High Dimension Multiple Imputation: Missing Blood Glucose Values in the Epidemiology of Diabetes Interventions and Complications Study

Mike Larsen, George Washington University Statistics and Biostatistics Center

John Lachin, GWU, Epi/Biostat, Statistics, BSC

Monday, June 14, 2010, Poster Presentation

Stoweflake Inn, Stowe, VT

Outline

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON DC

1. The study and its missing data
2. Relevance to sample surveys
3. Multiple imputation using chained equations (MICE)
4. Example from a single subject
5. Preliminary results
6. Issues in evaluations of imputations and future work

The DCCT/EDIC study

Randomized clinical trial 1984-1993 of 2 treatments for type 1 (insulin dependent) diabetes. Up to 9 years of data collection.

- n=729 conventional treatment
- n=711 intensive treatment

Baseline weight, height, and other factors.

Blood glucose (BG) profiles – 7 at home measurements during 1 day each quarter

Plus clinic visit measurements: glycated hemoglobin HbA1c quarterly, diabetic retinopathy (microvascular retinal changes) 2x year, nephropathy (kidney disease) 1x year. Genetic information also is available.

Subjects are still being followed for conditions/outcomes.

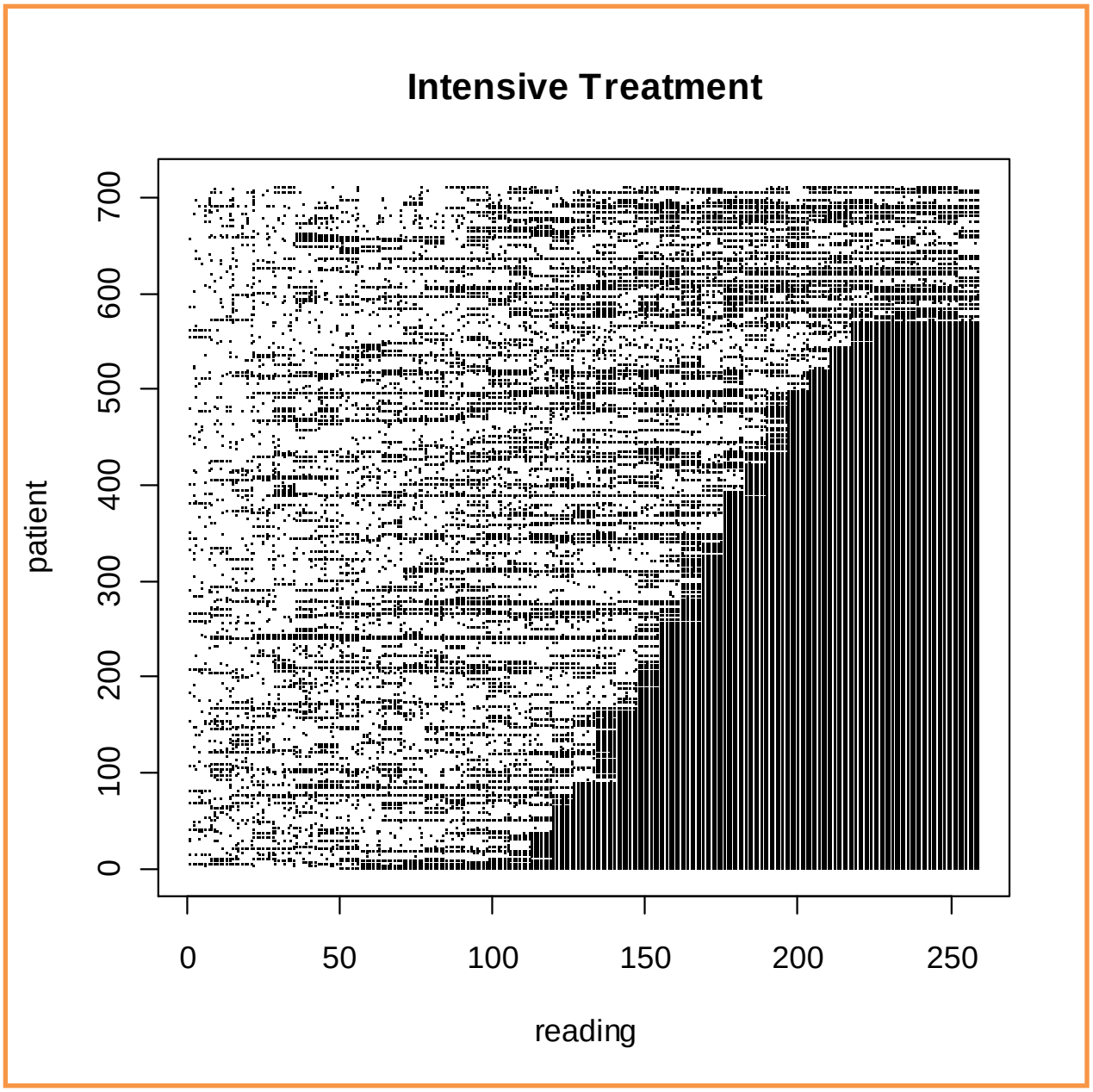
The DCCT/EDIC study

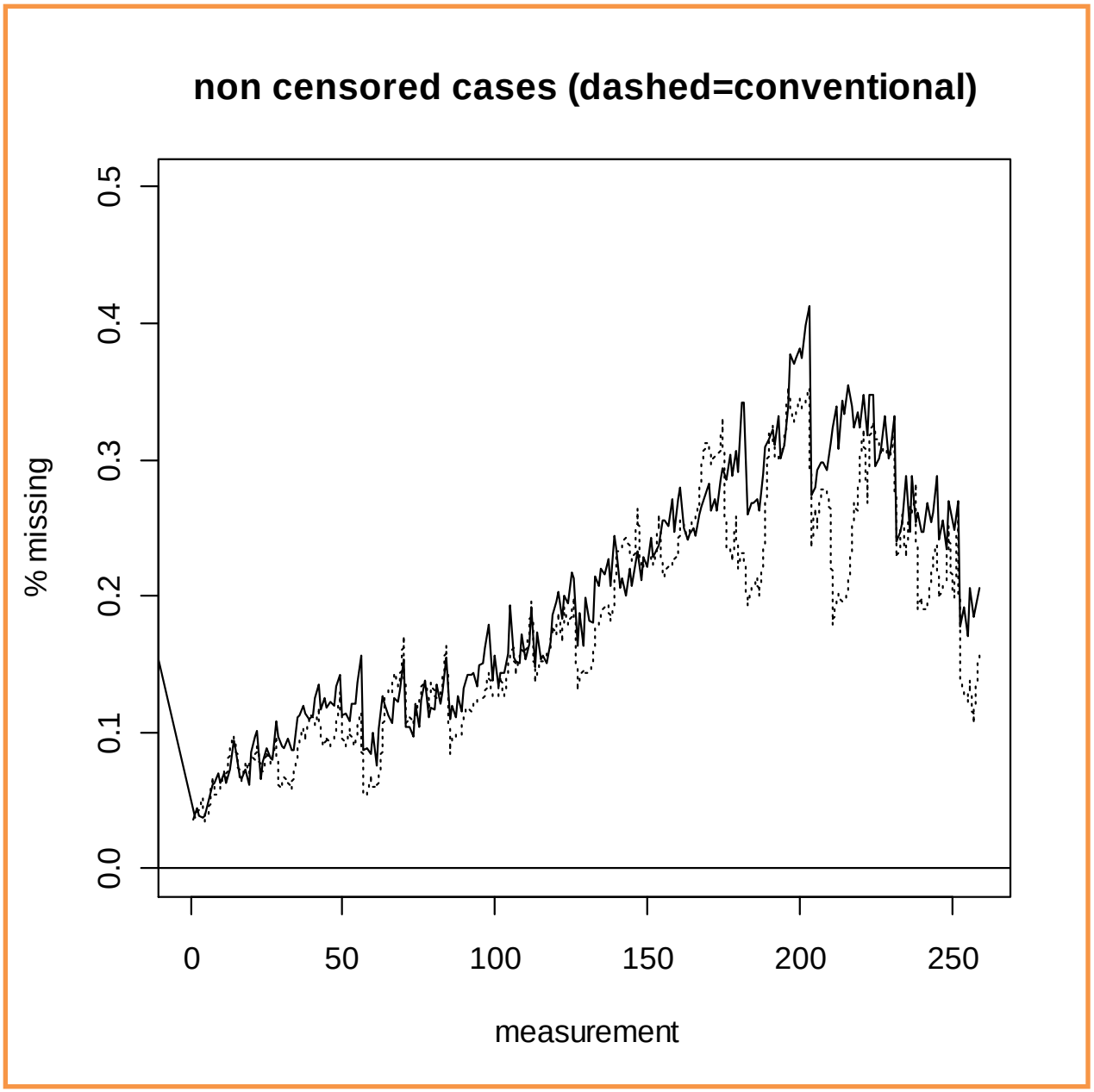
- **Reported findings** have had fundamental impacts on treatment in practice. Recent work has found genetic associations to health and diabetes-related conditions.
- **New hypotheses:** Both level and variability in blood glucose over time impact diabetes-related conditions.
- **Missing data** complicate analyses, reduce power, and make summaries (mean, SD) over time less useful for some subjects.

Missing data

- End-of-study censoring [late beginning for some]
- Missed/incomplete blood glucose profiles by qtr
- Screen for conditions biannually/yearly

n=729 cases originally in conventional treatment				
Observations through…	All cases not censored	And 60% of BG recorded	And 80% of BG recorded	And 90% of BG recorded
8 Quarters	728	708	657	547
16 Quarters	726	701	619	464
32 Quarters	168	155	132	95


Intensive Treatment
patient
0
100
200
300
400
500
600
700
0
50
100
150
200
250
reading


non censored cases (dashed=conventional)
% missing
0.0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200
250
measurement

Relevance to sample surveys

- Survey data are multivariate and have missing values for various reasons
- Longitudinal/panel studies have similar issues of drop-out (and return) and correlation over time
- Some health/medical surveys collect similar measurements, including genetics
- How to do *and evaluate* multiple imputation effectively is a common concern

Approach to Missing Data: MI/MICE

Goal: produce a filled-in data set and enable proper assessment of standard errors for many analyses.

Approach: multiple imputation (MI) using chained equation (MICE) modeling.

- Specify a statistical model for each variable given other variables as predictors
- Generate random values for missing measurements from the appropriate model conditional on other values
- Reflect variability due to missingness by creating multiple sets of imputations, repeatedly analyze completed data sets, and combine results for estimates & SEs.

About MI and MICE

- *Imputation:* fill-in missing with plausible values. Plausible values are generated from a statistical model.
- *Multiple imputation (MI):* create *M* (*M≥5*) imputes for each missing value. This in effect yields *M* completed data sets (observed data plus one set of imputes for missing values). Run the same analysis (a complete data analysis) on each of *M* completed data sets. Combine results to reflect uncertainty within and between the analyses. Doing so reflects uncertainty due to missing values.
- *Dimension:* 259+ variables are in the data set. Base-line plus 36 quarters for blood glucose (7 measurements each time) profile. HbA1c, retinopathy, nephropathy potentially each quarter. The latter two actually measured less.

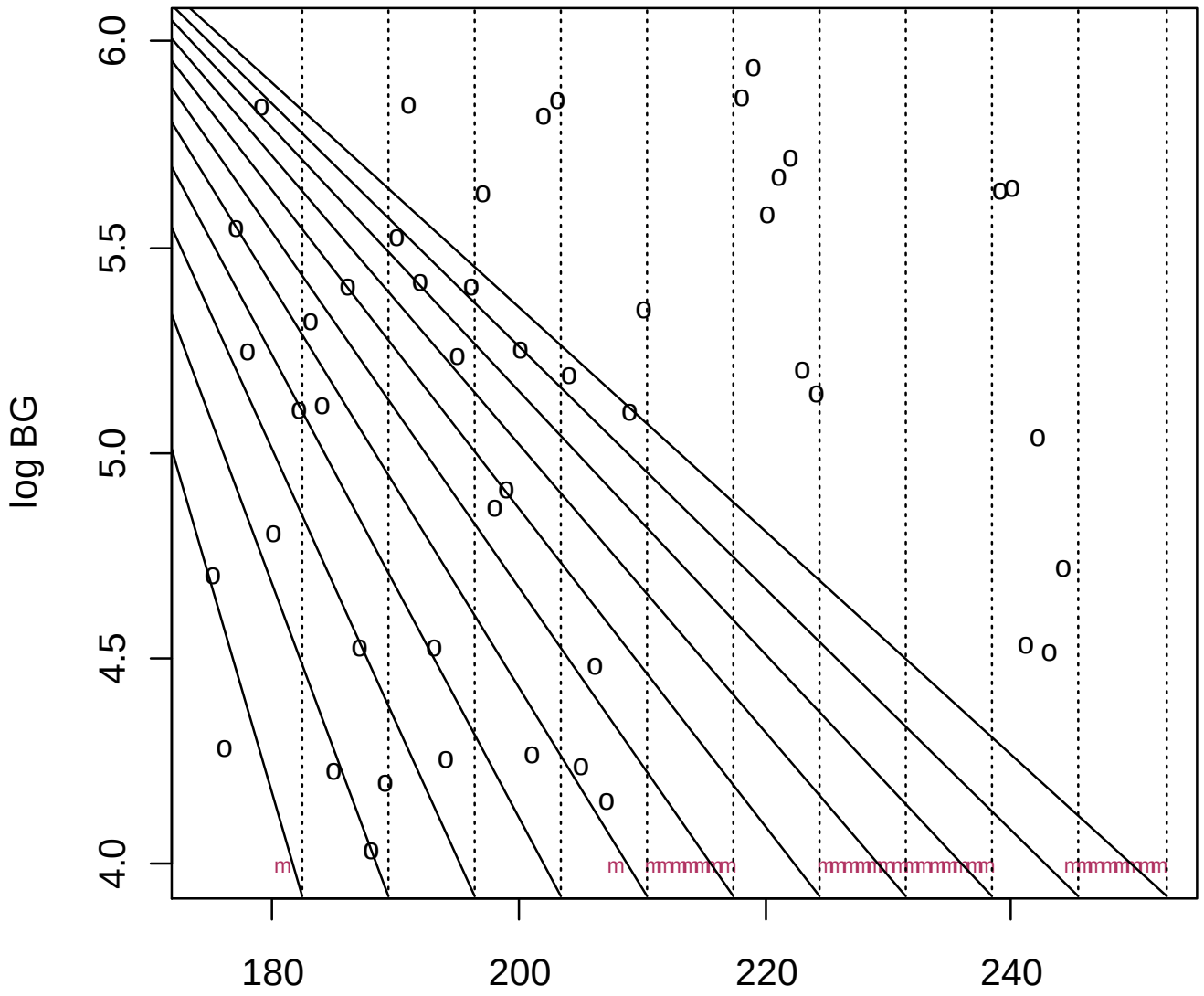
About MI and MICE

- *Chained equations:* fit model for one variable after another conditional on some of the variables measured close by in time; how big to make models is an issue. Specifying and fitting a model for one variable at a time is simpler than doing so for all variables at once.
- *Models:* log BG and log HbA1c look reasonably bivariate Normally distributed over time and in pair-wise plots; use linear regression models to predict one variable based on several others; other variables (e.g., genetic markers) are used as discrete predictors; binary variables predicted using logistic regression.

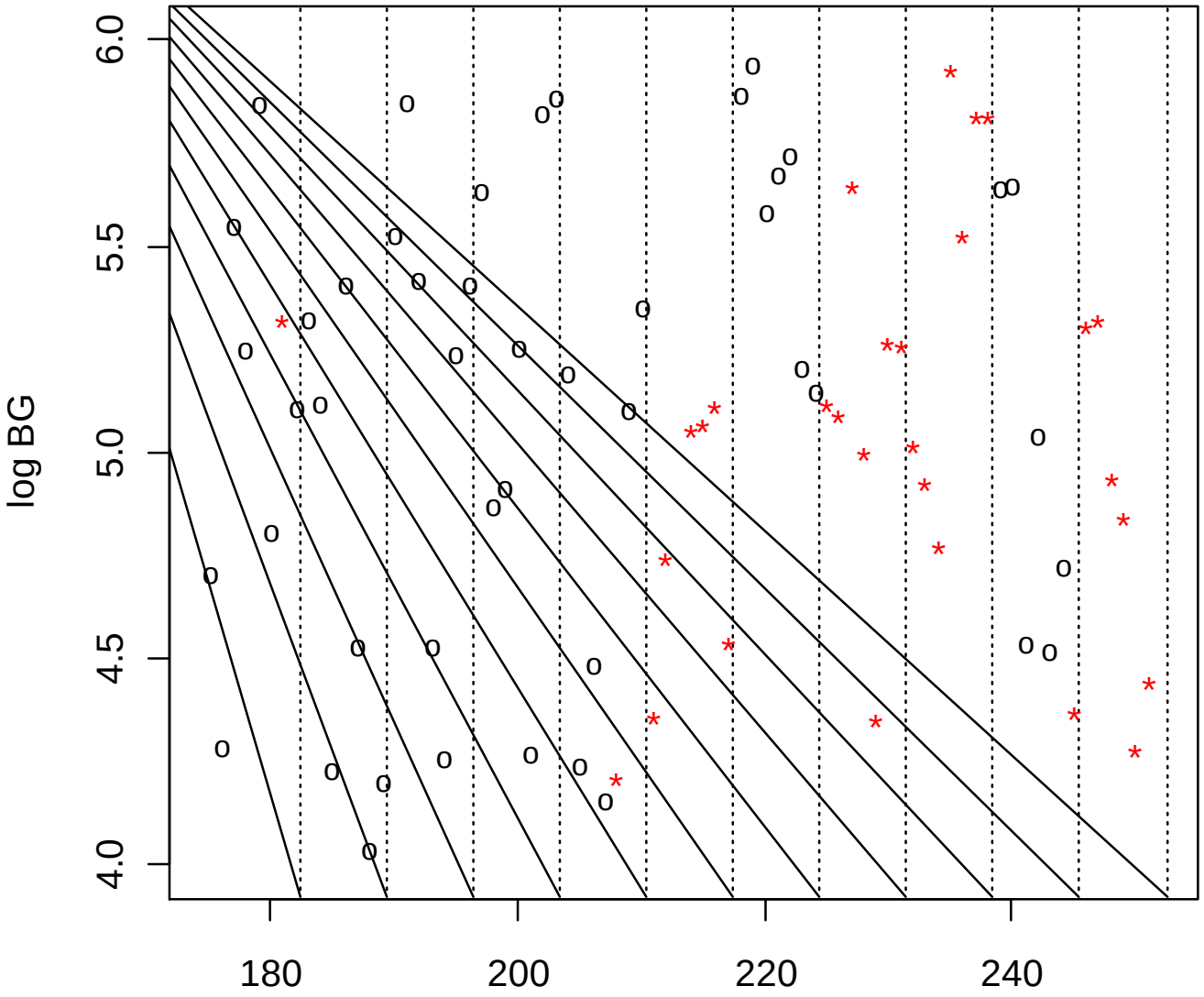
About MI and MICE

- *Software and Iterations:* The R package mice was used for computing. It cycles through each variable that needs imputation a few times, redoing imputations on each pass, until the distribution of values appears stable. There is some effort needed to check convergence of the algorithm and sensitivity of results to number of iterations, but the situation does not appear as complicated or to require as many iterations as Gibbs sampling.
- *Evaluating models and imputations*: basic distribution checks (means, standard deviations, percentiles, correlations) show imputations are reasonable; Posterior predictive checks (regular and cross validated) are being used for model selection; more work is planned in this area; in general this is an area of research.

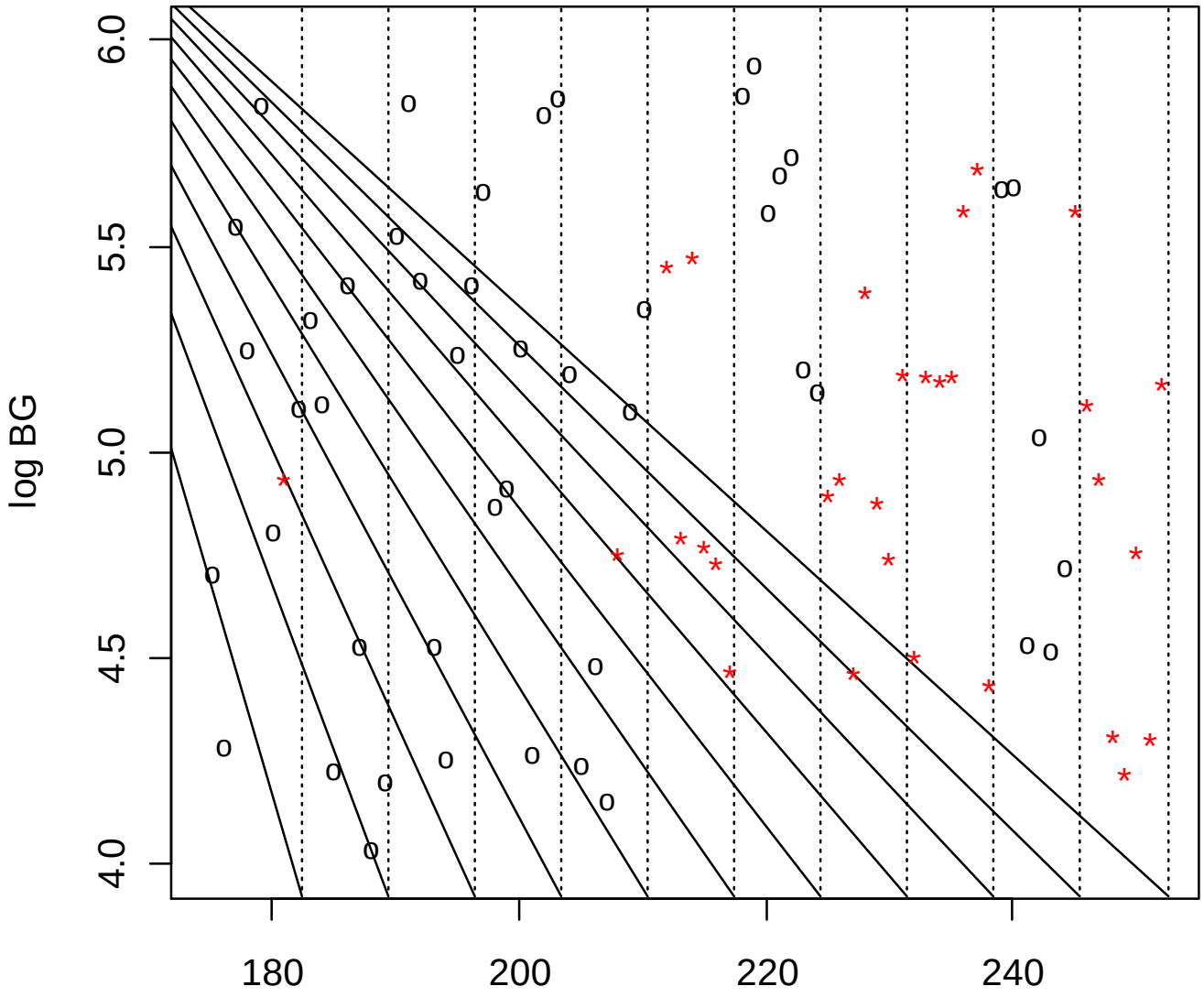
One subject, Blood Glucose over time



First imputed data set values

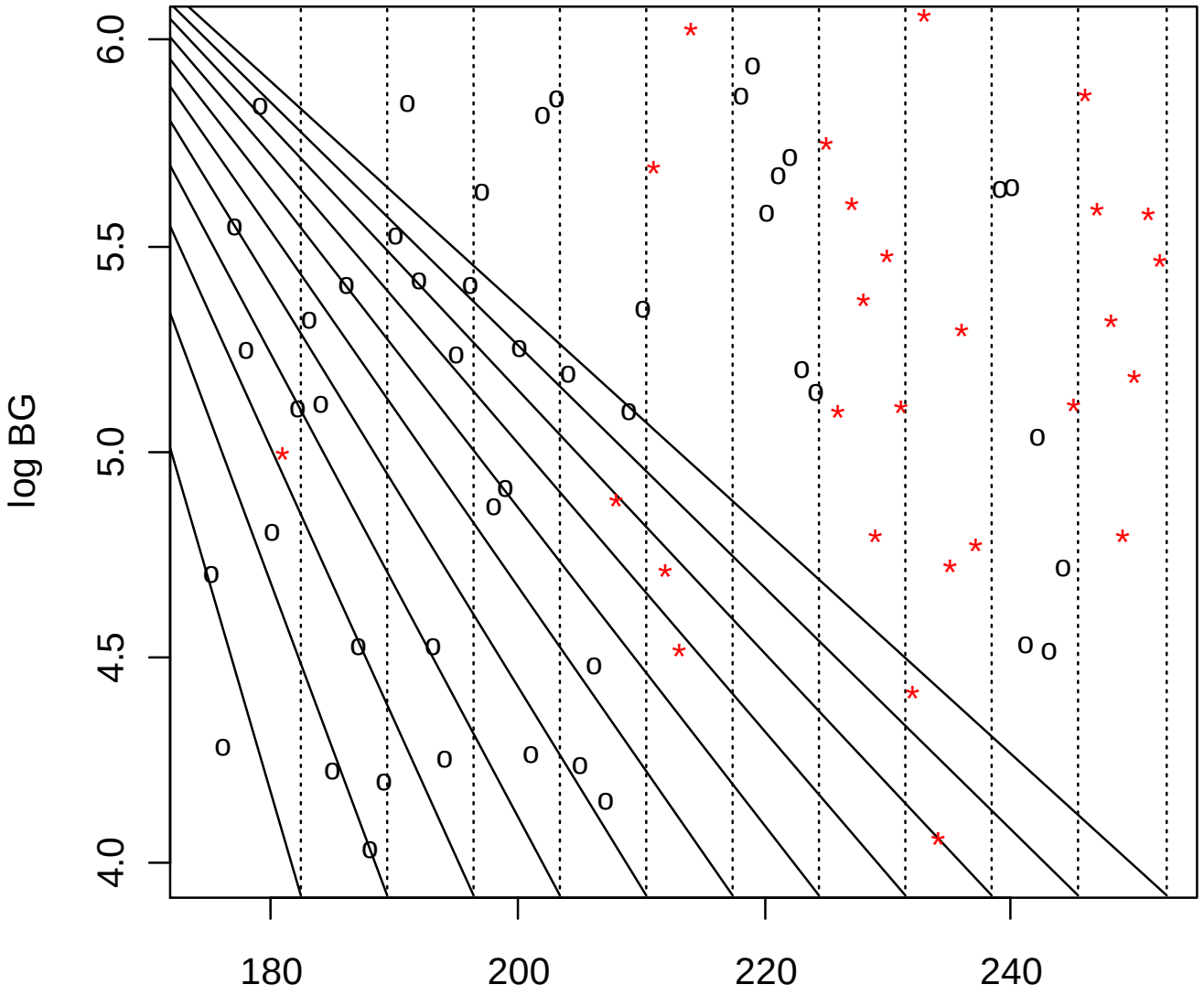


Second imputed data set values

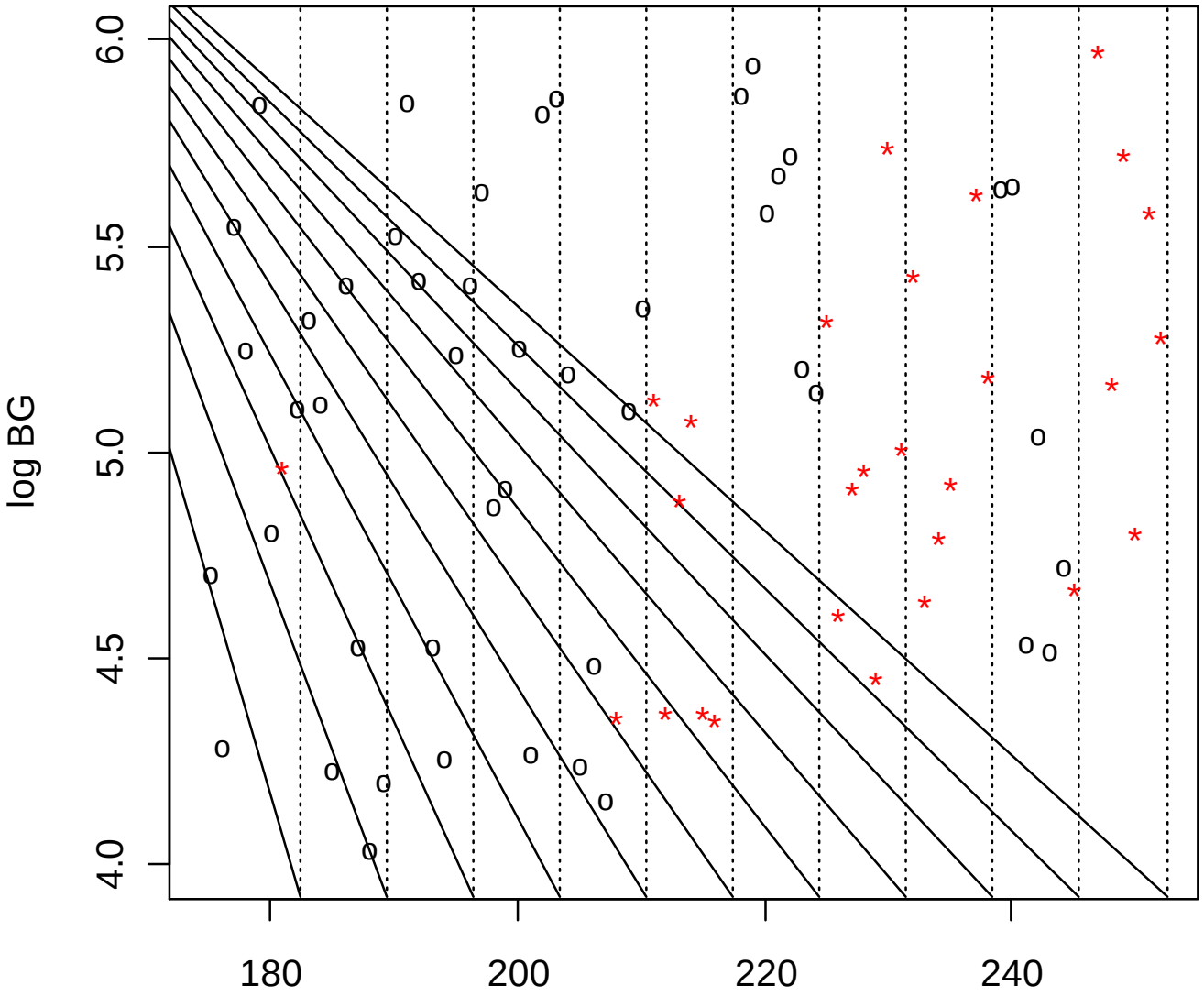


BG reading number (vertical lines divide quarters)
o=observed BG, *=imputed value

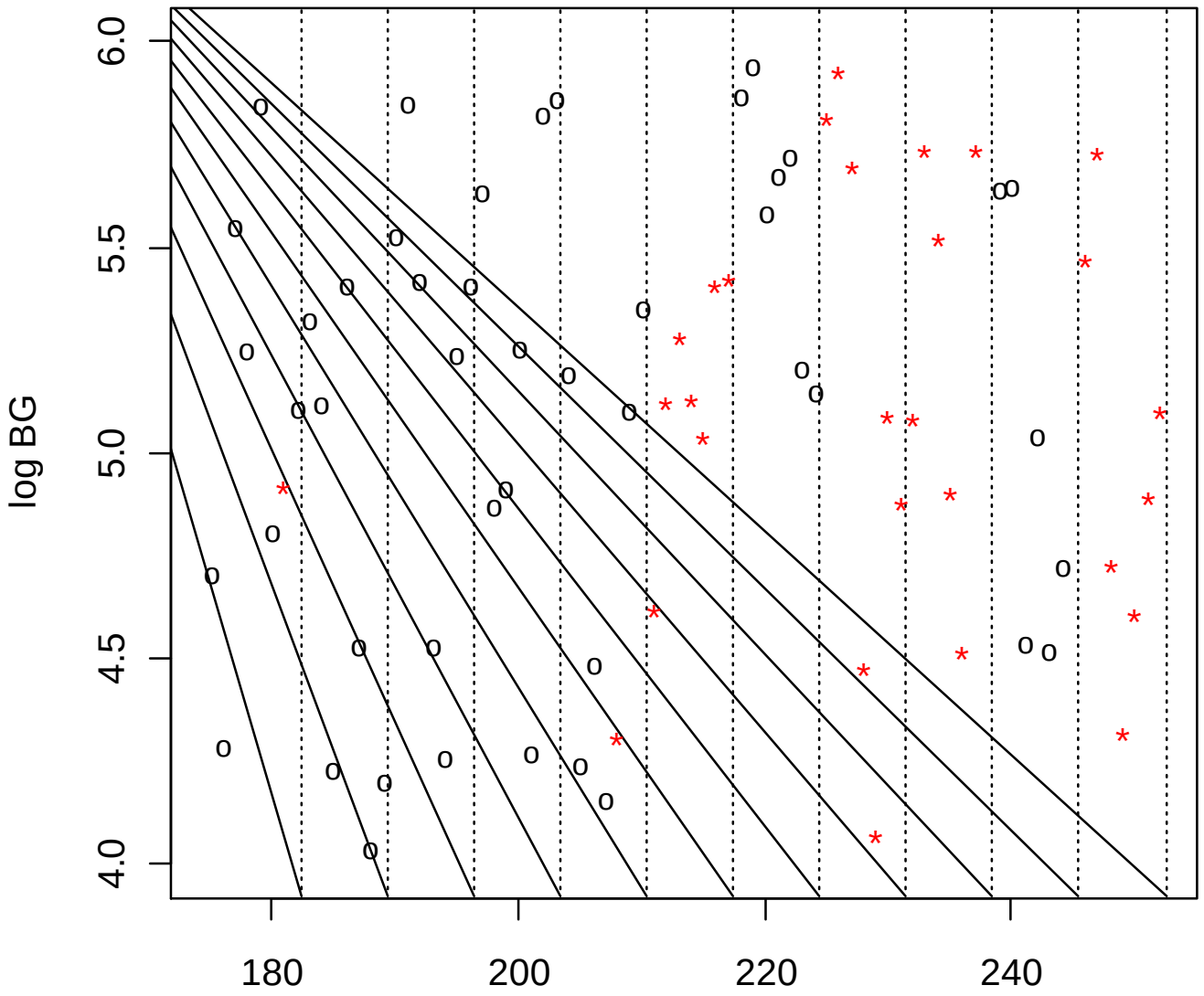
Third imputed data set values



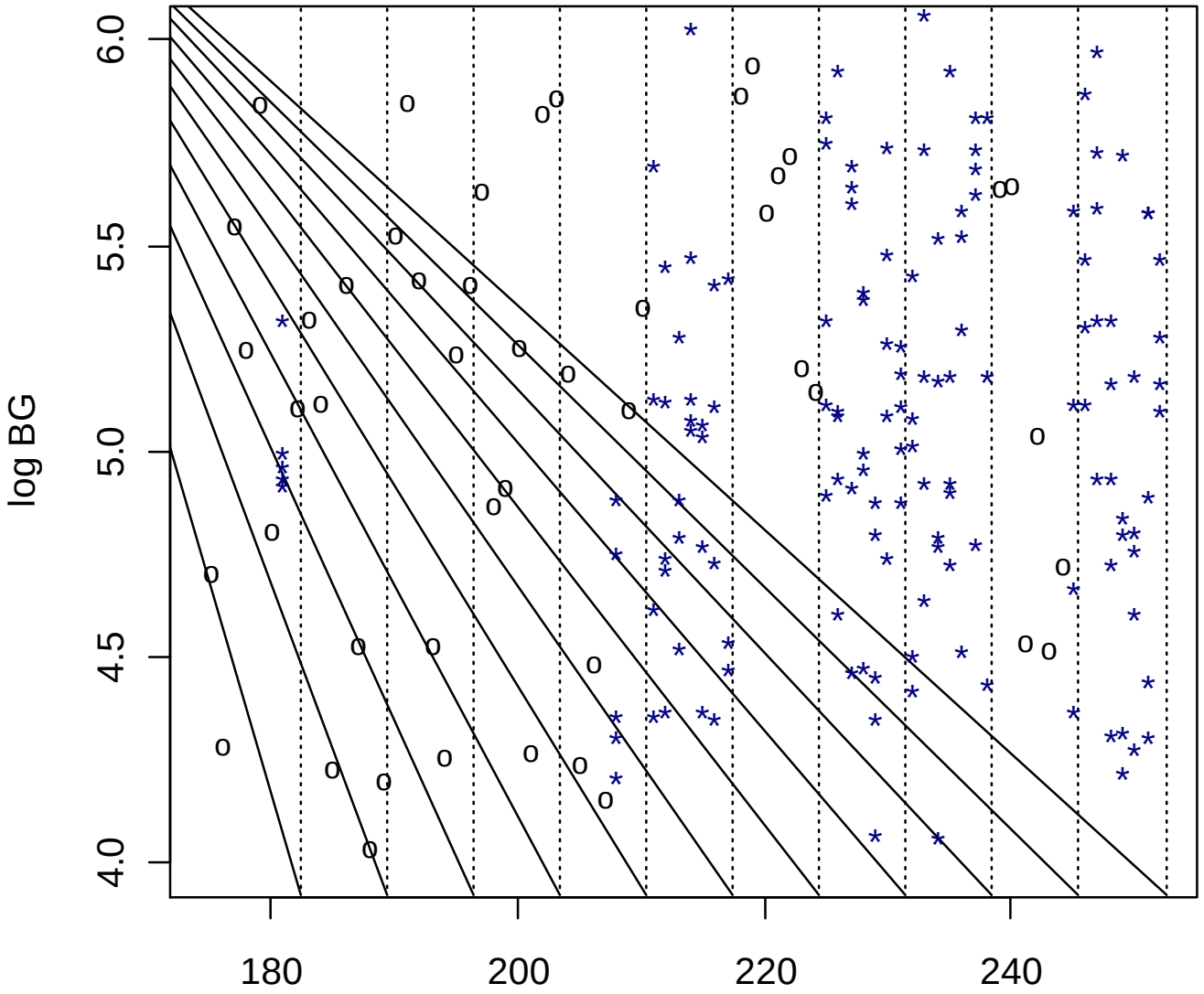
Fourth imputed data set values



Fifth imputed data set values



All imputed values



Why take the MI approach?

- *Completed data allow better description of cumulative exposure and variability over time!*
- *Multiple imputations allow expression of uncertainty due to missingness while enabling complete data analyses.*

Preliminary Results/Comments

- Restricting analysis to available cases with few missing BGP values increases correlation of mean BG with HbA1c and with an indicator of Retinopathy, but analyses have increased SE (decreased sample size) and results depend on selection criterion
- Multiple imputation analysis w/ BGP and HbA1c together in an imputation model shows similar higher correlation of mean BGP with HbA1c (but not as high as available cases), smaller SE (larger *n)* – this is desirable. Results not too sensitive to imputation models and procedures checked so far.

Future work

- Future work will incorporate more predictors (e.g., genetic indicators from diabetes, obesity, and other literatures that have been collected)
- Report more evaluations of imputations (e.g., more distribution checks, more comparisons of results under different imputation models)
- Posterior predictive checks (including cross validated ones) on imputations

Thanks!

mlarsen@bsc.gwu.edu

Multiple Imputation and MI with Chained Equations

Stuart, Azur, Frangakis ,Leaf. (2009), *American J Epidemiology,* 169(9): 1133-39.

Horton, Kleinman. (2007), *The American Statistician*, 61(1): 79-90.

Van Buuren. (2007), *Statistical Methods and Medical Research*, 16: 219-42.

Raghunathan et al. (2001), *Survey Methodology*, 27(2): 85-96.

R software MICE: http://cran.r-project.org/web/packages/mice/index.html

Van Buuren, Brand, Groothuis-Oudshoorn, and Rubin. (2006). *Journal of Statistical Computation and Simulation*, 76, 1049-1064

Rubin and Schenker. (1991) Multiple imputation in health-care databases. *Statistics in Medicine*, 10, 585-598

Rubin. (1996) MI after 18+ years (with discussion) . *JASA,* 91, 473-489

DCCT/EDIC references

http://www.bsc.gwu.edu, research projects (EDIC), completed projects (DCCT)

The DCCT Research Group (1997). Hypoglycemia in the Diabetes Control and Complications Trial. *Diabetes*, 45: 271-286.

The DCCT Research Group (1996). The absence of a glycemic threshold for the development of long-term complication: the perspective of the Diabetes Control and Complications Trial. *Diabetes*, 45: 1289-1298.

The DCCT Research Group (1995a). Adverse events and their association with treatment regimens in the Diabetes Control and Complications Trial. *Diabetes Care*, 18: 1415-1427.

The DCCT Research Group (1995b). The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the Diabetes Control and Complications Trial. *Diabetes*, 44: 968-983.

The DCCT Research Group (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *The New England Journal of Medicine*, 329: 977-986.

Jacobson A, Ryan CM, Cleary P et al. (2007). Long-Term Effect of Diabetes and Its Treatment on Cognitive Function. *The New England Journal of Medicine*, 356:1842-1852.

Musen G, Jacobson A, Ryan CM, Cleary PA et al. (2008). The Impact of Diabetes and its Treatment on Cognitive Function among Adolescents Who Participated in the DCCT. *Diabetes Care*, 31:1933-1938.