Statistical modelling of risk factors for hypoglycemic attacks

Paul McKeigue

Molecular, Genetic and Population Health Sciences
University of Edinburgh

HYPO RESOLVE

Hypoglycaemic events: modelling risk factors and prediction

- Aims
  - Examining risk factors for severe hypoglycaemia
- Objectives
  - Determine the relative contribution of known predictors of hypoglycaemia with greater precision
  - Identify new risk factors which can predict future episodes or are amenable to intervention

HYPO RESOLVE

Data sources

- All available trials in which hypoglycaemia events are recorded appropriately
- Studies can be analysed jointly or separately, depending on data availability and level of harmonisation possible.
- Harmonized dataset curated on Hypo-RESOLVE server (WP3)

HYPO RESOLVE

Eligibility and end points

- Inclusion/exclusion criteria: diagnosed with Type 1 or Type 2 Diabetes
- Entry times: point of entry into study
- Exit times: end of trial, end of follow-up or death
- Outcome: hypoglycaemia events, defined as
  - Major, minor, symptoms only
  - Score of hypoglycaemia based on intensity and duration

HYPO RESOLVE

Covariates to be modelled

Selection of covariates will be informed by systematic review

- 2 sets of covariates:
  - Variables previously reported in the literature
  - New hypothesised variables
- Missing data
  - Multiple imputations up to 20% missingness threshold

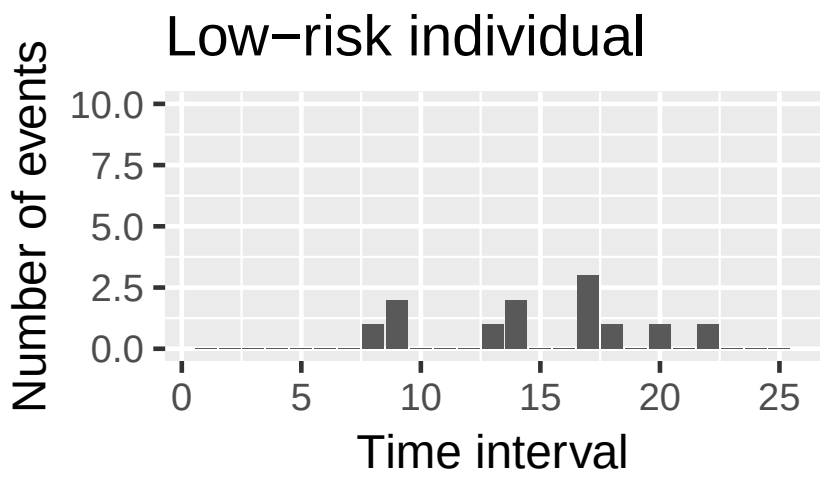
HYPO RESOLVE

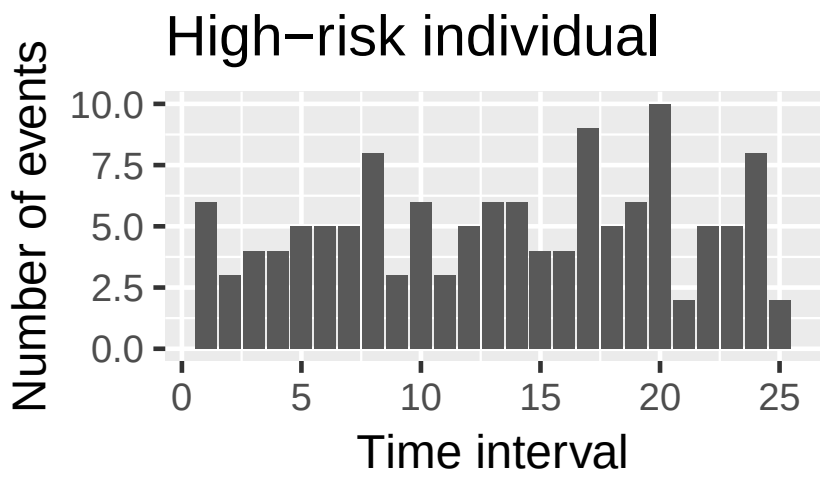
Hypoglycaemic events: modelling risk factors and prediction

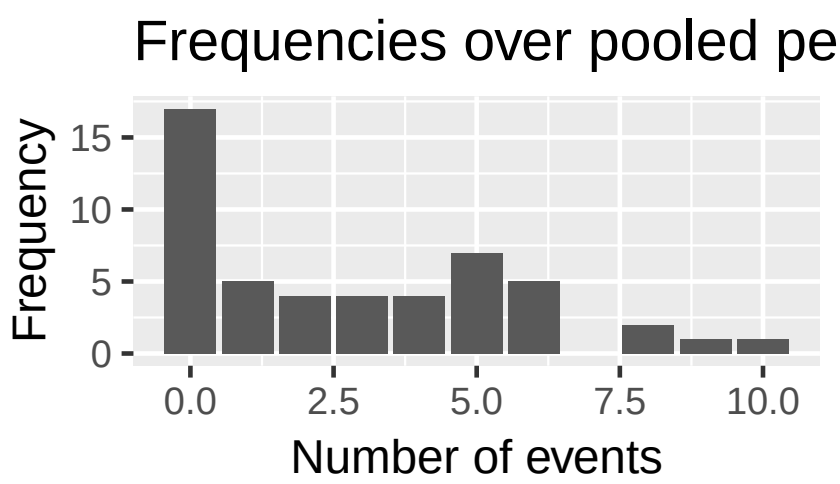
- Hypoglycaemic events occur repeatedly within each patient
  - To model time-updated covariates, data must be split into person-time intervals
  - Counts of events in each person-time interval are usually small with many zero values
    - cannot be approximated by a Gaussian (normal) distribution
  - Hazard rates vary between individuals

HYPO RESOLVE

Distribution of counts of events based on pooling person-time intervals over individuals







HYPO RESOLVE

Standard modelling approach: negative binomial regression

- Pooling person-time intervals from individuals with different hazard rates gives a distribution of counts of events that is *overdispersed*: variance $>$ mean
  - Poisson distribution has variance equal to its mean
- Negative binomial distribution is a convenient way to model an overdispersed distribution of counts of events:
  - parameterized with mean λ and and dispersion parameter α
  - variance is $\lambda\frac{\lambda+\alpha}{\alpha}$
  - equivalent to a mixture of Poisson distributions where the hazard rates are drawn from a gamma distribution with shape parameter α.
- Counts in ith person-time interval modelled as

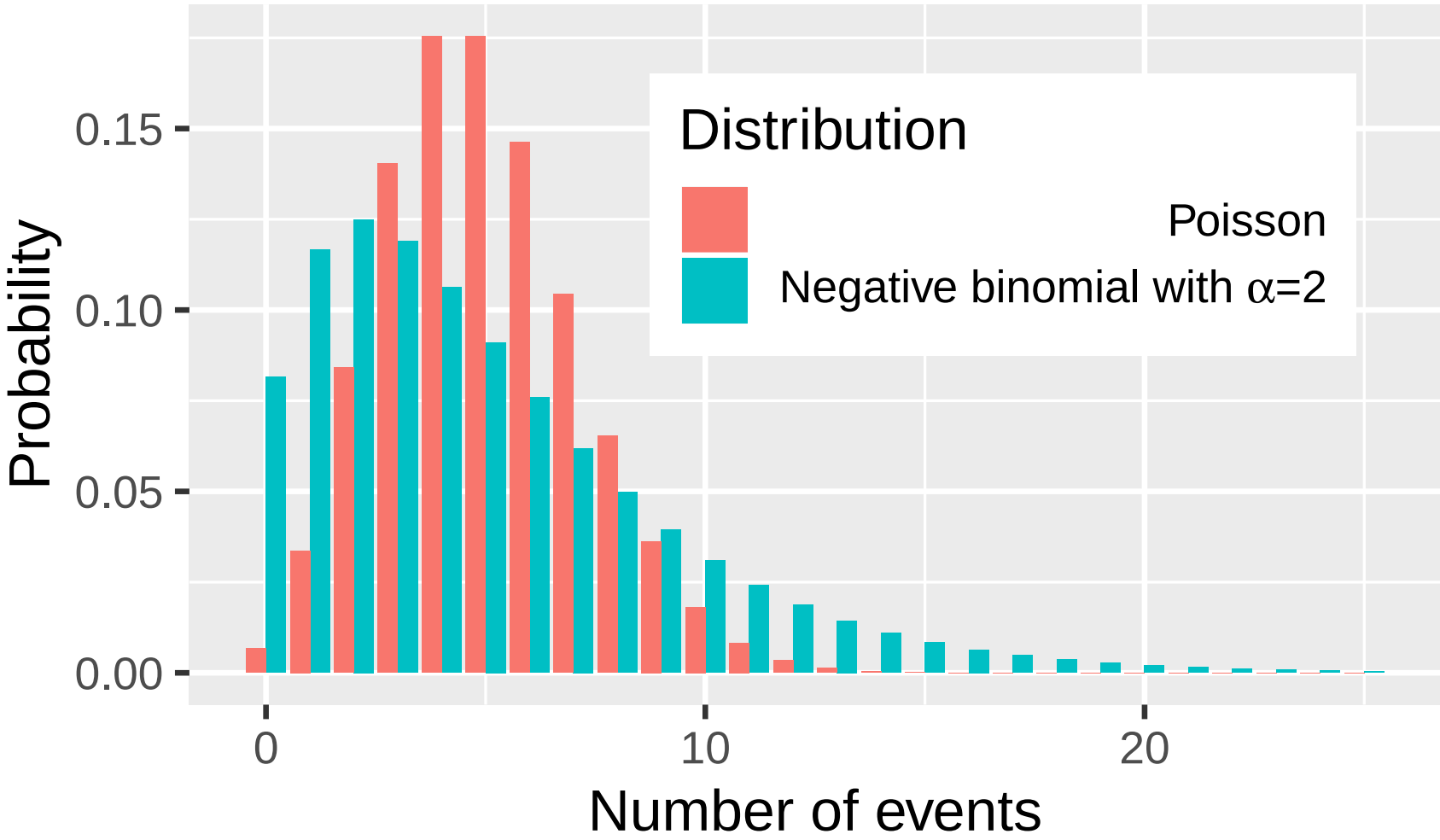
$$y_i \sim \text{Negative Binomial}\,(\lambda_i, \alpha)$$

$$\log \lambda_i = \beta_0 + \beta_1 x_{1i} + \ldots$$

.

HYPO RESOLVE

Comparison of Poisson and negative binomial distributions with mean 5



HYPO RESOLVE

Limitations of negative binomial model

- Negative binomial regression ignores information about which person-time intervals are repeat observations on the same individuals.
- Negative binomial regression assumes that the dispersion parameter is constant across person-time intervals given covariates (Luo and Qu, 2013), even where these person-time intervals are contributed by different persons.
- Regression coefficients are log ratios of population mean event rates between strata, not interpretable as log hazard rate ratios conditional on baseline risk.

HYPO RESOLVE

A simulation study of negative binomial regression for modelling event rates in a sample of individuals who vary in susceptibility

- Simulated 1000 samples each of 200 individuals from the following model:
  - log baseline hazard rates distributed as normal with mean zero and standard deviation 2
  - two covariates distributed as standard normal with regression coefficients -0.5, 0.
  - Events in each person-time interval distributed as Poisson with log hazard rate given by linear predictor
  - Average 11 person-time intervals observed per individual
- Results of fitting a negative binomial regression model:
  - algorithm failed to converge in 6% of draws
  - p-value for effect of covariate with zero true effect was < 0.01 in 62% of draws.

Statistical methods for modelling hypoglycemia as outcome in Hypo-RESOLVE

- To model repeat observations of a continuous variable on the same individuals, we need to specify the variation between individuals as **random effects**.
- Effects of covariates in a regression model are **fixed effects**
- A mixed model has fixed effects for the covariates, random effects for individuals.

HYPO RESOLVE

Methods for fitting a mixed model with Poisson likelihood

- Where the outcome is a continuous variable, we can specify a linear mixed model: easy to fit.
- Where the outcome is counts of events, we have to specify a **generalized** linear mixed model.
    - there is no exact method to calculate the integral that averages over the random effects.
- Methods for fitting a generalized linear mixed model, proposed in the Hypo-RESOLVE statistical analysis plan:
    - Approximate the integral over the random effects and maximize the likelihood of the fixed-effect parameters: implemented in the R package `lmer4`: fails with real data.
    - Bayesian approach: sample the posterior distribution of the regression coefficients: infeasible until recent development of efficient algorithms.

HYPO RESOLVE

Stan: a platform for Bayesan inference and imputation: Gelman, Lee and Guo (2015)

- Stan uses a *Hamiltonian Monte Carlo* algorithm (Duane, Kennedy, Pendleton & Roweth 1987) - to sample the posterior distribution given the data and the model
  - Hamiltonian Monte Carlo updates all parameters jointly: algorithms implemented in BUGS (1996) and JAGS (2007) can sample only one parameter at a time
  - Programs `PyMC3` and `pyro` implement the same sampling algorithm

HYPO RESOLVE

Stan or William?

Stan Ulam (Poland / USA, 1909-84)

- Markov chain Monte Carlo sampling algorithm
- Method of initiating a hydrogen bomb

William Rowan Hamilton (Ireland, 1805-65)

- Hamiltonian dynamics, variational principle of least action, quaternions

HYPO RESOLVE

Scottish Diabetes Research Network Type 1 Bioresource

- 6084 people clinically diagnosed as Type 1 diabetes or latent autoimmune diabetes of adulthood aged over 16 years at recruitment.
- C-peptide and autoantibodies measured at clinic visit

Follow-up for average 5.2 years through health records: clinic measurements including HbA1c and body mass index, hospital admissions

HYPO RESOLVE

Distribution of age at onset and duration

	Duration	0 to <5	5 to <15	15 to <25	25 to <35	35 -
Age at onset						
0 to <15		14	362	563	529	613
15 to <25		174	342	364	369	299
25 to <35		168	338	351	265	128
35 -		272	491	298	118	26

HYPO RESOLVE

Frequency of hypoglycaemic episodes requiring hospital admission in 120-day person-time intervals

Number of events	Frequency
0	97035
1	284
2	101
3	24
4	4
5	6
6	1
9	2
10	1

Overdispersion: variance is 2.2 $\times$ mean

HYPO RESOLVE

Logistic regression of >= 1 hypoglycemic episode during follow-up on baseline covariates

	Odds ratio	p-value
Intercept	0.03	8e-11
Gender	0.77	0.04
Age at diagnosis	1.02	1e-05
Duration (years)	1.03	3e-10
BMI (kg m^{-2})	0.95	1e-04
HbA1c (mmol/mol)	1.02	1e-07
C-peptide 5 to <30	0.74	0.1
C-peptide 30 to <200	0.55	0.008
C-peptide 200-	0.59	0.03

HYPO RESOLVE

Risk factors for >=1 hypoglycemic episode requiring hospital admission

- Baseline covariates associated with increased risk of hypoglycemia at follow-up:
  - Later age at diagnosis
  - Longer duration
  - Higher HbA1c
  - Lower body mass index
  - Absent / low residual C-peptide secretion
- Logistic regression is valid, but wastes information by ignoring length of follow-up and multiple hypoglycemic episodes.
- Cannot model effects of time-varying covariates with this approach.

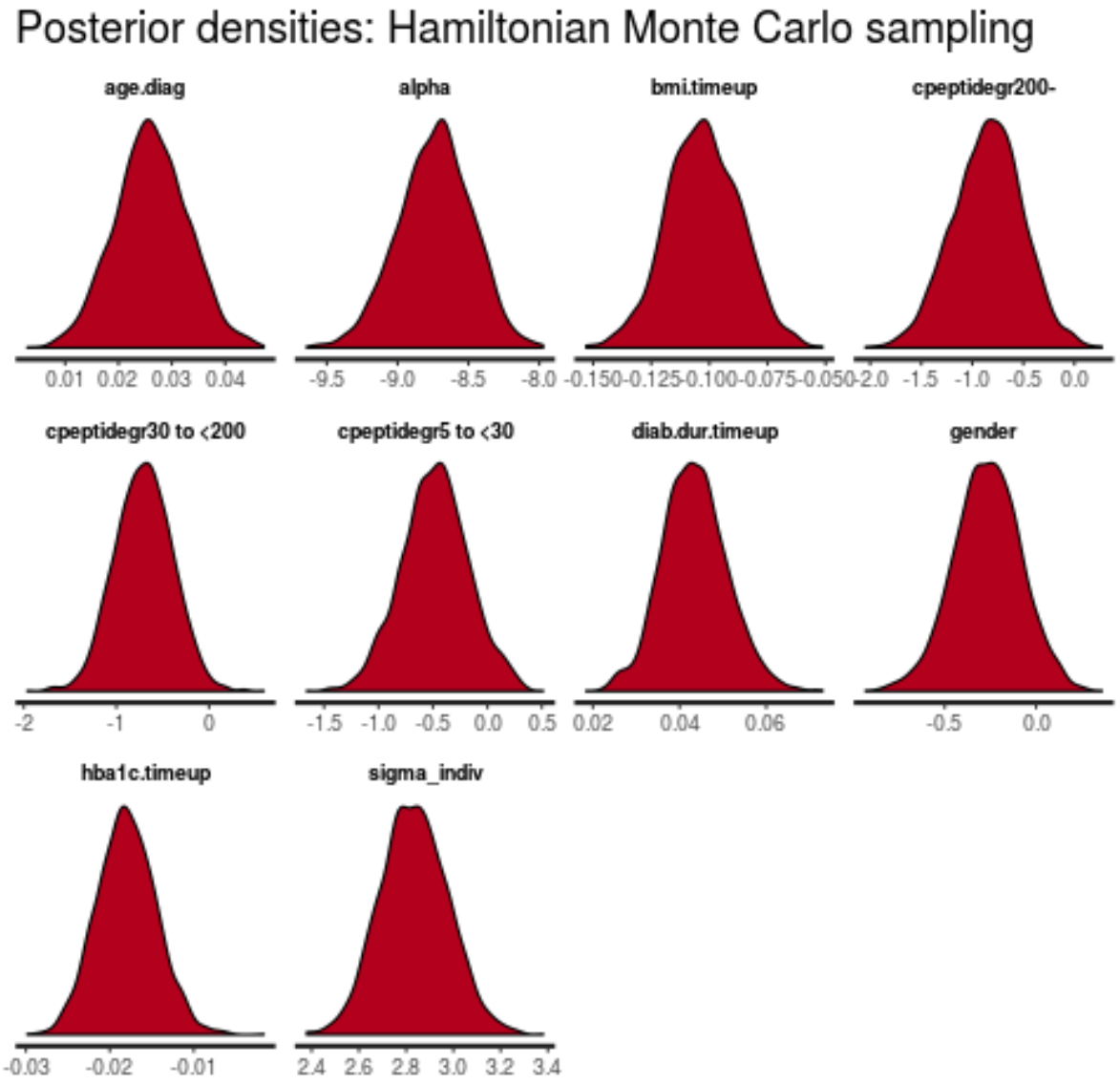
HYPO RESOLVE

Negative binomial regression of number of hypoglycemic episodes on time-updated covariates

	Ratio of means	p-value
Intercept	0.01	2e-18
Gender	0.96	0.7
Age at diagnosis	1.02	7e-04
***Duration (years)**	1.03	6e-09
***BMI (kg m^{-2})**	0.92	2e-08
***HbA1c (mmol/mol)**	1.01	0.002
C-peptide 5 to <30	0.78	0.2
C-peptide 30 to <200	0.64	0.04
C-peptide 200-	0.72	0.1

HYPO RESOLVE

Determinants of hypoglycemic episodes: Bayesian generalized linear regression



HYPO RESOLVE

Maximum likelihood estimate and p-value calculated from the posterior density

	Hazard ratio	p-value
Gender	0.75	0.1
Age at diagnosis	1.03	5e-05
***Duration (years)**	1.05	6e-09
***BMI (kg m^{-2})**	0.9	1e-11
***HbA1c (mmol/mol)**	0.98	1e-05
C-peptide 5 to <30	0.59	0.08
C-peptide 30 to <200	0.5	0.05
C-peptide 200-	0.42	0.009

- time-updated covariates

HYPO RESOLVE

Conclusions (1) - effect of residual C-peptide secretion on rates of serious hypoglycemic episodes

- Even very low levels of residual C-peptide secretion ($<$ 30 pmol/l) are enough to reduce the rate of serious hypoglycemic episodes by about 40%.
- This supports use of C-peptide levels as a surrogate end-point in trials of therapy to slow / reverse progression of Type 1 diabetes

HYPO RESOLVE

Conclusions (2) - statistical methods

- The standard statistical method – negative binomial regression – for modelling rates of severe hypoglycemia in clinical trials and observational studies should no longer be used.
  - ignores information about which observations are on the same individual
  - gives seriously misleading results on simulated data
- New tools for statistical computation make it possible to fit a mixed model with Poisson likelihood even to large and complex datasets.
  - outputs are Bayesian posterior distributions of the parameters of interest
  - classical estimates and p-values can easily be obtained where readers or regulatory agencies require them.

HYPO RESOLVE