H2020-MSCA-ITN-2018-813545

HELICAL

Health Data Linkage for Clinical Benefit

Deliverable D1.2 Statistical Software Package and Code

This deliverable reflects only the authors' views, and the European Commission Research Executive Agency is not responsible for any use that may be made of the information it contains.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813545

R package dlvarsel for filtering and quantifying the impact of environmental exposures on rare disease episodes.

There has been much recent interest in modelling the impact of environmental exposures on adverse human health and in the context of vasculitis, this is a core aim of Helical WP1. As sensor measurement, scientific and data analytic methods advance to make the study of such phenomena possible, there is a need for rigorous and bespoke modelling strategies with the capacity to address and overcome specific challenges associated with these application domains.

In the context of vasculitis, we identified three key challenges of clinical importance, which lead to scenarios at odds with the usual canon of statistical assumptions. In identifying potential environmental predictors of vasculitis flare episodes based on data comprising hospital encounters we must consider:

(i) **Imbalance:** high disease activity (flare) is seen less frequently than other levels of flare activity in hospital visits, in an already small cohort of vasculitis patients (small *n*)
(ii) **Many potential predictors:** we may know little about which (if any) environmental factors could play a role in triggering or contributing to a flare episode (large *p*)
(iii) **Exposure profile:** if an environmental variable does have predictive value, we know nothing about the cumulative exposure profile which could be useful for predicting future out-of-sample episodes.

The relevance of challenge (iii) to the clinical context is in characterising the *prodrome* period i.e. the period between a trigger and the patient requiring treatment.

Quantifying correlation between flare episode cases and environmental variables is a first step towards new insights about disease mechanism. In doing so, (i), (ii) and (iii) must be appropriately considered and addressed. The open source **dlvarsel** package written for the statistical computing language **R** was developed to meet these challenges. The package is currently available through github at: https://github.com/jsnwyse/dlvarsel.

The package implements a novel Bayesian model which combines ideas from Bayesian quantile regression [1], distributed lag modelling [2] and stochastic model search [3] to address the challenges. In addressing (iii) we explore all possible prodrome shapes through modelling cumulative exposure curves non-parametrically using splines. A key output of the package computations is a Bayesian posterior probability calibrating whether an environmental measurement is correlated with flare episodes. This interpretable output is simple to communicate and useful for determining what might warrant further thought or exploration.

The package has been developed through Helical and in the vasculitis context, however, the core methods implemented in the package can be of use for other rare disease researchers investigating environmental exposures. The package is being further developed to include random effects modelling and modelling over spatial domains.

References

[1] Benoit D, and Van den Poel D. (2017) “BayesQR : A Bayesian Approach to Quantile Regression.” *JOURNAL OF STATISTICAL SOFTWARE*, vol. 76, no. 7, Foundation for Open Access Statistic, 2017, pp. 1–32, doi:10.18637/jss.v076.i07.

[2] Gasparrini A, Armstrong B, Kenward MG. (2010) “Distributed lag non-linear models.” STATISTICS IN MEDICINE, vol. 29, no. 21, 2224-2234, doi:10.1002/sim.3940.

[3] Holmes C, and Held L. (2006) “Bayesian auxiliary variable models for binary and multinomial regression.” BAYESIAN ANALYSIS, vol. 1, 145-168, doi:10.1214/06-BA105.