Predicting Patient Reported Outcome Measures (PROMs) Using Routinely Collected Healthcare Data

Joseph Taggart[1], Helen Parsons[2], Martyn Plummer[1], Rhiannon K. Owen[3], Keith R. Abrams[1,2]

[1]Department of Statistics, University of Warwick; [2]Warwick Medical School, University of Warwick; [3]Population Data Science, Swansea University Medical School, University of Swansea

Background:

- In Health Technology Assessment, quality of life evidence from trials is used to evaluate cost-effectiveness of new treatments by agencies such as National Institute for Health and Care Excellence (NICE).

- Trials are often not inclusive; women, minority ethnic groups, the elderly, and individuals living with multiple long-term conditions are often under-represented [1,2], contributing to health inequalities.

- We could use electronic health records instead to access a broader population, but they typically do not contain the quality of life PROMs that we need.

- If PROMs were routinely available in EHR data (or could be predicted), decision makers would have access to a more representative patient population with which to make decisions.

Objective:

Predict the quality of life PROM CASP-19 (Control, Autonomy, Self-Realisation, Pleasure) from routinely collected variables (sex, ethnicity, body mass index, comorbidities, Index of Multiple Deprivation) using the English Longitudinal Study of Ageing (ELSA), with approximately 10,000 individuals and 31,000 observations.

Methods:

Models were fit and estimated using Markov Chain Monte Carlo (MCMC) using JAGS (Just Another Gibbs Sampler) to allow for full quantification of the uncertainty around parameter estimates and predictions.

Model 1 consisted of a linear mixed-effects model, with overall CASP-19 score as the longitudinal outcome:

$$y_i(t) = \mu_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$$
$$\mu_i(t) = \beta_0 + b_i + \beta_1 t + \boldsymbol{\beta}^T \mathbf{X}_i(t)$$

with $b_i \sim \mathcal{N}(0, \sigma_b^2)$ an individual level random intercept for individual i, and $\mathbf{X}_i(t)$ a vector of covariates for individual i

Model 2 then extended this by dividing the longitudinal component into the four sub-dimensions of CASP-19:

$$y_i(t) = \sum_{j=1}^{4} \mu_{i,j}(t) + \epsilon_i(t), \quad \epsilon_i(t) \sim \mathcal{N}(0, \sigma^2)$$
$$\mu_{ij}(t) = \beta_{0j} + b_{ij} + \beta_{1\mathrm{j}} t + \boldsymbol{\beta}_j^T \mathbf{X}_i(t)$$

with $b_{ij} \sim \mathcal{N}\left(0, \sigma_{b_\mathrm{j}}^2\right)$ an individual level random intercept, for sub-dimension j.

Model 3 models each of the 19 questions simultaneously, using ordered logistic regression. The CASP-19 total score is the sum of the 19 item responses:

$$y_i(t) = \sum_{q=1}^{19} y_{iq}(t)$$
$$p_{cq} = \mathbb{P}(y_{iq}(t) \leq c \mid \mathbf{X}_i(t))$$
$$\mathrm{logit}(p_{cq}) = k_{cq} - \mu_{iq}(t)$$
$$\mu_{iq}(t) = b_{iq} + \beta_{1q} t + \boldsymbol{\beta}_q^T \mathbf{X}_i(t)$$

Funded by NIHR | National Institute for Health and Care Research

Results:

Predictive accuracy was assessed by mean squared error on the predicted CASP-19 total score. Each model was fit with and without the routinely collected covariates.

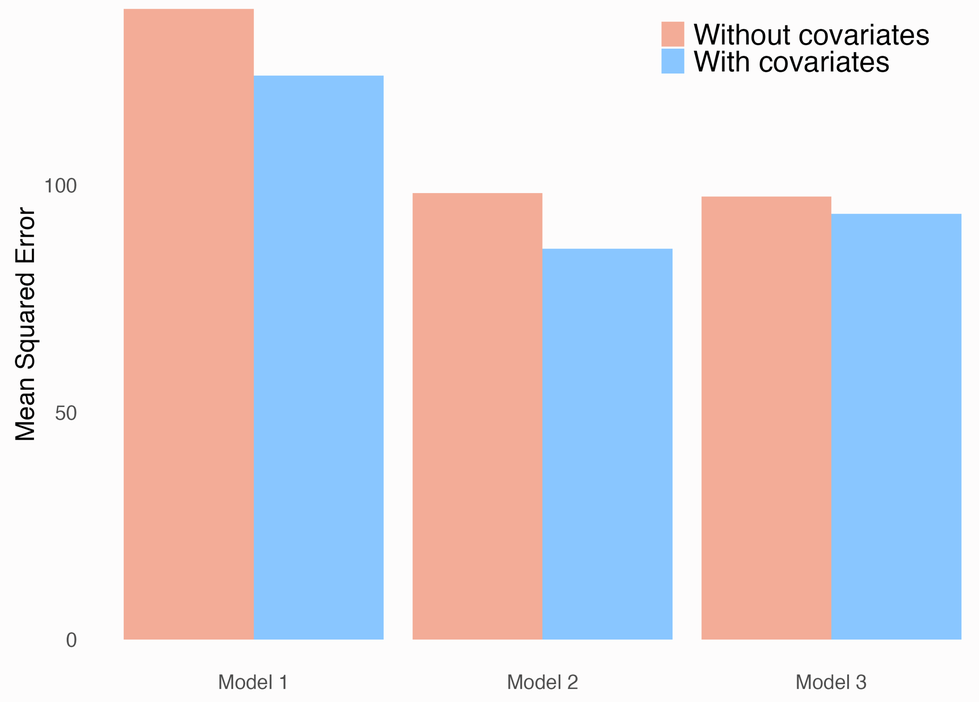

Figure 1: Model comparison by mean squared error.

Adding covariates increases MSE across all models. Model 2 achieved the lowest error (MSE = 86.0, RMSE = 9.27 on the 0-57 CASP-19 scale). The item-level ordered logistic model did not improve on the simpler mixed-effects model, indicating that modelling the ordinal items adds little to total score prediction.

Limitations

The biggest limitation of the Bayesian approach is the computational cost:

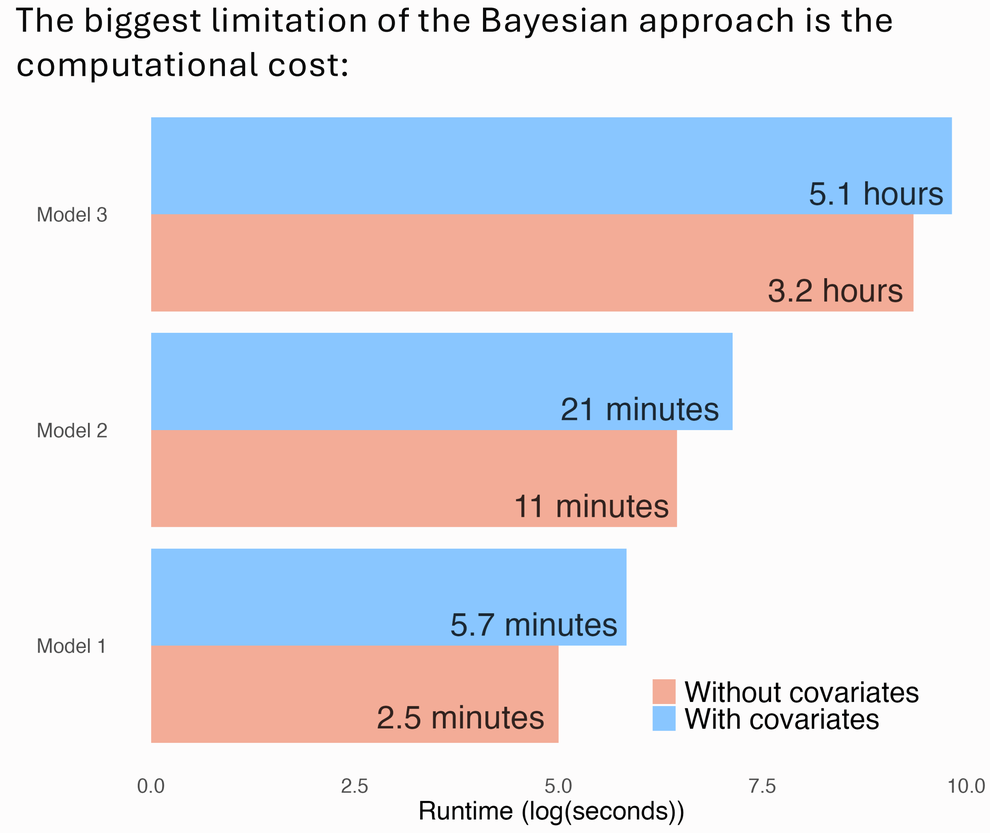

Figure 2: Model comparison by runtime until convergence.

Models incorporating a survival sub-model to account for informative dropout showed potentially increased accuracy but were too computationally intensive and struggled to converge.

Future Work

1. Future application to routinely collected population primary and secondary care data using SAIL Databank (NIHR Doctoral Award application) both for specific diseases (e.g. MS) and general population.

2. Joint survival-longitudinal modelling (to allow for both informative dropout and observations) using INLA or Variational Inference to avoid the high computational cost of MCMC.

[1] Vitale, C., Fini, M., Spoletini, I., Lainscak, M., Seferovic, P. and Rosano, G.M., 2017. Under-representation of elderly and women in clinical trials. *International journal of cardiology*, *232*, pp.216-221.
[2] Redwood, S. and Gill, P.S., 2013. Under-representation of minority ethnic groups in research—call for action. *The British Journal of General Practice*, *63*(612), p.342.