbf{y}}^{(k)}(t_0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \quad (\text{F.2})$$

Where:

- $\boldsymbol{\mu}_0$ is the vector of population mean of the initial latent score for each severity item (dimension 9×1).
- $\Sigma_0 = \text{diag}(\boldsymbol{\sigma}_0)\Omega_0 \text{diag}(\boldsymbol{\sigma}_0)$ is the 9×9 covariance matrix of the initial latent score.
- $\boldsymbol{\sigma}_0$ is the vector of marginal population standard deviation of the latent score for each severity item (dimension 9×1).
- Ω_0 is the 9×9 correlation matrix for the initial latent score.

Trend

We model the trend of the latent score using exponential smoothing [129]:

$$\mathbf{b}^{(k)}(0) = \mathbf{0} \quad (\text{F.3})$$

$$\mathbf{b}^{(k)}(t) = \boldsymbol{\phi} * (\hat{\mathbf{y}}^{(k)}(t) - \hat{\mathbf{y}}^{(k)}(t-1)) + (\mathbf{1} - \boldsymbol{\phi}) * \mathbf{b}^{(k)}(t-1) \quad (\text{F.4})$$

Where:

- $*$ denotes the element-wise multiplication.
- $\boldsymbol{\phi}$ is the vector containing the smoothing trend parameter for each severity item (dimension 9×1). If $\phi_i = 0$ then the trend is constant. If $\phi_i = 1$ then the trend is not smoothed.

It is worth noting that we do not define a double exponential smoothing as the measurement process already acts as if the latent score was a smoothed version of the observations. Moreover, to keep things simple, we do not assume any damping or that the smoothing parameter is patient-dependent.

Daily treatment usage inference

The dataset contains the information of whether treatment “was used within the past two days”, that we deconvolve to obtain daily treatment usage information.

For any of the two treatments (we drop the subscript j indexing treatments out of concision), let:

- u_{W2} be the time-series of “treatment usage within the past two days”
- u the time-series of daily treatment usage
- $\hat{u}(t) = P(u(t) = 1)$ the probability that treatment was used at time t

We can use logic to deterministically infer some values of \hat{u} given the time-series u_{W2} . More specifically, we identify three cases where deterministic inference is possible:

$$u_{W2}(t) = 0 \quad \Rightarrow \quad \hat{u}(t-2) = \hat{u}(t-1) = \hat{u}(t) = 0 \quad (\text{F.5})$$

$$u_{W2}(t) = 0 \quad \& \quad u_{W2}(t+1) = 1 \quad \Rightarrow \quad \hat{u}(t+1) = 1 \quad (\text{F.6})$$

$$u_{W2}(t) = 1 \quad \& \quad u_{W2}(t+1) = 0 \quad \Rightarrow \quad \hat{u}(t-2) = 1 \quad (\text{F.7})$$

The values of $\hat{u}(t)$ that cannot be inferred deterministically are treated as parameters to be inferred by the model, given the likelihood:

$$P(u_{W2}(t) = 1) = 1 - (1 - \hat{u}(t))(1 - \hat{u}(t - 1))(1 - \hat{u}(t - 2)) \quad (\text{F.8})$$

In addition, we assume a patient-dependent Markov Chain likelihood (when known) / hyperprior (when unknown) for $\hat{u}(t)$, in order to reduce the parameter space:

$$\hat{u}(t + 1) = p_{11}^{(k)}\hat{u}(t) + p_{01}^{(k)}(1 - \hat{u}(t)) \quad (\text{F.9})$$

With \hat{u} initialised to the steady state distribution of the Markov Chain:

$$\hat{u}(t_0) = \frac{p_{01}^{(k)}}{p_{01}^{(k)} + p_{10}^{(k)}} \quad (\text{F.10})$$

F.1.2 Measurements

We assume two measurement distributions:

- For PO-SCORAD items:

$$y_i^{(k)}(t) \sim \text{OrderedLogistic}(\hat{y}_i^{(k)}(t), (\sigma_y)_i, \boldsymbol{\delta}_i) \quad (\text{F.11})$$

- For oSCORAD items (except subjective symptoms, as they are the same regardless of whether the score is self-assessed or assessed by a clinician):

$$q_i^{(k)}(t) \sim \text{OrderedLogistic}(\hat{y}_i^{(k)}(t) + \lambda_i(t), (\sigma_q)_i, \boldsymbol{\delta}_i) \quad (\text{F.12})$$

Where:

- $(\sigma_y)_i$ and $(\sigma_q)_i$ are the standard deviation of the logistic distribution.
- $\boldsymbol{\delta}_i$ is a vector of normalised distances between consecutive cutpoints.
- $\lambda_i(t)$ are the measurement biases between SCORAD and PO-SCORAD (see next section).

Calibration

To calibrate PO-SCORAD measurements using SCORAD, we assume:

- SCORAD measurements are more precise than PO-SCORAD measurements (but they

are not perfect)³:

$$\sigma_q = 0.5 \sigma_y \quad (\text{F.13})$$

- The biases $\lambda(t)$ decrease exponentially from $\lambda_i(t_0)$, with a characteristic time τ_i (in days):

$$\lambda_i(t) = \lambda_i(t_0) \exp\left(-\frac{t - t_0}{\tau_i}\right) \quad (\text{F.14})$$

As a rule of thumb, when $t - t_0 = 2\tau_i$, then the bias is equal to 14% of the original bias. Considering that SCORAD measurements occur every four weeks, the last one being at week 12, as a rule of thumb, if $\tau_i < 15$ (days), then the bias at the second measurement is almost null; and if $\tau_i > 200$ (days), we can consider the bias to be almost constant over the duration of the study.

F.2 Priors

F.2.1 Power prior

Background

According to Bayes' theorem, the posterior of a (set of) parameter θ given the data \mathcal{D} is:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta) \quad (\text{F.15})$$

- $p(\mathcal{D}|\theta)$ is the likelihood.
- $p(\theta)$ is the prior.
- $p(\mathcal{D})$ is the evidence (normalisation constant).

If we have previously obtained the posterior distribution $p(\theta|\mathcal{D}_0)$ after observing the historical data \mathcal{D}_0 , we can use this posterior as a new prior:

$$p(\theta|\mathcal{D}, \mathcal{D}_0) \propto p(\mathcal{D}|\theta)p(\theta|\mathcal{D}_0) \propto p(\mathcal{D}|\theta)p(\mathcal{D}_0|\theta)p(\theta) \quad (\text{F.16})$$

³We may wonder why we do not let σ_q be inferred by the model. First, when calibrating measurements, we often assume the calibrated measurements are perfect, i.e. $\sigma_q = 0$. As such, we do not think deciding that $\sigma_q = 0.5 \sigma_y$ is more arbitrary than deciding the measurements are perfect, and is motivated by the fact that we know clinician measurements are not perfect (cf. Section 2.2.5). Second, there is an identifiability issue between σ_q and σ_y , as y and q are mathematically equivalent. Indeed, ignoring the bias term for the sake of argument, when q is observed, the latent score \hat{y} will be determined by y , q , and the prior that the latent score is around the previous latent score (itself informed by previous observations of y). As a result, the latent score will mechanically be closer to y a priori, but specifying the ratio $\frac{\sigma_q}{\sigma_y}$ indicates to the model “how close” the latent scores should be to q as opposed to y .

If we assume that data \mathcal{D} contains N observations, data \mathcal{D}_0 contains N_0 observations, and the observations are independently distributed, then $p(\mathcal{D}|\theta)$ contains N terms and $p(\mathcal{D}_0|\theta)$ contains N_0 terms. Since observations in \mathcal{D} and \mathcal{D}_0 are weighted equally, the relative contribution of data \mathcal{D} to the posterior is $\frac{N}{N+N_0}$.

The power prior [112] introduces a parameter a_0 that weights the contribution of the historical data \mathcal{D}_0 :

$$p(\theta|\mathcal{D}_0) \propto p(\mathcal{D}_0|\theta)^{a_0} p(\theta) \quad (\text{F.17})$$

a_0 quantifies how much information is borrowed from the historical data, with $a_0 = 0$ implying no borrowing and $a_0 = 1$ implying full borrowing. Put it another way, as a rule of thumb, the relative contribution of data \mathcal{D} to the posterior in the presence of the power prior is $\frac{N}{N+a_0N_0}$.

Construction of the power prior

To construct the power prior, we made several assumptions and approximations that we describe in this section.

As historical data, we used “dataset 1” from Chapter 7, i.e. the data from an already published study investigating the role of an emollient in 337 children with AD [189]. The historical data contains 9943 patient-day observations of PO-SCORAD, compared to 1136 patient-day observations in our dataset.

To ensure that the posterior is mostly determined by our dataset rather than the historical dataset, we chose $a_0 = 0.04$ (Fig. 8.3A).

To construct the power prior, first we used the historical data to fit the ScoradPred model, consisting of independent state-space models with ordered logistic measurement distribution and latent random walk dynamic.

Then, we exponentiate the posterior distribution rather than the likelihood to construct the power prior, resulting in initial priors being weighted by $1 + a_0$ instead of 1 (increasing weights results in sharper distribution). Considering that $a_0 \ll 1$ and that the initial priors are weakly informative, this choice has little influence on the posterior, but is more convenient to implement. We also restrict the posterior distribution to the marginal distribution of the population parameters $(\sigma_1, \delta_i, \sigma_y, \mu_0, \sigma_0)$ rather than the full joint distribution. Finally, since distribution is represented by samples (cf. MCMC), we approximated the marginal distributions by Gaussian distributions with moment matching.

We used the same initial priors for $(\sigma_1, \delta_i, \sigma_y, \mu_0, \sigma_0)$ as defined in Chapter 7.

F.2.2 Correlations between severity items

We used an LKJ prior for the correlation matrix Ω of the changes in latent severity items and for the correlation matrix Ω_0 of the initial latent severity items. We selected a shape parameter greater than 1, thus penalising the complexity of the model by putting more density towards the identity correlation matrix (cf. independent severity items):

$$\Omega, \Omega_0 \sim \text{LKJ}(10) \quad (\text{F.18})$$

F.2.3 Trend

For the trend component, we only need to define the prior for the smoothing parameter ϕ and chose a prior that penalised slightly the complexity of the model, i.e. assuming more density toward a constant trend ($\phi = 0$):

$$\phi_i \sim \text{Beta}(1, 3) \quad (\text{F.19})$$

F.2.4 Inference of daily treatment usage

In the following, since we used the same priors for both corticosteroids and emollient cream, we drop the subscript j indexing treatment, out of concision.

We used the same priors for the Markov Chain parameters $p_{10}^{(k)} = 1 - p_{11}^{(k)}$ and $p_{01}^{(k)}$. We defined a hierarchical prior for $p_{10}^{(k)}$ and $p_{01}^{(k)}$, and chose weakly informative priors for the population mean (μ_{10}, μ_{01}) and standard deviation $(\sigma_{10}, \sigma_{01})$ parameters, that translate to approximately uniform distribution for $p_{10}^{(k)}$ and $p_{01}^{(k)}$:

$$p_{10}^{(k)} \sim \text{logit } \mathcal{N}(\mu_{10}, \sigma_{10}^2) \quad (\text{F.20})$$

$$p_{01}^{(k)} \sim \text{logit } \mathcal{N}(\mu_{01}, \sigma_{01}^2) \quad (\text{F.21})$$

$$\mu_{10}, \mu_{01} \sim \mathcal{N}(0, 1) \quad (\text{F.22})$$

$$\sigma_{10}, \sigma_{01} \sim \mathcal{N}^+(0, 1.5) \quad (\text{F.23})$$

F.2.5 Treatment effects

We used the same priors for both treatment and all severity items, assuming that treatment effects are approximately within $\pm 20\%$ the range of the score:

$$\Theta_{i,j}/M_i \sim \mathcal{N}(0, 0.1^2) \quad (\text{F.24})$$

F.2.6 Calibration

We assumed the prior for initial bias $\lambda_i(t_0)$ to be within $\pm 20\%$ the range of the severity item:

$$\lambda_i(t_0)/M_i \sim \mathcal{N}(0, 0.1^2) \quad (\text{F.25})$$

We assumed a lognormal prior for the characteristic learning time τ_i , such as most of its mass is between 1 day and ≈ 250 days, which allows very fast learning (bias null for at the second SCORAD measurement) or no learning (constant bias for the duration of the study):

$$\tau_i \sim \log \mathcal{N}\left(1.2 \log(10), (0.6 \log(10))^2\right) \quad (\text{F.26})$$

F.3 Treatment recommendations

F.3.1 Utility function

We consider two binary actions (using or not using topical corticosteroids TC or emollient cream EC), not mutually exclusive, that we denote with $a = \{TC, EC\} \in \{0, 1\}^2$.

We define the utility function as a function of the predicted SCORAD \hat{y} and the action a by:

$$U(\hat{y}, a) = -(\hat{y} + \text{cost}(a)) \quad (\text{F.27})$$

Where $\text{cost}(a)$ is the “perceived” cost of action a , that we define by:

$$\text{cost}(a) = \text{cost}_{TC} \times TC + \text{cost}_{EC} \times EC + \text{cost}_{both} \times TC \times EC \quad (\text{F.28})$$

- cost_{TC} and cost_{EC} are the costs of using topical corticosteroids and emollient cream, respectively.

- $cost_{both}$ is the additional cost when using the two treatments.

It is worth noting that:

- The cost of using “no treatment” is set to 0 without loss of generality.
- $cost(a)$ is in the same unit as \hat{y} (i.e. SCORAD), and can be interpreted as the minimum improvement in y the patient would require to use treatment.

F.3.2 Objective function

We chose to maximise the following objective function, which includes a risk-sensitive criterion:

$$E(U(\hat{y}, a)) - q\sqrt{V(U(\hat{y}, a))} \quad (\text{F.29})$$

Where q quantifies the tolerance to risk (uncertainty):

- q can be interpreted as a z-score, assuming the utility is normally distributed. For example, if $q = 1.96$, the objective function is the lower bound of the 95% CI of the utility, i.e. the 2.5% quantile.
- $q > 0$ corresponds to a patient that is risk-averse (penalising uncertainty) or pessimistic (maximise the worst case).
- $q < 0$ corresponds to a patient that is risk-seeking (welcoming uncertainty) or optimistic (maximise the best case). $q < 0$ encourages the exploration of new treatments with uncertain effects.
- $q = 0$ corresponds to a patient that is risk neutral.

F.3.3 Decision profiles

In our situation, the decision profile of a patient is fully determined by its sensitivity to risk q and its “perceived” cost of using treatment, parametrised by $cost_{TC}$, $cost_{EC}$ and $cost_{both}$.

We defined three risk profiles (risk-averse, risk-neutral and risk-seeking) and three cost profiles (no cost, normal cost, high cost) for a total of nine decision profiles (Table F.1). The cost profiles assume the same costs for topical corticosteroids and emollient cream.

Even though we believe these decision profiles to be relevant, they are only illustrative and do not claim to represent the preferences of an actual patient. In particular, the decision parameters may be changed to reflect changes in patients’ preferences. For example, q could

be negative at the beginning of a trial to encourage the exploration of new treatments and be increased gradually to exploit a treatment that has been proved to be effective.

Table F.1: Decision profiles for treatment recommendations

Decision profile	Cost profile	Risk Profile	Treatment costs			Risk tolerance
			$cost_{TC}$	$cost_{EC}$	$cost_{both}$	q
1	No cost	Risk averse	0	0	0	1.5
2	No cost	Risk seeking	0	0	0	0
3	No cost	Risk neutral	0	0	0	-1.5
4	Normal cost	Risk averse	0.5	0.5	0	1.5
5	Normal cost	Risk seeking	0.5	0.5	0	0
6	Normal cost	Risk neutral	0.5	0.5	0	-1.5
7	High cost	Risk averse	3	3	3	1.5
8	High cost	Risk seeking	3	3	3	0
9	High cost	Risk neutral	3	3	3	-1.5

F.4 Supplementary figures

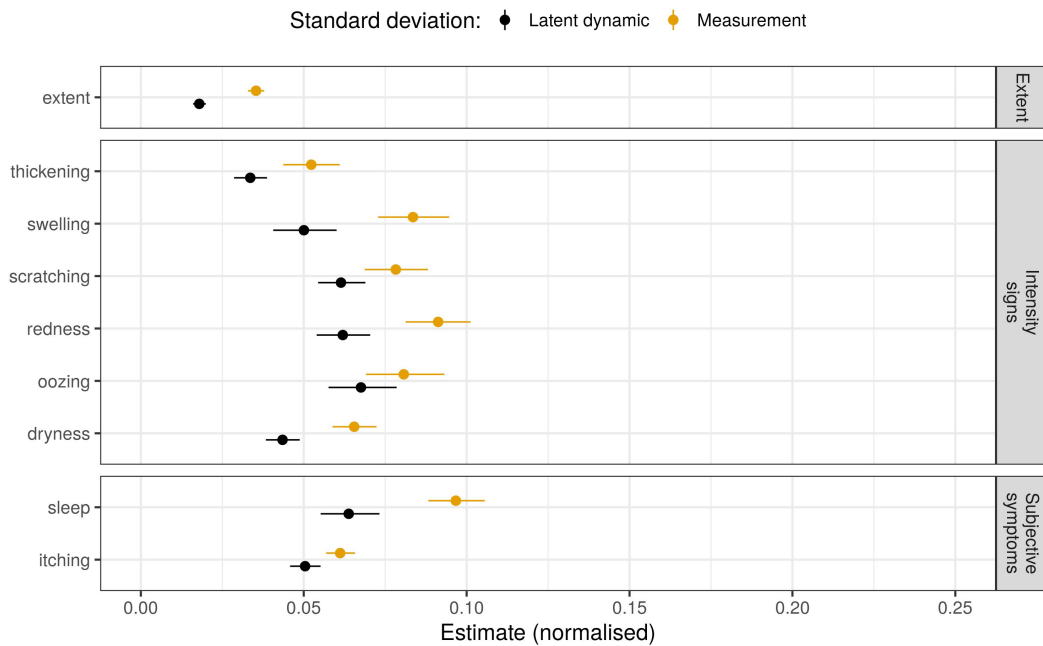


Figure F.1: Estimates of the measurement (σ_y) and latent dynamic (σ_1) standard deviations for all severity items. Estimates are normalised by the range of the score.

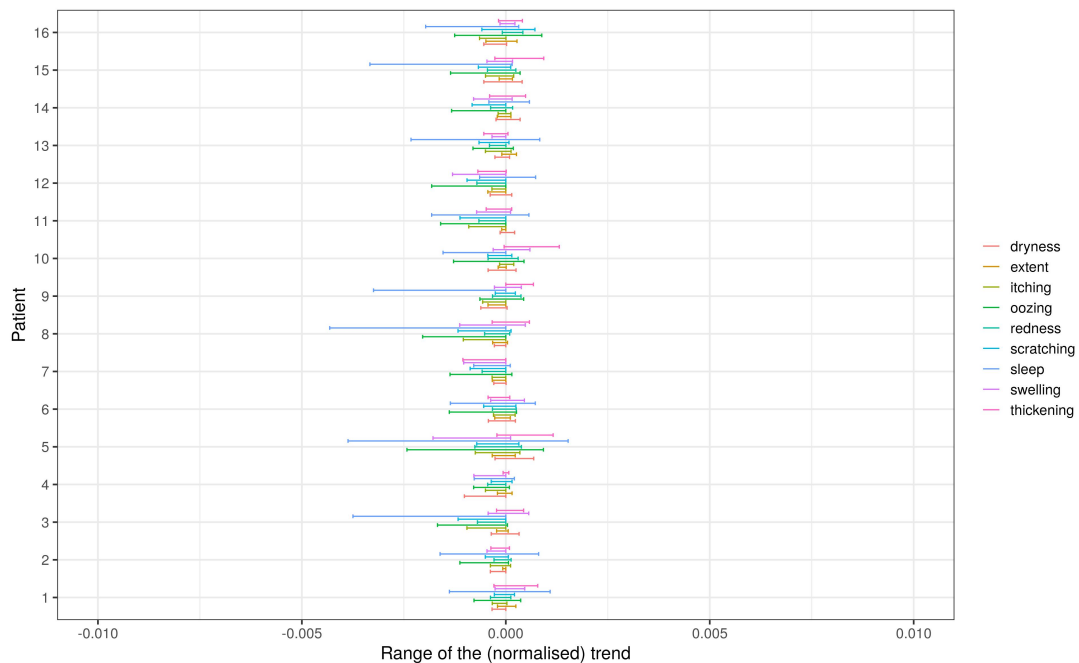


Figure F.2: Minimum and maximum of the expected trend component, for each patient and each severity item. The estimates are normalised by the range of the score. For example, if the maximum is 0.01, for extent (defined in $[0, 100]$) this would mean that the maximum expected trend is $0.01 \times 100 = 1$. We can consider that the trend is always zero as the order of magnitude of the amplitude of the trend is around 0.001.

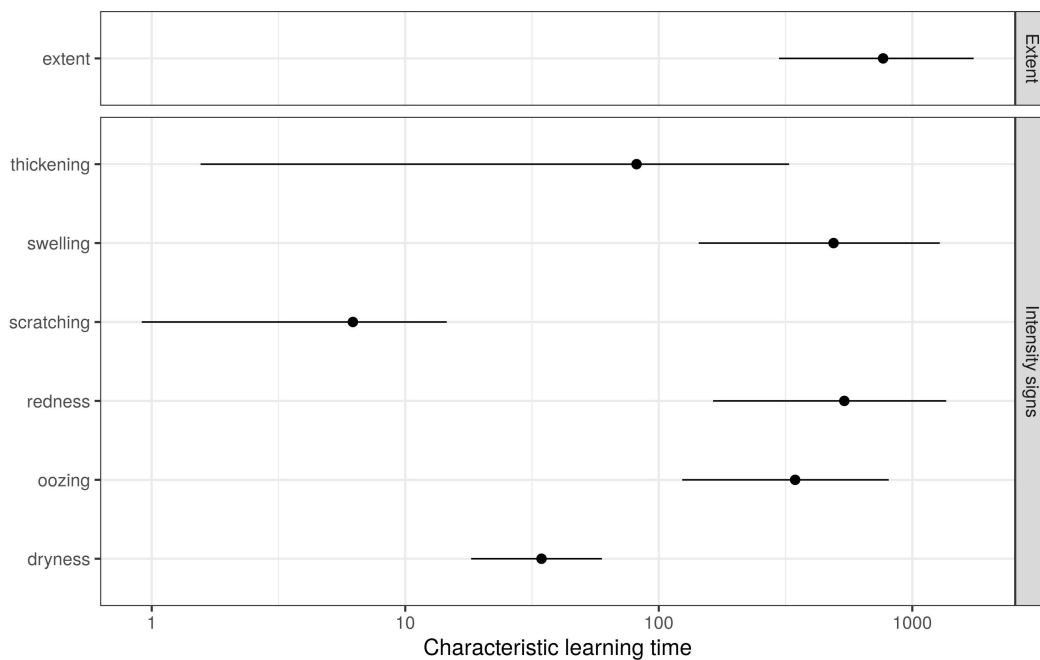


Figure F.3: Mean and 90% credible interval of the characteristic learning time τ of the calibration process (in days). Since SCORAD measurements are every four weeks, estimates less than ≈ 15 days translate to a bias of 0 for the second measurement. Any estimates greater than ≈ 200 days (longer than the study follow-up) can be interpreted as a constant bias (no learning).

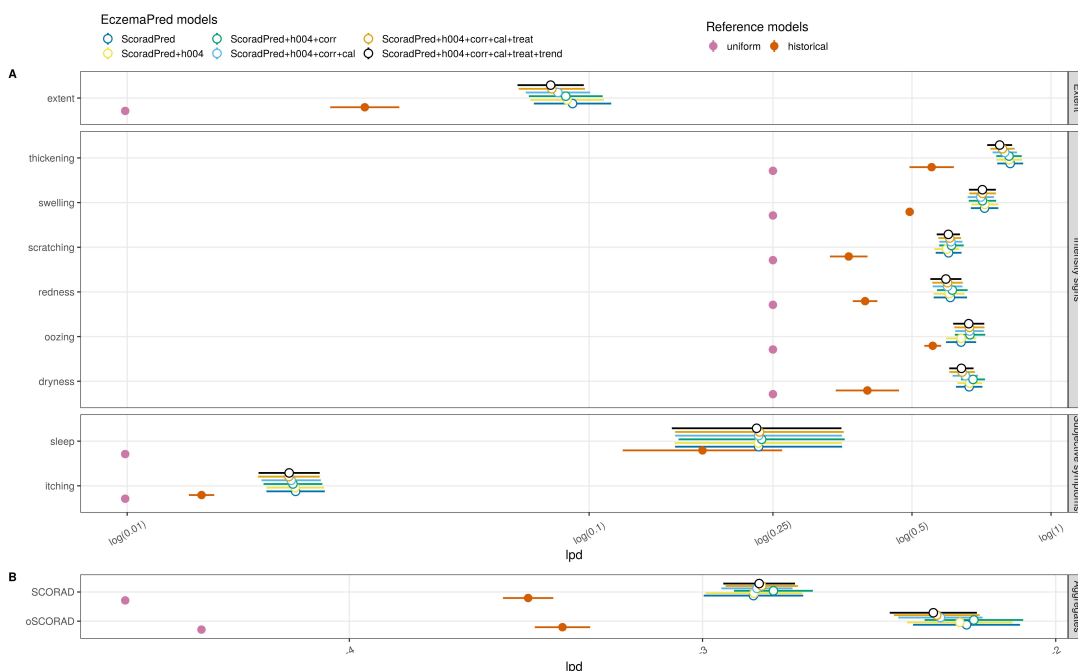


Figure F.4: Predictive performance (lpd) estimates (mean \pm SE) for four-days-ahead predictions after training the model with 65 days (79%) of data. “ScoradPred” corresponds to the base model with independent state-space models for each severity item, no power prior, calibration, treatment effects or trend. The suffixes indicate the additions made to the base model, where “h004” corresponds to the power prior with $a_0 = 0.04$, “corr” to the correlation between severity items, “cal” to the calibration data, “treat” to treatment effects and “trend” to the trend component. A) Severity items predictive performance. B) (o)SCORAD predictive performance.