

Clinical impact of residual C-peptide secretion in Type 1 diabetes on glycemia and microvascular complications

Anita Jeyam 1 , Helen Colhoun 1 , Stuart McGurnaghan 1 , Luke Blackbourn 1 , Timothy J McDonald 3 , Colin N A Palmer 10 , John A McKnight 6 , Mark W J Strachan 6 , Alan W Patrick 5 , John Chalmers 7 , Robert S Lindsay 4 , John R Petrie 4 , Sandeep Thekkepat 8 , Andrew Collier 11 , Sandra MacRury 12 , Paul M McKeigue 2 *, on behalf of SDRNT1BIO investigators

1 Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital Campus, Crewe Road, Edinburgh, UK

2 Usher Institute of Population Health and Informatics, University of Edinburgh, Old Medical School, Teviot Place, Edinburgh EH8 9AG, UK

3 Medical School, University of Exeter, Exeter, UK

4 Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK

5 Royal Infirmary of Edinburgh, Edinburgh, UK

6 Metabolic Unit, Western General Hospital, Edinburgh, UK

7 Diabetes Centre, Victoria Hospital, Kirkcaldy, UK

8 David Matthews Diabetes Centre, Monklands Hospital, Airdrie, UK

3 Medical School, University of Exeter, Exeter, UK

10 Medical School, University of Dundee, Dundee, UK

11 Glasgow Caledonian University, Glasgow, UK

12 NHS Highland Diabetes Centre, Inverness, UK

* Corresponding author: paul.mckeigue@ed.ac.uk

Appendix – statistical principles and computational techniques

Mixed models specify covariates as “fixed effects” and encode variation between individuals by specifying “random effects”, typically with a Gaussian distribution and zero mean. Inference about fixed effects depends upon integrating out (averaging over) the random effects to compute the marginal likelihood as a function of the fixed effects. For linear regression models, where the outcome is a continuous variable, and the residuals are modelled as having a Gaussian distribution, random effects that have a Gaussian distribution can be integrated out analytically. Such linear mixed models can be fitted using standard statistical packages. For generalized linear models, which have a non-Gaussian likelihood and a link function_ that relates the expected value of the outcome to a linear function of the covariates, the random effects cannot be integrated out analytically. Although approximate methods for evaluating the integral over the random effects and maximizing the marginal likelihood of the fixed-effect parameters are implemented in standard statistical packages, in many situations these approximations are not accurate and the algorithms fail to converge. To evaluate the marginal likelihood of the fixed effect parameters, we used the Bayesian program Stan [27] to sample the the posterior distribution of the fixed effects while marginalizing over the random effects.

For each parameter, the likelihood function was evaluated by fitting a kernel density (using the R function `density` with default bandwidth) to the posterior samples weighted by the inverse of the prior density. A linear regression was used to fit a quadratic function to the logarithm of this likelihood function. Maximum likelihood estimates and classical significance tests were obtained from this quadratic approximation to the log-likelihood. The results obtained by this procedure do not depend on the prior distribution specified for the parameter under study, only on the prior distribution specified for nuisance variables.

For mixed models fitted with Stan, the fixed-effect regression parameters were rescaled with a QR reparameterization to improve computational stability. Gaussian priors with zero mean and standard deviation 3 were specified for these rescaled parameters, and the QR reparameterization was reversed to present the results in the original units. Random effects for individuals were specified as standard normal variates shifted by the intercept and scaled by the standard deviation: this non-centred parameterization makes sampling more efficient. For the standard deviation of random effects a half-Cauchy prior centred at zero with scale parameter 5 was specified. For each model, four sampling chains were run, each with 1000 iterations, of which the last 800 were used as posterior samples. Less than 20 iterations were required for the log posterior to stabilize. Diagnostics for the sampling algorithm were the rate of divergent transitions, the mixing of the chains evaluated as the R-hat statistic, and the effective sample size for each parameter. The target acceptance rate during the adaptation phase of the sampling algorithm was set to 0.95, which was sufficient to limit the rate of divergent transitions to 1%.