Fully Conditional Specification: Past, present and beyond

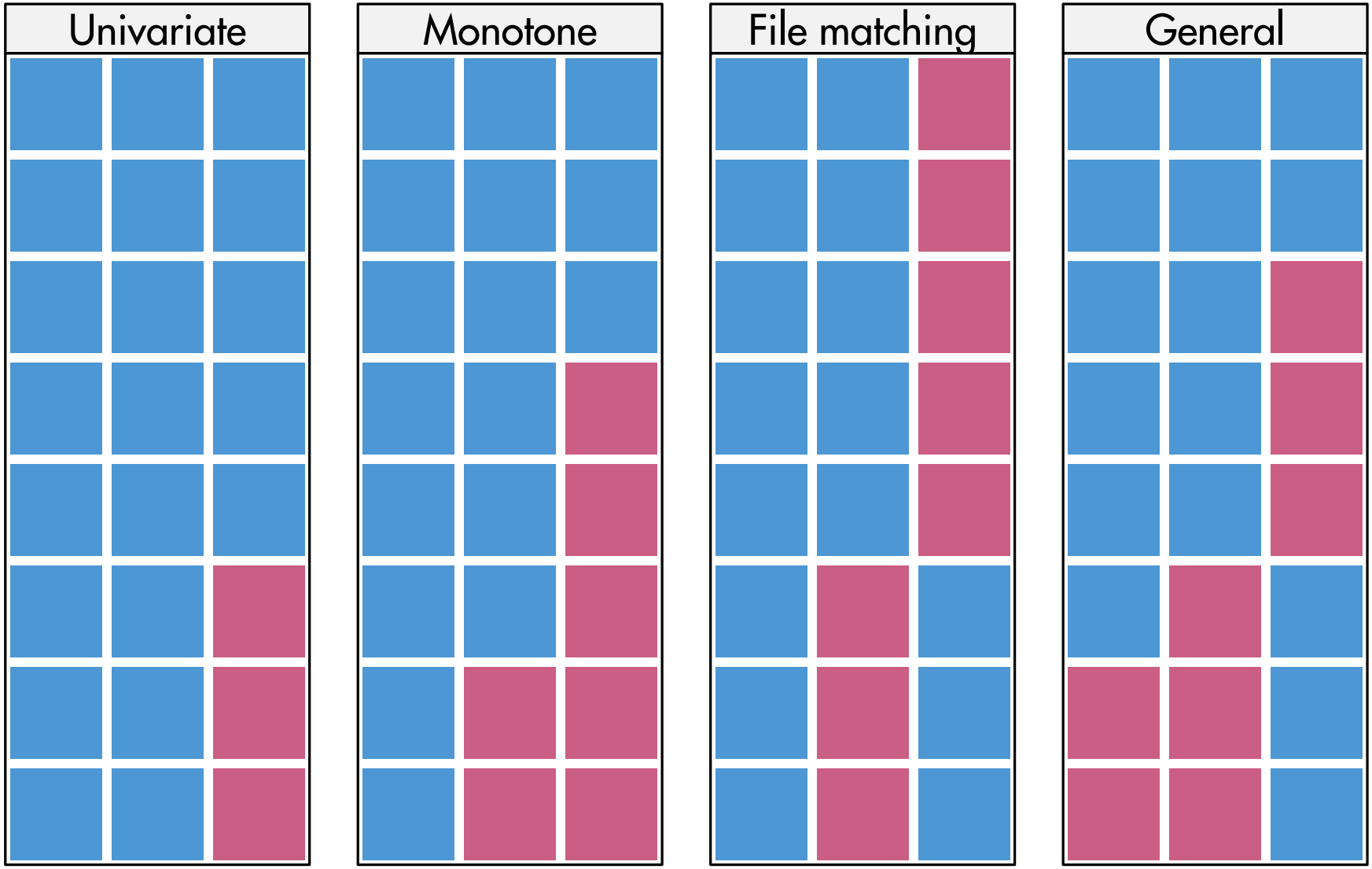
Stef van Buuren[1,2]

[1]Methodology and Statistics, FSBS, Utrecht University

[2]Netherlands Organization for Applied Scientific Research TNO, Leiden

Rennes, June 18, 2015

Missing data patterns



Problems in multivariate imputation

- Predictors themselves can be incomplete
- Mixed measurement levels
- Order of imputation can be meaningful
- Too many predictor variables
- Relations could be nonlinear
- Higher order interactions
- Impossible combinations

Three general strategies

- Monotone data imputation
- Joint modeling
- Fully conditional specification (FCS)

Joint Modeling (JM)

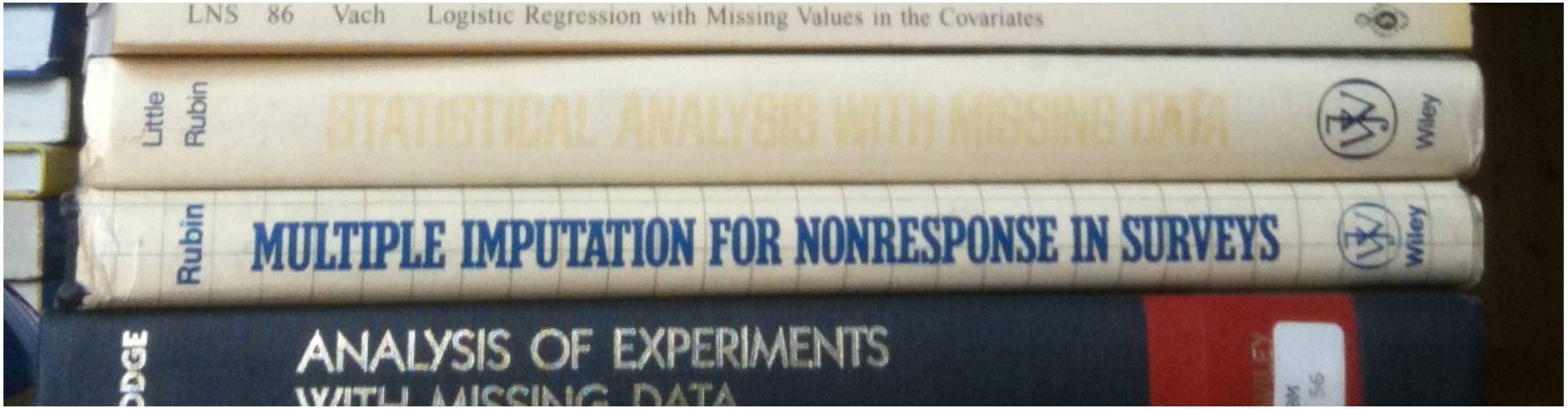
1. Specify joint model $P(Y, X, R)$
2. Derive $P(Y_{\text{mis}}|Y_{\text{obs}}, X, R)$
3. Use MCMC techniques to draw imputations $\dot{Y}_{\text{mis}}$

Fully Conditional Specification (FCS)

1. Specify $P(Y_{\text{mis}}|Y_{\text{obs}}, X, R)$
2. Use MCMC techniques to draw imputations $\dot{Y}_{\text{mis}}$

Past: the legacy

Rubin 1987


LNS 86 Vach Logistic Regression with Missing Values in the Covariates
Little
Rubin
STATISTICAL ANALYSIS WITH MISSING DATA
Wiley
Rubin
MULTIPLE IMPUTATION FOR NONRESPONSE IN SURVEYS
Wiley
ANALYSIS OF EXPERIMENTS

Rubin 1977 (published 2004, The American Statistician)

The Design of a General and Flexible System for Handling Nonresponse in Sample Surveys

Donald B. Rubin

1. INTRODUCTION

The general approach to nonresponse (missingness) in sur-

choose for imputation). How to obtain such a distribution will be discussed later. The question addressed in this section is: how

Rubin 1977 (published 2004, The American Statistician)

The masterplan

The current plan is to (1) model the respondent's data, (2) obtain for each nonrespondent, the conditional distribution of missing data (given observed data), as if he were a respondent (i.e., the predictive distribution under the model that says each nonrespondent is just like a respondent with the same values of observed variables), and (3) alter this distribution in various ways to allow for nonresponse bias that may change from nonrespondent to nonrespondent (e.g., the bias could change as type of nonresponse changes from refusal to not home or don't know).

Rubin 1977 (published 2004, The American Statistician)

Invents multiple imputation

Of course (1) imputing *one* value for missing datum can't be correct in general, and (2) in order to insert sensible values for a missing datum we must rely more or less on some model relating unobserved values to observed values. Hence, I see the best approach to be one where we can (1) insert more than one value for a missing datum, and (2) the inserted values reflect a variety of models for the dataset.

Rubin 1977 (published 2004, The American Statistician)

Invents monotone data imputation

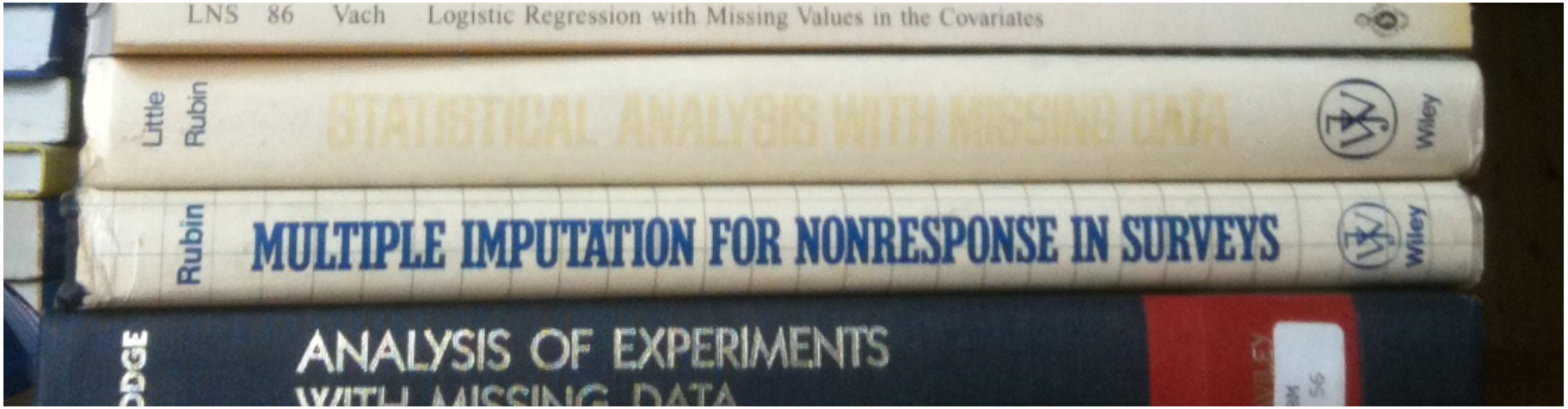
eral case. Let $Y = (Y_1, Y_2)$ where Y_1 represents essentially continuous variables and Y_2 represents categorical variables; more specifically, Y_1 given (Y_2, X) will be modeled as a multivariate normal linear regression and Y_2 given X will be modeled as a conditional contingency table (i.e., the joint distribution of (Y_2, X) can be represented by a contingency table). (Note:

Rubin 1977 (published 2004, The American Statistician)

Claim: Monotonicity is not essential

Display 2 for the data structure. This restriction is not essential to the ideas being presented, but it does simplify discussion and reduce the need for exotic computer programs.

Rubin 1987


LNS 86 Vach Logistic Regression with Missing Values in the Covariates
Little
Rubin
STATISTICAL ANALYSIS WITH MISSING DATA
Wiley
Rubin
MULTIPLE IMPUTATION FOR NONRESPONSE IN SURVEYS
Wiley
ANALYSIS OF EXPERIMENTS

One of the great academic growth industries

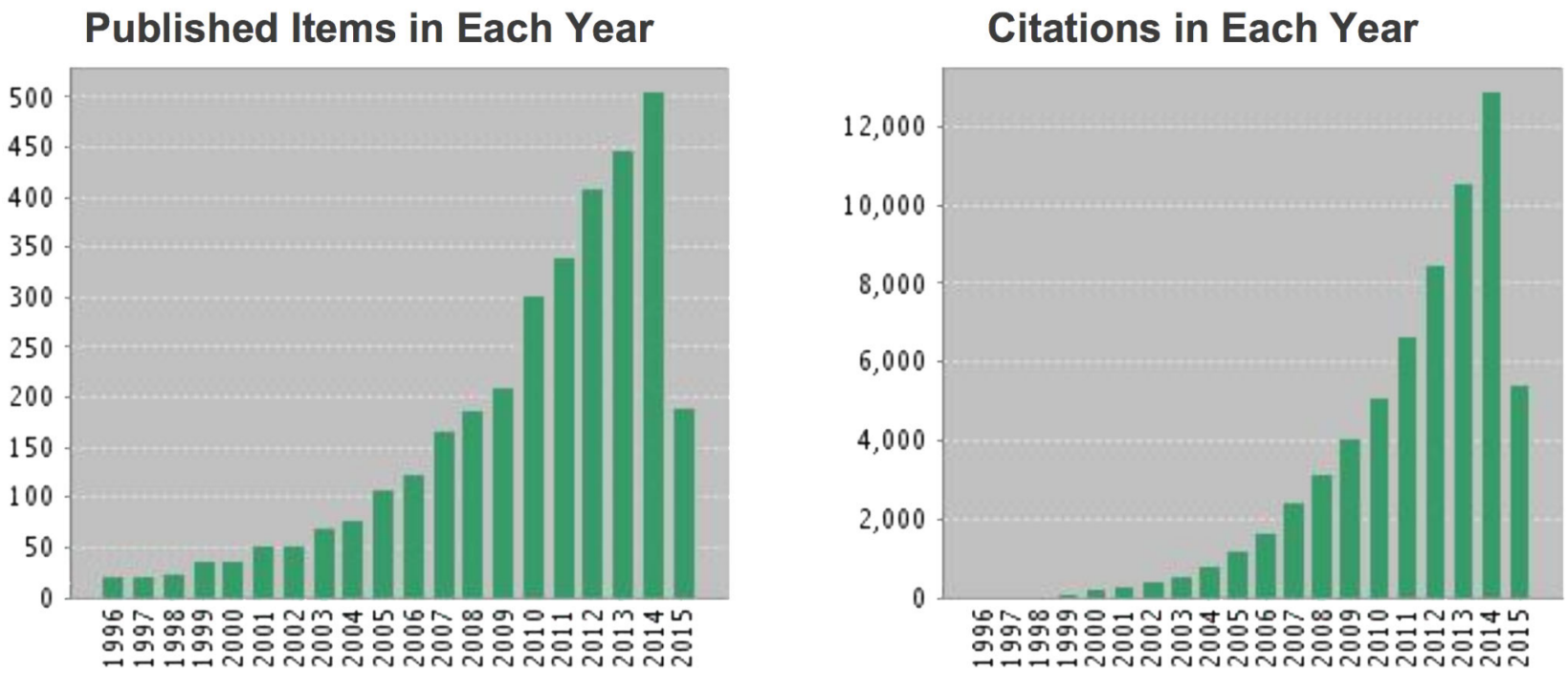
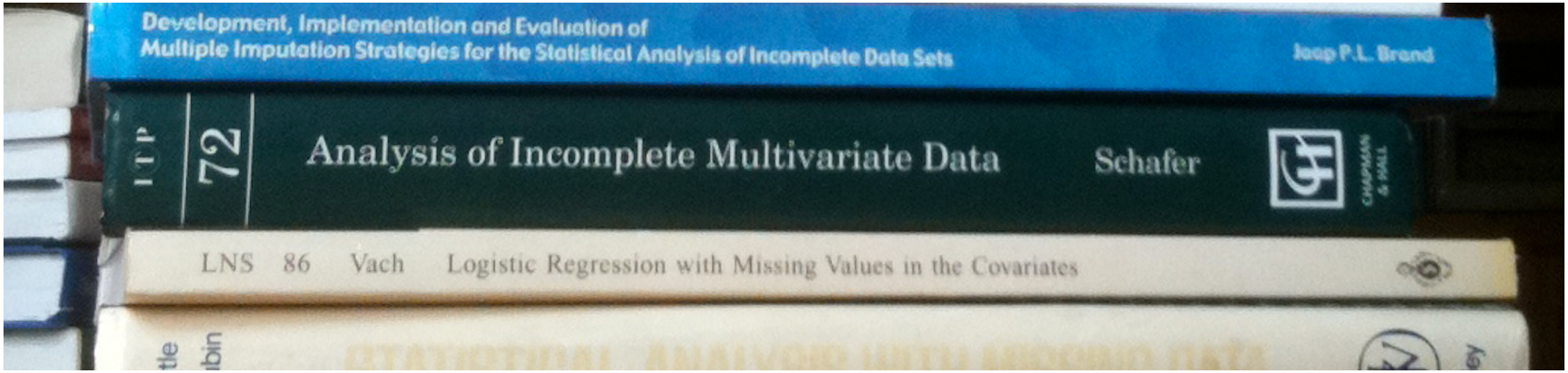


Figure 1: The number of published items and the number of citations that contain the keyword 'multiple imputation' over the last 20 years (Web of Science™ Citation Reports, retrieved on June 11 2015)

Schafer 1997



Joint Modeling (JM)

1. Specify joint model $P(Y, X, R)$
2. Derive $P(Y_{\text{mis}}|Y_{\text{obs}}, X, R)$
3. Use MCMC techniques to draw imputations $\dot{Y}_{\text{mis}}$

Joint modeling: Software

R/S Plus	`norm`, `cat`, `mix`, `pan`, `Amelia`
SAS	`proc MI`, `proc MIANALYZE`
STATA	`MI` command
Stand-alone	`Amelia`, `solas`, `norm`, `pan`

Joint Modeling: Pro's

- Yield correct statistical inference under the assumed JM
- Efficient parametrization (if the model fits)
- Known theoretical properties
- Works very well for parameters close to the center
- Many applications

Joint Modeling: Con's

- Lack of flexibility
- May lead to large models
- Can assume more than the complete data problem
- Can impute impossible data

Rubin & Schafer 1990, ASA

EFFICIENTLY CREATING MULTIPLE IMPUTATIONS FOR INCOMPLETE MULTIVARIATE NORMAL DATA

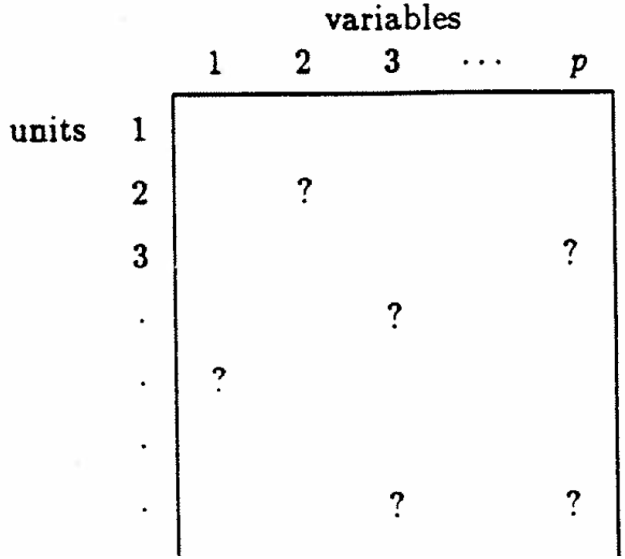
Donald B. Rubin and Joseph L. Schafer *

Joseph L. Schafer, Department of Statistics, 1 Oxford Street, Cambridge, MA 02138

Key Words: Factored likelihoods, Gibbs sampling, missing data, SIR algorithm, stochastic relaxation.

1 INTRODUCTION

The multivariate normal distribution is a common baseline model for many data analyses. When datasets have missing values, it may be reasonable to use the multivariate normal model to create multiple imputations (Rubin 1987); until now, however, no straightforward, general-purpose algorithm has been available to create proper multiple imputations under this model.



Rubin & Schafer 1990, ASA

Stochastic relaxation allows one to simulate a joint probability distribution by repeatedly drawing from conditional distributions. Let

$$Z = (Z_1, Z_2, \ldots, Z_J)$$

be a random variable with distribution $P(Z)$, where each Z_j is a random variable. Stochastic relaxation simulates values of Z by an iterative procedure. Given the value of Z at step t, say

$$Z^{(t)} = (Z_1^{(t)}, Z_2^{(t)}, \ldots, Z_J^{(t)}),$$

we obtain the value of Z at step $t+1$,

$$Z^{(t+1)} = (Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_J^{(t+1)}),$$

by successively drawing from the distributions

$$\begin{aligned} Z_1^{(t+1)} &\sim P(Z_1 \mid Z_2^{(t)}, Z_3^{(t)}, \ldots, Z_J^{(t)}) \\ Z_2^{(t+1)} &\sim P(Z_2 \mid Z_1^{(t+1)}, Z_3^{(t)}, \ldots, Z_J^{(t)}) \\ &\vdots \\ Z_J^{(t+1)} &\sim P(Z_J \mid Z_1^{(t+1)}, Z_2^{(t+1)}, \ldots, Z_{J-1}^{(t+1)}) \end{aligned} \tag{5}$$

in a slight abuse of notation. In other words, we draw from the conditional distributions of Z_1, Z_2, up to Z_J, conditioning each time on the most recently drawn values of all other variables. After the

Fully conditional specification

Journal of Statistical Computation and Simulation
Vol. 76, No. 12, December 2006, 1049–1064

Taylor & Francis
Taylor & Francis Group

Fully conditional specification in multivariate imputation

S. VAN BUUREN*†‡, J. P. L. BRAND§, C. G. M. GROOTHUIS-OUDSHOORN¶
and D. B. RUBIN‖

†Department of Statistics, TNO Quality of Life, Wassenaarseweg 56,
P.O. Box 2215, 2301 CE Leiden, The Netherlands
‡University of Utrecht, The Netherlands
§Pennington Biomedical Research Center, Baton Rouge, LA 70808, USA
¶Roessingh Research and Development, Enschede, The Netherlands
‖Harvard University, Cambridge, MA 02138, USA

(Received 15 January 2004; in final form 22 March 2005)

Names for the idea

- stochastic relaxation (Kennickell 1991)
- variable-by-variable imputation (Brand 1999)
- regression switching (van Buuren et al. 1999)
- chained equations (MICE) (van Buuren and Oudshoorn 2000)
- sequential regressions (Raghunathan et al. 2001)
- ordered pseudo-Gibbs sampler (Heckerman et al. 2001)
- partially incompatible MCMC (Rubin 2003)
- iterated univariate imputation (Gelman 2004)
- fully conditional specification (FCS) (van Buuren 2006)
- partially incompatible Gibbs Sampler (PIGS) (Rubin 2014)

Fully Conditional Specification (FCS)

1. Specify $P(Y_{\text{mis}}|Y_{\text{obs}}, X, R)$
2. Use MCMC techniques to draw imputations $\dot{Y}_{\text{mis}}$

Multivariate Imputation by Chained Equations (MICE)

- MICE algorithm
- Specify imputation model for each incomplete column
- Fill in starting imputations
- And iterate
- Model: Fully Conditional Specification (FCS)

Fully Conditional Specification: Con's

- Theoretical properties only known in special cases
- Cannot use computational shortcuts, like sweep-operator
- Joint distribution may not exist (incompatibility)

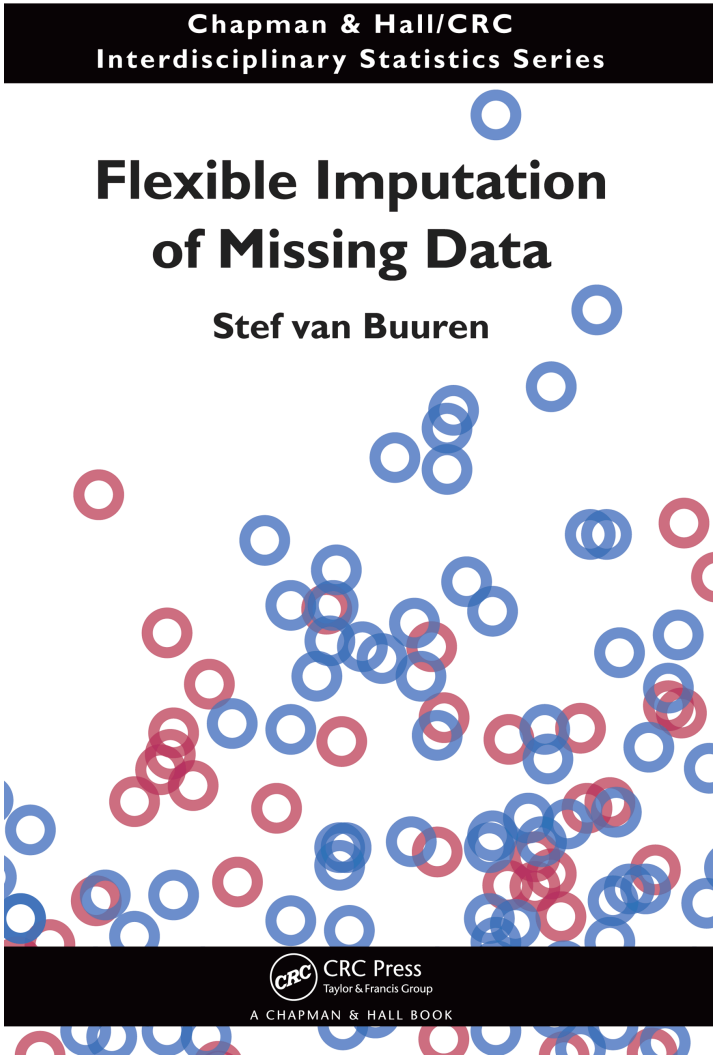
Fully Conditional Specification: Pro's

- Easy and flexible
- Imputes close to the data, prevents impossible data
- Subset selection of predictors
- Modular, can preserve valuable work
- Works well, both in simulations and practice

Fully Conditional Specification (FCS): Software

R	mice, transcan, mi, VIM, baboon
SPSS V17	procedure multiple imputation
SAS	IVEware, SAS 9.3
STATA	ice command, multiple imputation command
Stand-alone	Solas, Mplus

Flexible Imputation of Missing Data (FIMD)



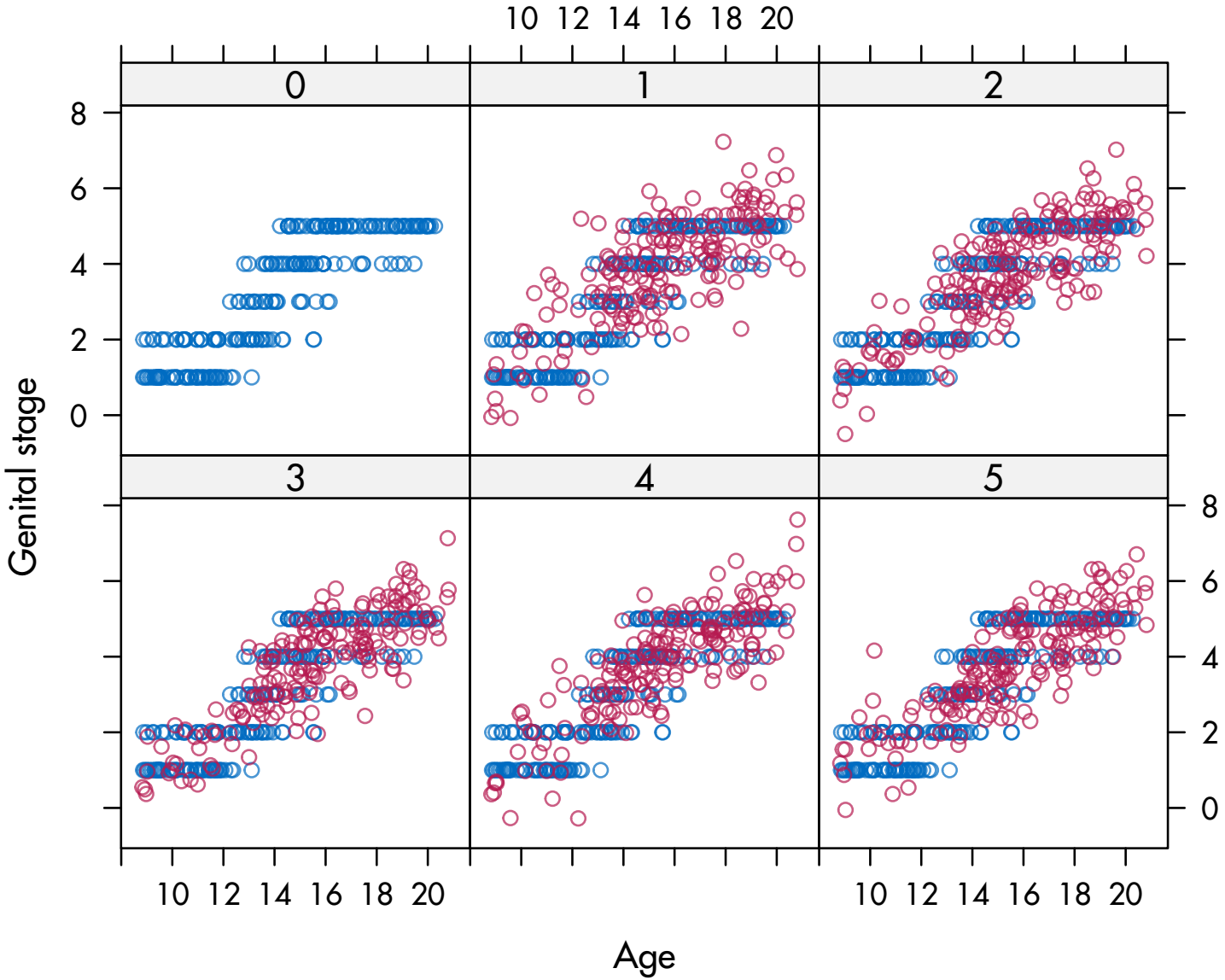
Joint model vs Fully conditional specification

- Fourth Dutch Growth Study 1997
- 22000 children between ages 0 and 21
- Tanner maturation stages
- Boys 8–21 years
- Genital development (5 stages)
- 42% missing data
- How does the probability per stage change with age?

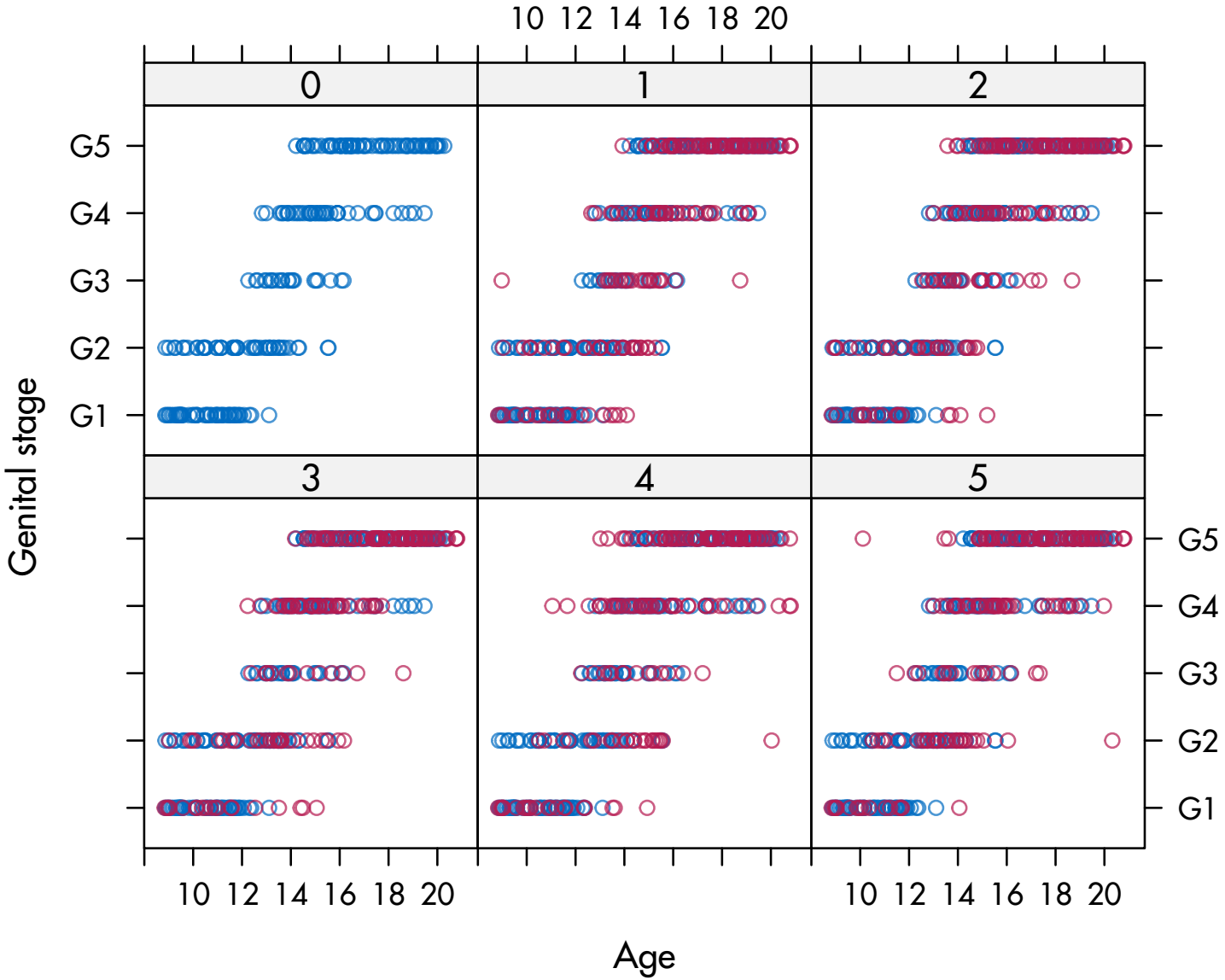
Imputation methods

- JM: multivariate normal
- JM: rounded
- FCS: predictive mean matching
- FCS: proportional odds model

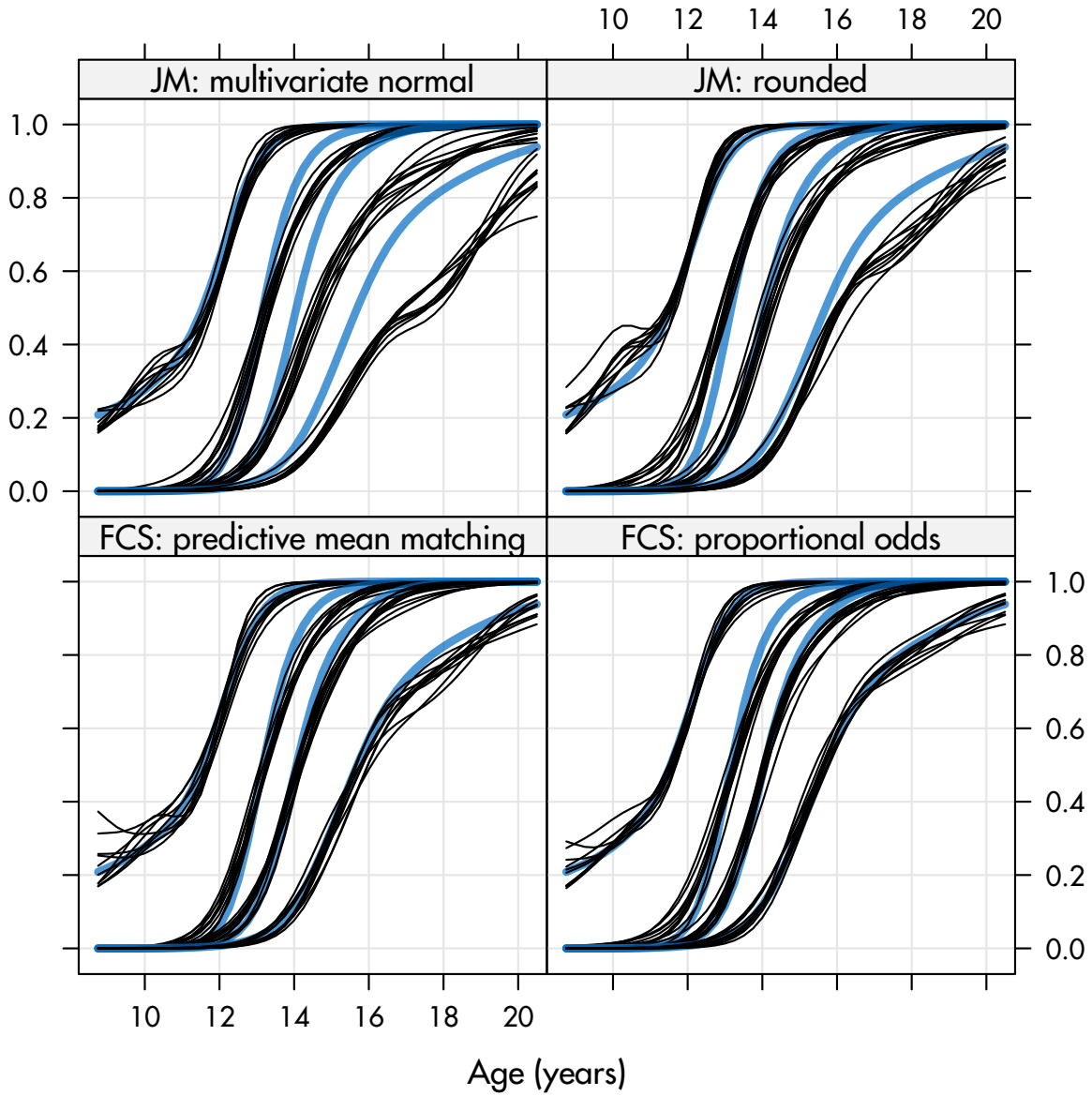
JM: Multivariate normal model



FCS: Proportional Odds model



JM vs FCS


JM: multivariate normal
JM: rounded
FCS: predictive mean matching
FCS: proportional odds
1.0
0.8
0.6
0.4
0.2
0.0
10
12
14
16
18
20
Age (years)

Incompatibility

Compatibility of conditionals is a theoretical requirement for the Gibbs sampler

What happens if the model is clearly incompatible?

Simulation setup

- Generate bivariate normal data, correlation 0.6
- Scientific interest on β in $Y_1 = \alpha + \beta Y_2 + \varepsilon$
- Generate 50% missing per variable, 75% incomplete cases (MCAR): three mechanisms

Three imputation models

MI compatible bivariate linear

- $Y_1^* \sim N(\phi+\theta\, Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma+\delta\, Y_1, \sigma_1^2)$

$E(\theta/\sigma_2^2) = E(\delta/\sigma_1^2)$

MI incompatible quadratic

- $Y_1^* \sim N(\phi+\theta\, Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma+\delta\, Y_1^2, \sigma_1^2)$

$E(\theta/\sigma_2^2) \neq E(\delta/\sigma_1^2)$

MI incompatible log

- $Y_1^* \sim N(\phi+\theta\, Y_2, \sigma_2^2)$
- $Y_2^* \sim N(\gamma+\delta\, \log(Y_1), \sigma_1^2)$

$E(\theta/\sigma_2^2) \neq E(\delta/\sigma_1^2)$

Incomplete data - Stef van Buuren April 8, 2010 67

Simulation setup

- Bivariate normal data, correlation 0.6
- Generate 50% missing per variable, 75% incomplete cases (MCAR)
- 500 replications
- Scientific interest on β in $Y = \alpha + \beta X + \varepsilon$

- MICE implementation with a derived variable
- $Y^* \sim N(\phi + \theta X, \sigma_X^2)$
- $Z = \log(Y)$ passive imputation
- $X^* \sim N(\gamma + \delta Z, \sigma_Y^2)$

Mechanism	Missing data method	FMI	E(*b*)	Cov
	Theoretical values		0.600	95
MARRIGHT	Complete case analysis		0.597	93
	MI compatible linear	0.63	0.595	95
	MI incompatible quadratic	0.63	0.589	95
	MI incompatible log	0.64	0.582	95
MARMID	Complete case analysis		0.678	79
	MI compatible linear	0.75	0.613	94
	MI incompatible quadratic	0.75	0.601	94
	MI incompatible log	0.75	0.579	94
MARTAIL	Complete case analysis		0.556	78
	MI compatible linear	0.50	0.596	94
	MI incompatible quadratic	0.50	0.590	94
	MI incompatible log	0.50	0.590	95

Incomplete data - Stef van Buuren April 8, 2010 69

Incompatibility - conclusion

Compatibility of conditionals is a theoretical requirement for the Gibbs sampler

Unclear what exactly happens when the conditions is not met

- Gibbs sampler may not converge
- results could depend on sequence of variables

- MICE appears to be robust against incompatibility, at least in the cases studied
- The incompatible model was superior to CCA for MARMID and MARTAIL mechanisms
- More work is needed

Masterplan step 3

- Now: between-variation is driven by predictive uncertainty of only one model
- Masterplan 3: Multiple models, each reflection a potentially different cause for the missing data
- MNAR: Make perturbations on the model to reflect missing data cause
- Examples
  - δ-adjustment in Van Buuren et al (1999)
  - Ofer Harel (2009) Methods for two types of missing data
  - Shahab Jolani (2012): Random indicator method for MNAR modelling

Which predictors?

1. Include all variables that appear in the complete-data model
2. In addition, include the variables that are related to the nonresponse
3. In addition, include variables that explain a considerable amount of variance
4. Remove from the variables selected in steps 2 and 3 those variables that have too many missing values within the subgroup of incomplete cases.

Function `quickpred()` and `flux()`

Model specification

- Now: Congeniality/compatibility is assumed rather than diagnosed
- Future: Ways to diagnose uncongeniality and incompatibility in practice
- Future: Ways to attack these problems
- Liu, Gelman et al. 2011: Propose conditions D1-D5 under which an FCS model will approximate a joint model
- Bartlett, Seaman et al. 2014: Write substantive model and imputation model as a joint model (SMC-FCS)
- Vink, Van Buuren 2013: Adapt the imputation method, e.g. for quadratic relations
- Xie, Meng 2015: Inference is valid when the imputer's model is more saturated (richer) than the substantive model

Comparability of simulation research

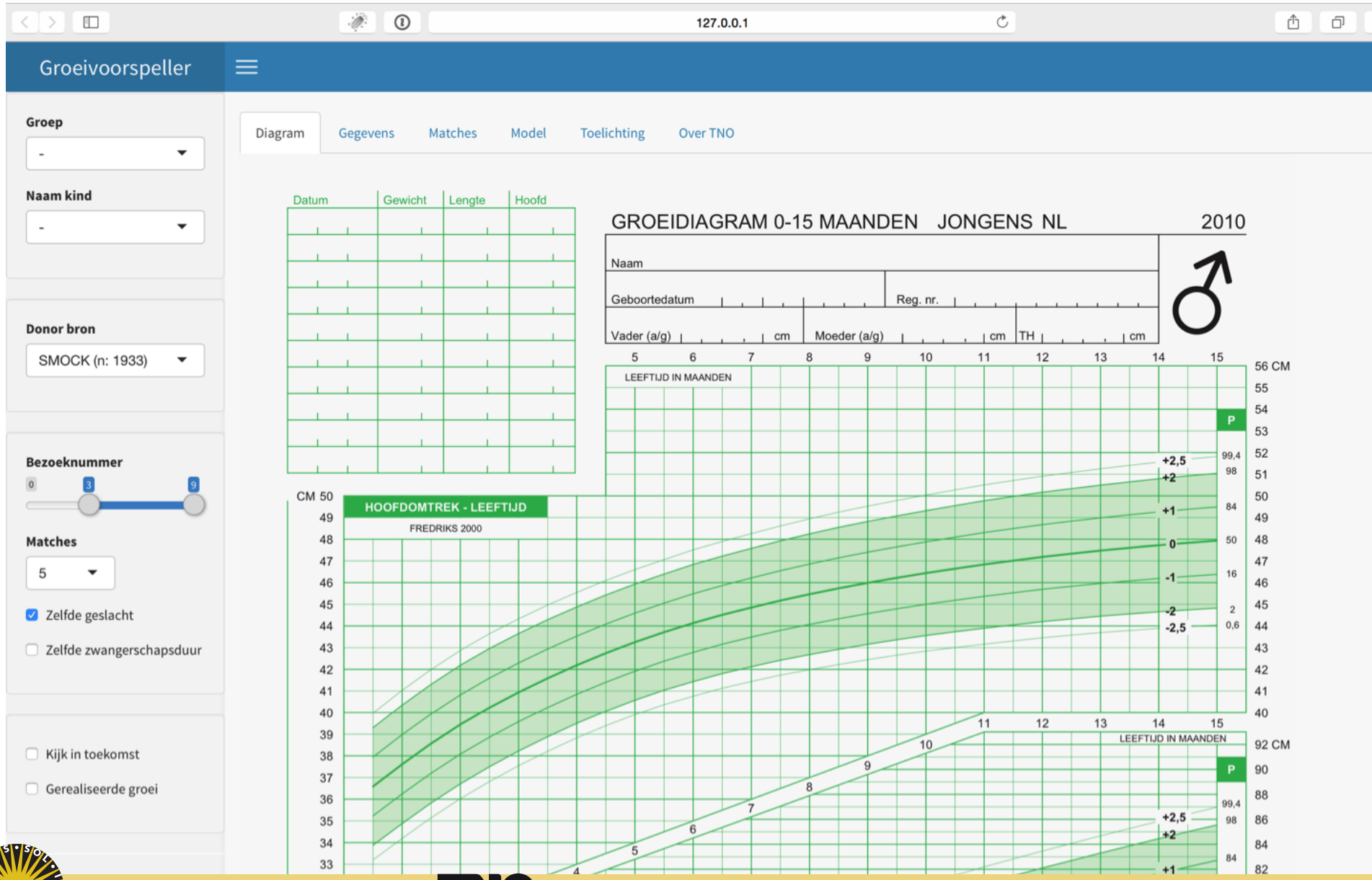
- About 10-20 papers per year about new imputation methods
- Validated in ways that makes it difficult to compare
  - Complete-data models differ
  - Missing data mechanisms differ
  - Proportions missing data differ
  - Evaluation criteria differ
  - Ratio sampling uncertainty / missing data uncertainty differ
- How to compare these new methods to existing?
- Solution: Benchmark platform

Individual causal effects

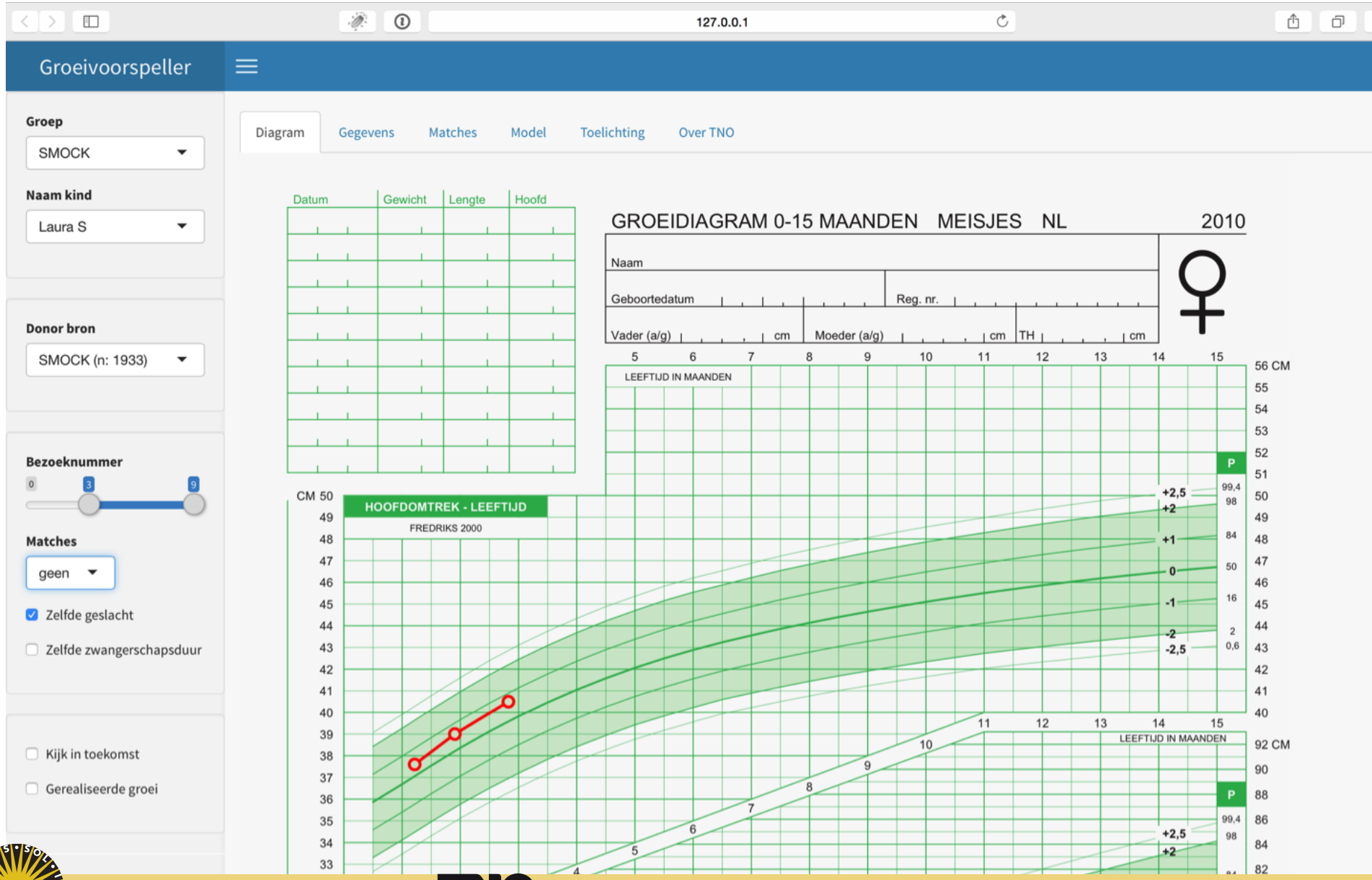
Individual causal effects

- Unit-level causal effect: $Y_i(1) - Y_i(0)$
- We can observe either $Y_i(1)$ or $Y_i(0)$, but never both
- Multiply-impute unobserved outcome
- We need assumptions about $r(Y(1), Y(0))$

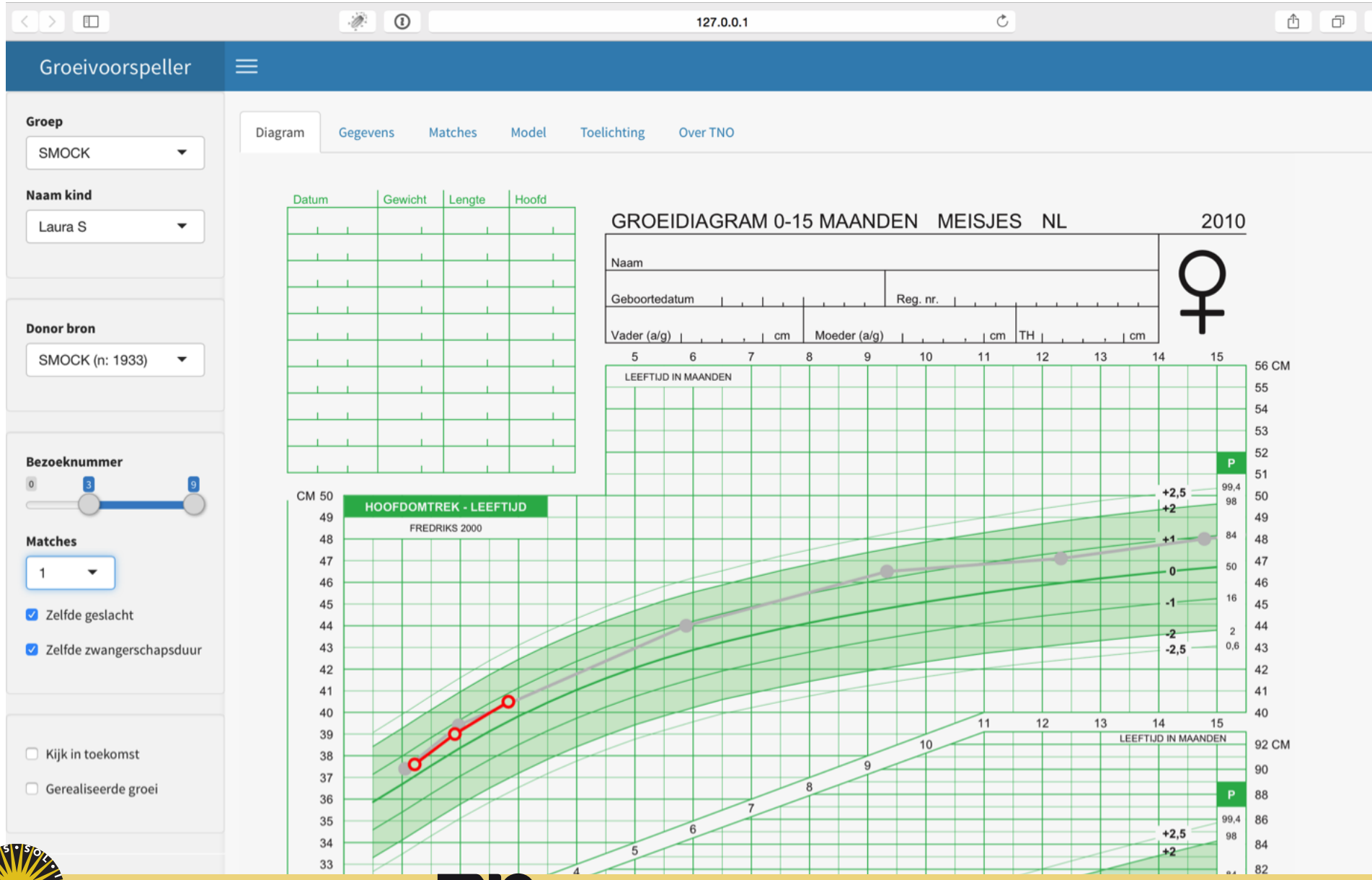
Individual causal effects - curve matching



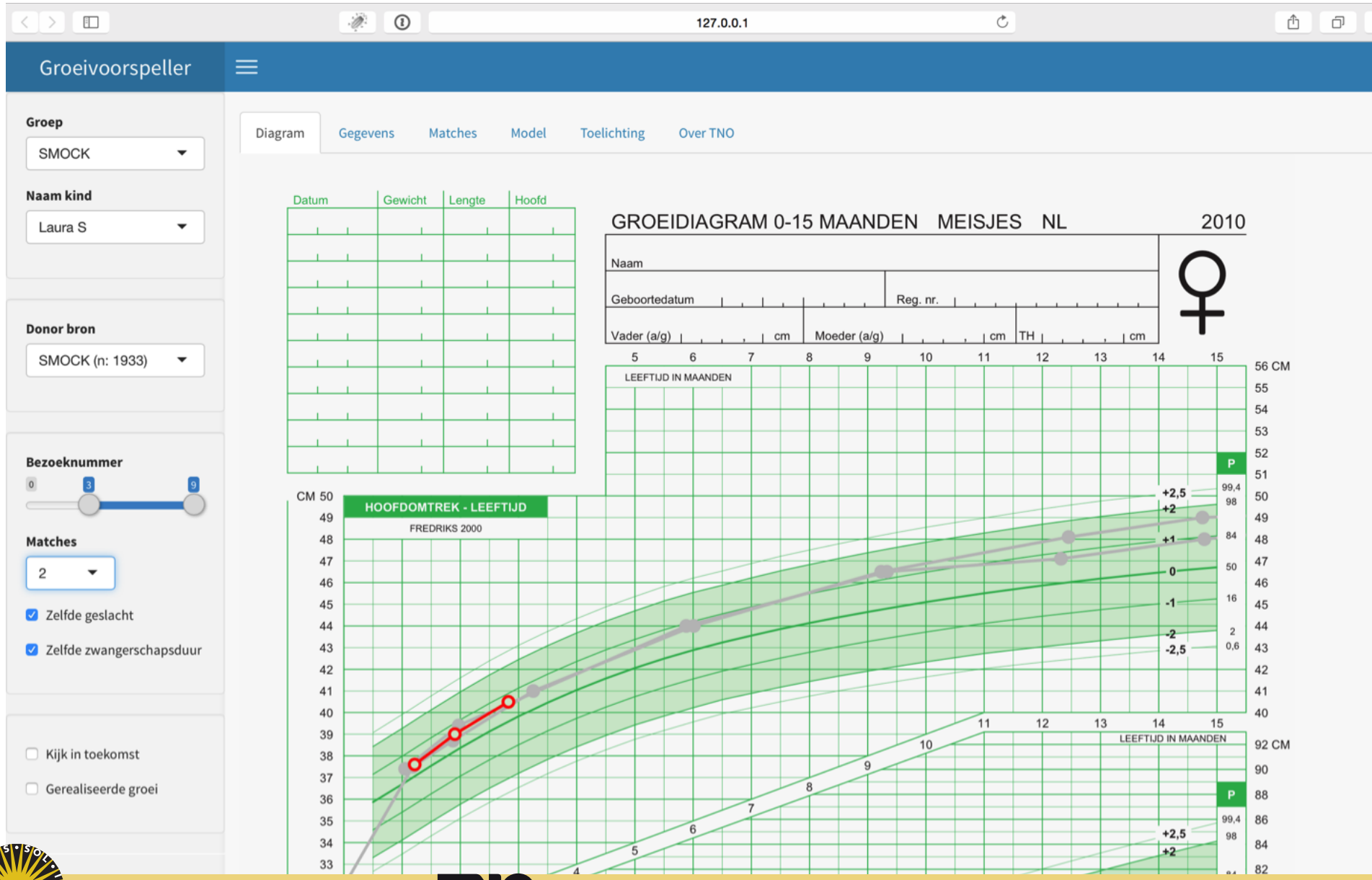
Individual causal effects - curve matching



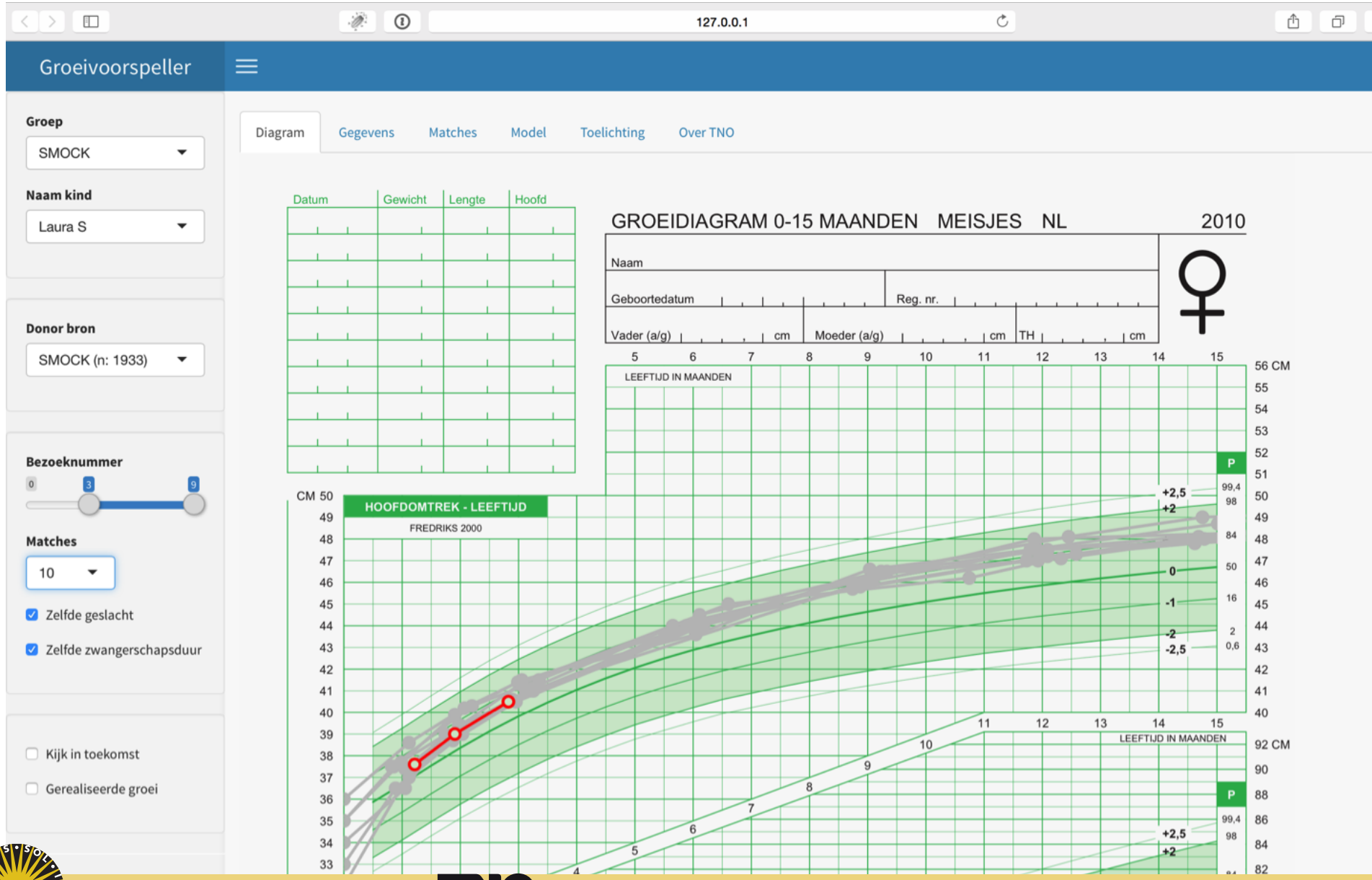
Individual causal effects - curve matching



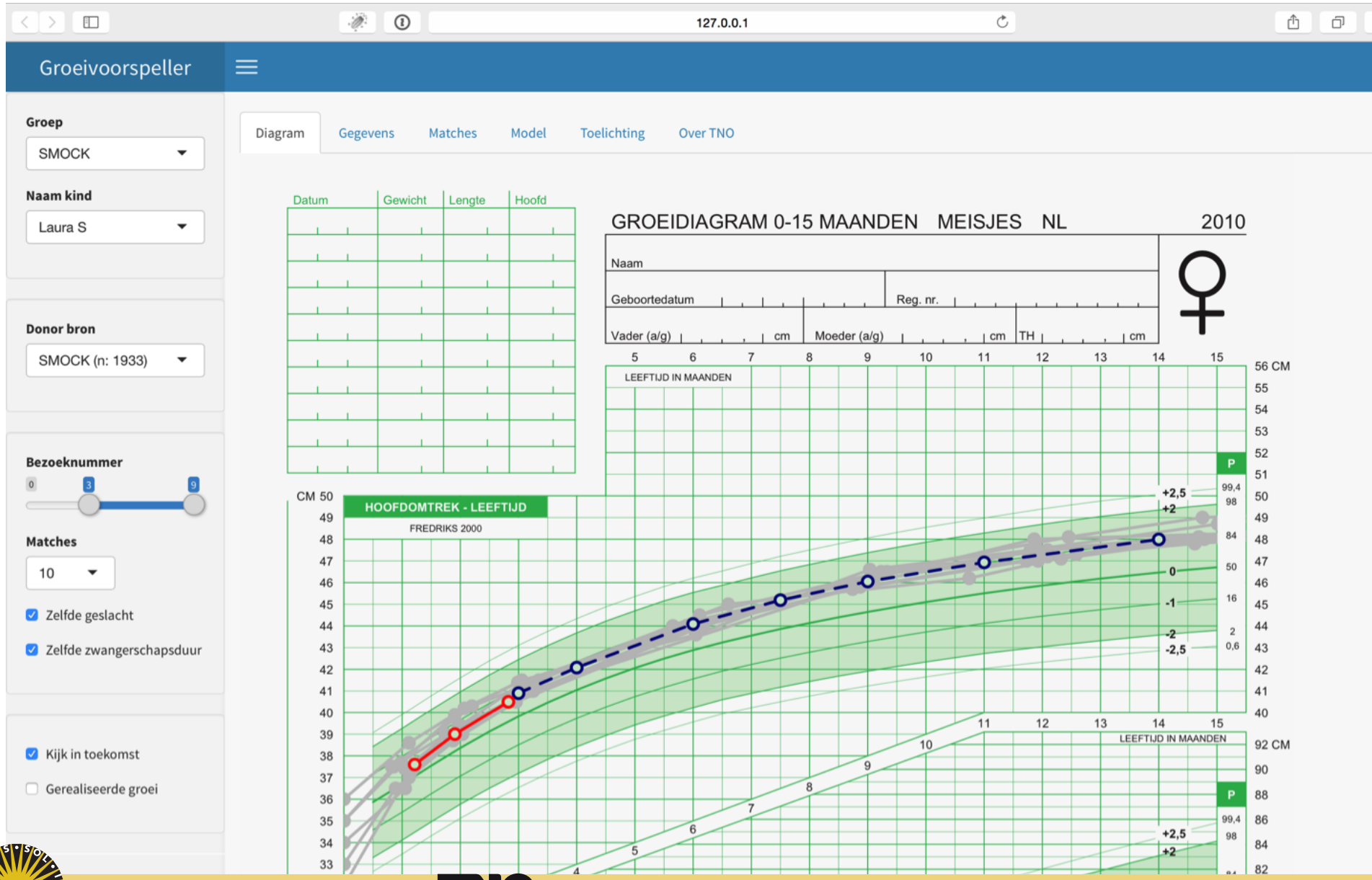
Individual causal effects - curve matching



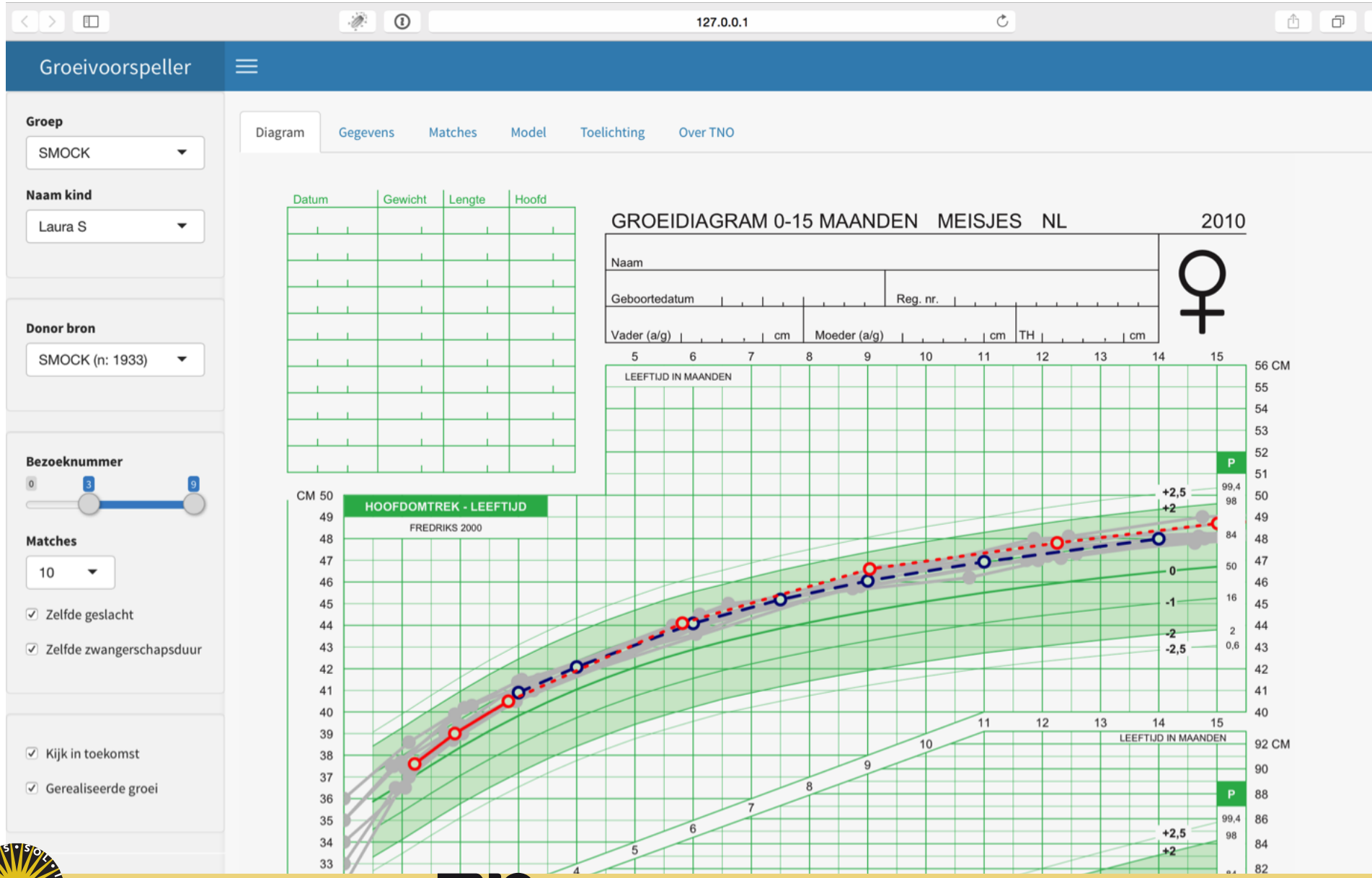
Individual causal effects - curve matching



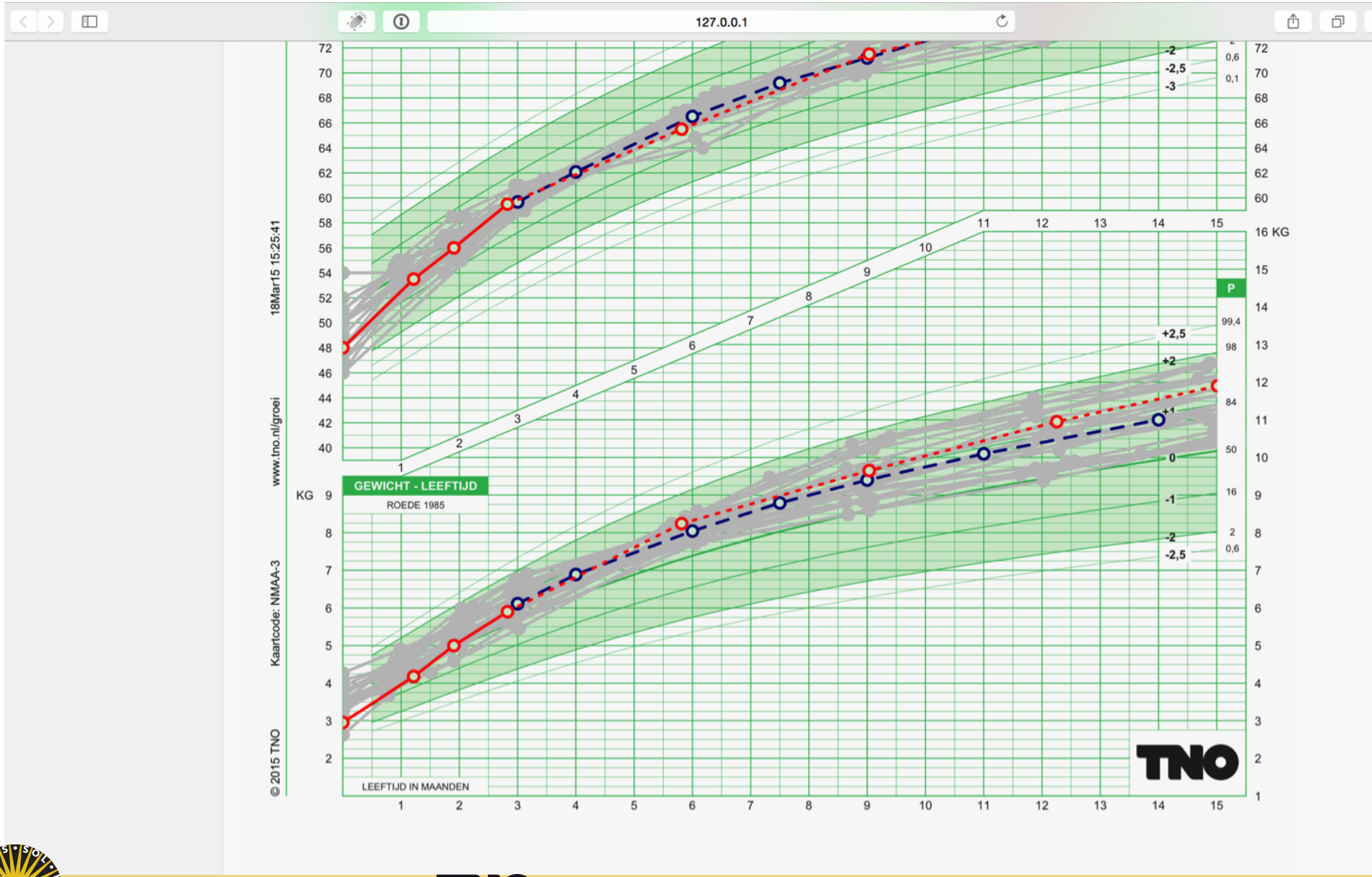
Individual causal effects - curve matching



Individual causal effects - curve matching



Individual causal effects - curve matching



Conclusion

- Multiple imputation is now extremely successful
- Both JM and FCS are powerful methodologies to impute multivariate data
- It is an open methodology, many ideas remain to be explored