Group based trajectory modeling (GBTM)

A primary aim of this research was to describe the natural history of glycaemia (as measured by HbA1c) over 12 years. The evolution of an outcome over time is its developmental trajectory, and it was hypothesised that there were groups of individuals within the cohort that follow distinctive developmental trajectories that were not identifiable prior to analysis.[1] GBTM was used to identify latent trajectory groups for HbA1c levels from age 26 to 38.[2]

GBTM is a specialised application of finite mixture modeling and involves a procedure which gathers individuals into meaningful subgroups that show statistically similar trajectories. It provides a statistical method to identify (rather than assuming a priori) groups of distinctive trajectories which are summarised by a finite set of different polynomial functions of age or time, as determined by maximum likelihood estimation.[1,2] The maximisation is performed using a general quasi-Newton procedure.[3] Rather than prescribing the existence of trajectories of a specific form ex ante on the basis of an individual trait or traits, the method allows the trajectories to emerge from the data itself. This offers an alternative to the limitations of using assignment rules based on inherently subjective categorisation criteria; it determines the form and number of groups that best fit the data; and it provides a metric for evaluating the precision of group assignments.[1] GBTM predicts the trajectory of each group, the form of each trajectory, estimates the probability for each individual of group membership and assigns them to the group for which they have the highest probability.

GBTM handles missing data by fitting the model using maximum likelihood estimation. This will generate asymptotically unbiased parameter estimates assuming the data are missing at random.[4] Data are considered to be missing at random (MAR) if the "missingness" is not related to the measured outcome. There were potentially three data points for each participant – at 26, 32 and 38. It was decided to include those who had data collected at two or more ages, and exclude those with fewer than two.

Trajectory groups are latent strata; that is, they are groups of individuals following approximately the same developmental course. Individuals do not actually *belong* to trajectory groups; rather, they are assigned a *probability* of group membership. Groups should not be reified (that is, should not be regarded as concrete or real). The cohort naturally follows a continuous rather than a discrete distribution; the model should then be regarded as a

convenient statistical device, rather than a state of being, for summarizing trajectories in distinctive regions of the distribution.[4] The number of trajectory groups is not immutable, and individuals do not follow the group-level trajectory in lock step.[5]

GBTM analyses were undertaken using Stata IC 12.0 for Windows (StataCorp 2011, *Stata Statistical Software: Release 12*, College Station, Tx, USA). The GBTM used a Stata Plugin for estimating the group-based trajectory model.[3,6] The Plugin generates parameter estimates which allow the calculation of a) the probability of group membership[a]; b) the predicted trajectory for each group[b]; and c) the posterior probabilities of group membership[c]. Mean HbA1c was modelled using the censored normal distribution.[6] Censors were set at values that were well beyond the range of any data values (minimum HbA1c = 10mmol/mol and maximum HbA1c = 150mmol/mol). A generalisation of the GBTM model was used to link baseline characteristics to the probability of group membership.[2,4]

We used the Bayesian information criterion (BIC)[d] as the criterion for model selection. However, this was moderated by (a) a preference for a useful parsimonious model which fitted the data well; (b) close correspondence between each group's estimated probability and the proportion of Study members classified to that group according to the maximum posterior probability assignment rule; (c) an average posterior probability (AvePP) value >0.7 for each group; (d) adequate sample numbers in each group; (e) reasonably narrow confidence intervals; and (f) the odds of correct classification based on the posterior probabilities of group membership >5 for each group.[4]

Identification of HbA1c GBTM groups

Identification of HbA1c GBTM groups began with 897 Study members who had two or more HbA1c assays over the 12 years. The four most extreme outliers were also removed because they gave rise to an analytically intractable 4-person group.

---

[a] The proportion of the population that belongs to each group.

[b] The capture of the essential features of a complex reality by a finite number of trajectory groups, each of a specific form, whether zero-order, linear, quadratic, cubic or higher.

[c] The collective measurement of each individual's probability of belonging to each trajectory group.

[d] The BIC (Bayesian information criterion) was introduced as an alternative to the Akaike information criterion (AIC) in 1978.[7] The two criteria are closely related, are both model selection criteria, and feature the same goodness-of-fit term. Generally the BIC penalises free parameters more strongly than does the AIC, so the BIC favours more parsimonious models. The model with the highest (least negative) value of BIC and AIC is preferred.

GBTM began with the choice of the number of groups to include in the model. (Table A1). We started with 2-group models, testing zero-order, linear and quadratic specifications for the trajectory shapes. Extra groups were added (3-, 4- and 5-groups) until the best fitting model was established.

Although the (1 2 2) model had a slightly higher BIC than the (2 2 2) model, the latter fitted the data better, and was found to capture the essential features of the data in a more parsimonious, comprehensible and analytically tractable manner. Both the 4-group and the 5-group models gave rise to one or more groups with a very small proportion of the observations (ten or less individuals). Thus the 3-group model with three quadratic trajectories (2 2 2) was chosen.

Table A1. BIC for HbA1c GBTM according to number of groups and trajectory shapes.

Number of groups	Trajectory shapes[1]	BIC[2] (N = 896)	BIC[3] (N = 2520)
2	0 0	-7205.68	-7207.75
2	0 1	-6912.80	-6915.38
2	0 2	-6898.46	-6901.56
2	1 1	-6781.01	-6784.11
2	1 2	-6839.65	-6843.26
2	2 2	-6828.91	-6833.05
3	0 0 0	-6879.95	-6883.05
3	0 1 1	-6371.67	-6375.80
3	0 1 2	-6360.70	-6365.34
3	0 2 2	-6335.71	-6340.88
3	1 1 1	-6346.80	-6351.45
3	1 1 2	-6338.70	-6343.86
3	1 2 1	-6314.04	-6319.21
3	1 2 2	-6311.06	-6316.74
3	2 1 0	-6382.19	-6386.84
3	2 1 1	-6342.69	-6347.85
3	2 1 2	-6337.07	-6342.75
3	2 2 1	-6313.65	-6319.33
3	2 2 2	**-6311.47**	**-6317.67**
4	0 0 0 0*	-6882.62	-6878.49
5	0 0 0 0 0*	-6890.45	-6885.29

[1]Trajectory shapes; 0 = zero-order; 1 = linear; 2 = quadratic.
[2]BIC = Bayesian information criterion (for the total number of participants)
[3]BIC = Bayesian information criterion (for the total number of observations)
*One or more of the groups had a very small proportion of the observations.

The model had an adequate proportion and sample number in each group: “Low” 11.0%, “Medium” 54.0%, and “High” 35.0% (Figure A1). The matrix of the observed and predicted values showed that the model fitted the data well and confidence intervals were narrow for each group (Table A2). The average posterior probability (AvePP) value was 0.84 or more for each group; well above the recommended minimum AvePP value of 0.70 (Table A3). The odds of correct classification based on the posterior probabilities of group membership were over 5.0 for all three groups, indicating the model had good assignment accuracy (Table A3). Finally, there was very close correspondence between each group’s estimated probability and the proportion of Study members assigned to it according to the maximum posterior probability assignment rule (Table A4).

Fig A1. HbA1c trajectory groups

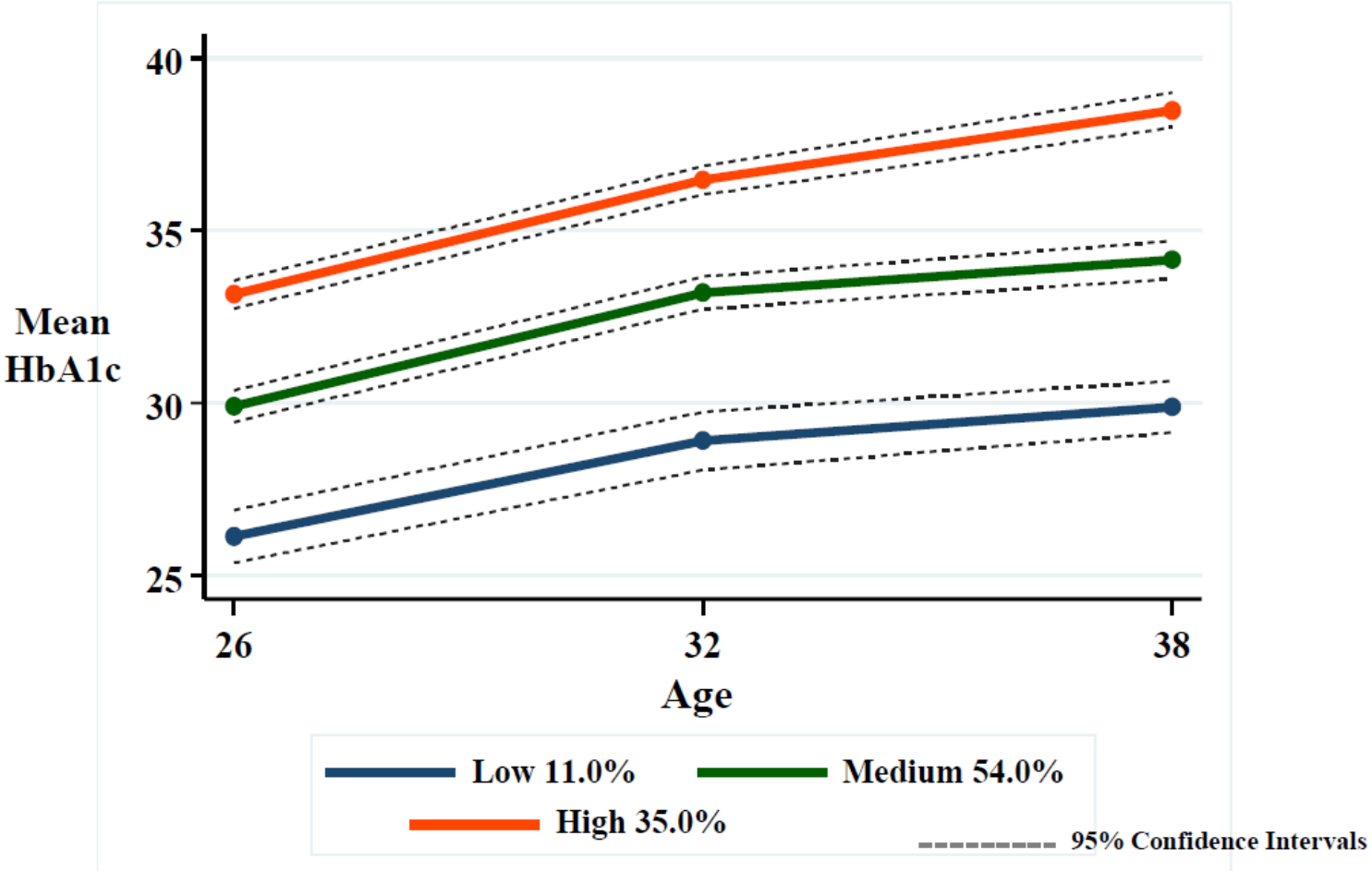

Table A2. Matrix of the observed and predicted values for mean HbA1c GBTM groups

	HbA1c trajectory group		
	Low	**Medium**	**High**
Age 26			
Observed values	26.13	29.91	33.14
Predicted values (95% CI)	26.13 (25.37, 26.89)	29.91 (29.45, 30.36)	33.14 (32.73, 33.55)
Age 32			
Observed values	28.90	33.20	36.45
Predicted values (95% CI)	28.90 (28.06, 29.74)	33.20 (32.73, 33.66)	36.45 (36.04, 36.87)
Age 38			
Observed values	29.89	34.15	38.49
Predicted values (95% CI)	29.89 (29.14, 30.63)	34.15 (33.59, 34.70)	38.49 (37.98, 38.99)

Table A3. Average posterior probability (AvePP) value and odds of correct classification for HbA1c GBTM groups

	HbA1c trajectory group		
	Low	**Medium**	**High**
Average posterior probability value	0.84	0.86	0.87
Odds of correct classification	43.3	5.2	12.6

Table A4. HbA1c trajectory groups' estimated probability and the proportion of Study members classified to each group according to the maximum posterior probability assignment rule

Group	**Estimated group probability**	**Proportion assigned to group according to the maximum posterior probability assignment rule**
Low	11.3	11.0
Medium	52.5	54.0
High	36.1	35.0

1. Nagin DS. Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol Methods* 1999;42(2):139-57.

2. Nagin DS. Group-Based Modeling of Development. Cambridge, MA: Harvard University Press, 2005.

3. Jones BL, Nagin DS. A Stata plugin for estimating group-based trajectory models. https://www.andrew.cmu.edu/user/bjones/ 2012.

4. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol* 2010;6:109-38.

5. Nagin DS, Tremblay RE. Developmental trajectory groups: Fact or a useful statistical fiction? *Criminology* 2005;43(4):873-904.

6. Jones B, Nagin D. A Note on a Stata Plugin for Estimating Group-based Trajectory Models. *SMR* 2013;42(4):608-13.

7. Schwarz G. Estimating the Dimension of a Model. *Ann. Statist* 1978;6(2):461-64.