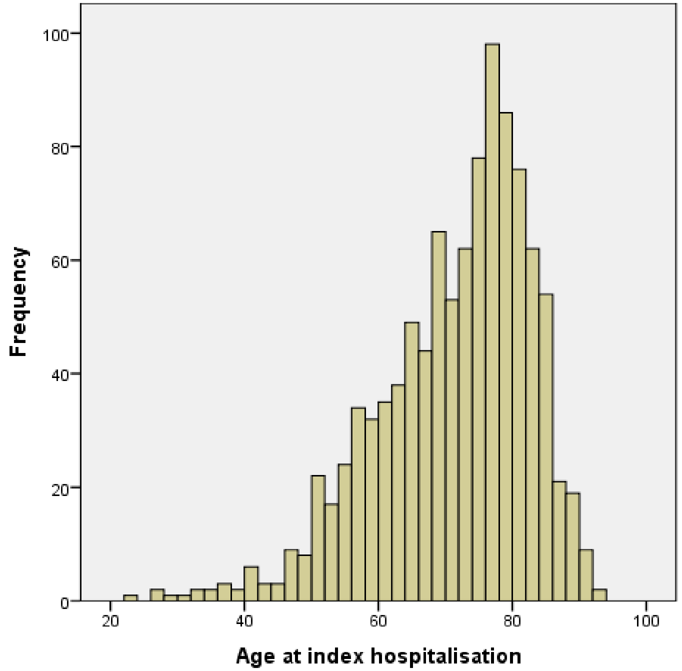
Visualizing data: the often neglected first step

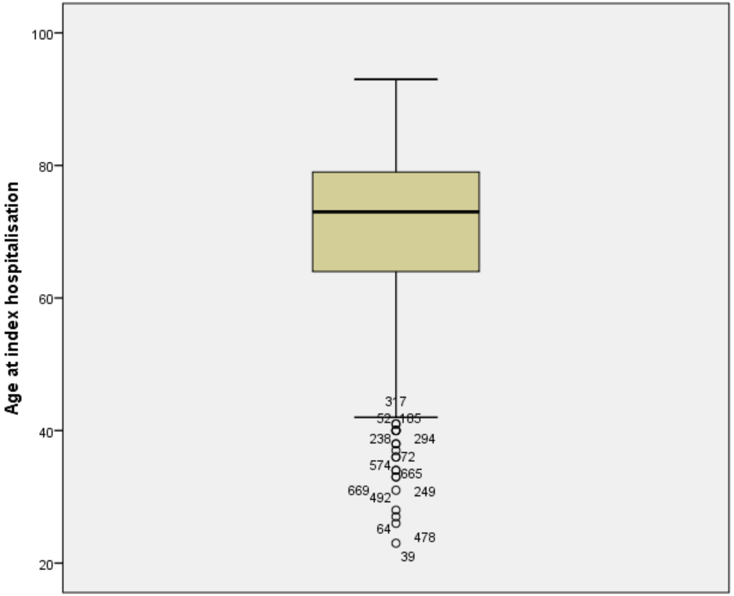
Dr. Douwe Postmus (Unit Medical Statistics, Dept Epidemiology)

university of groningen

University Medical Center Groningen

Data cleaning and validation

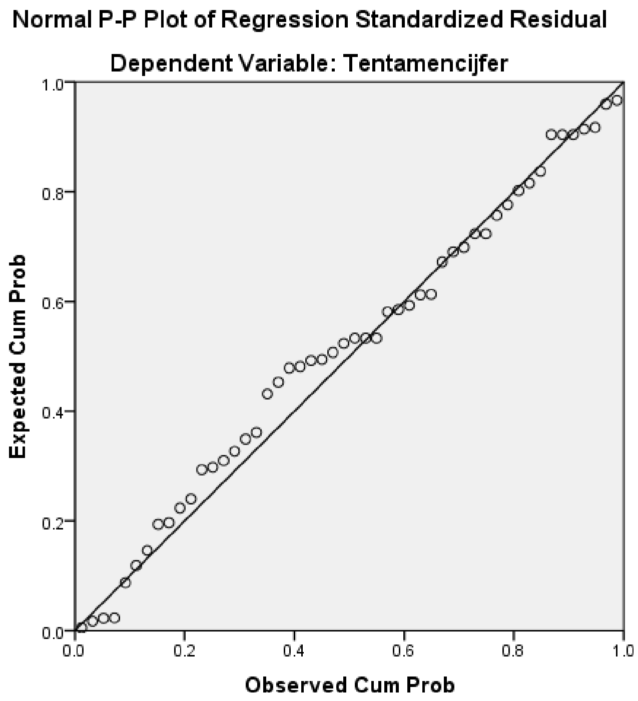


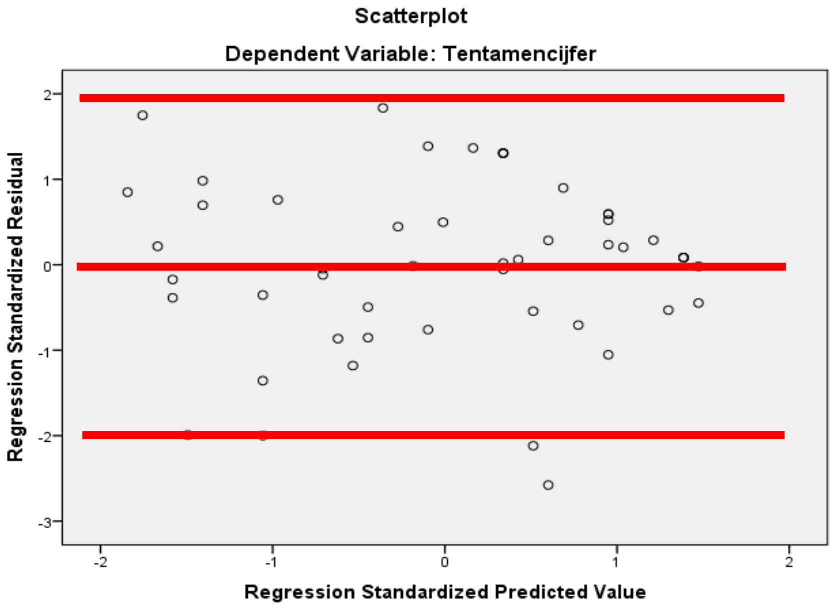


university of groningen

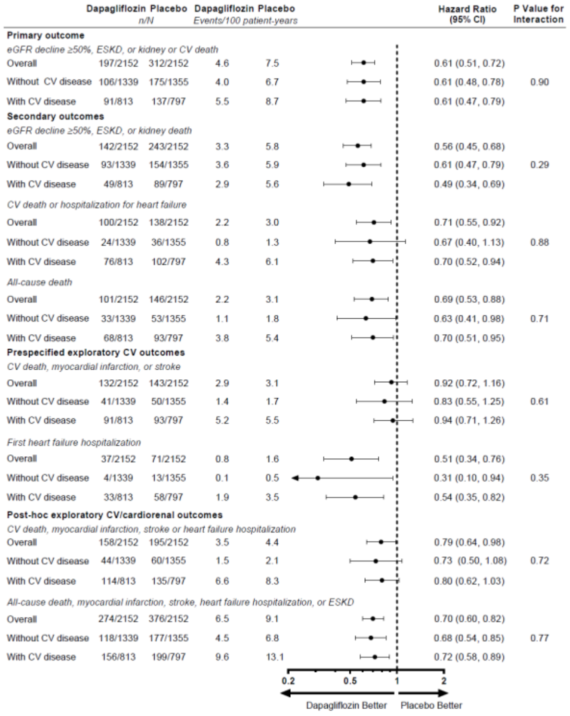
University Medical Center Groningen

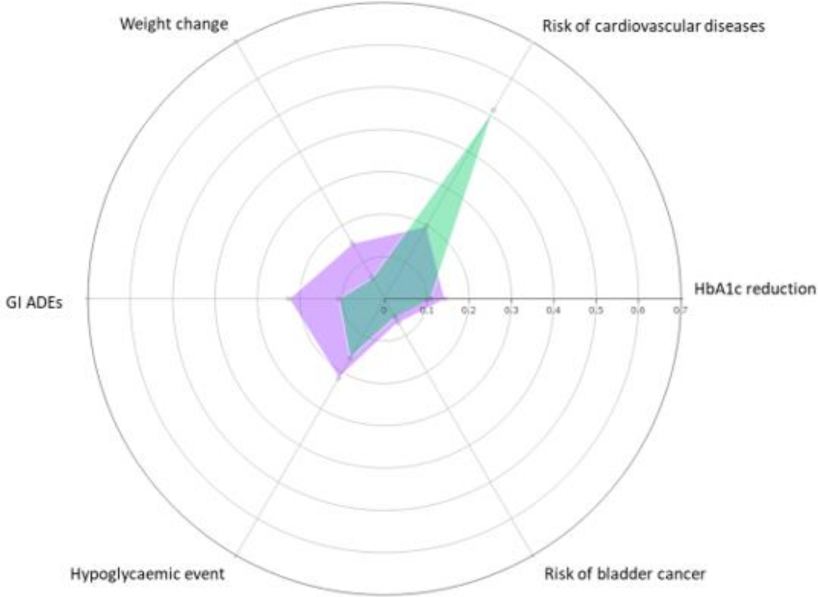
Post-analysis: goodness of fit





Post-analysis: results visualization





Today

- Use of graphical displays to summarize the patterns that are present in the data

- Helps with interpreting the results of a subsequent statistical test

- Supports statistical model building

Example 1: How does a woman's behavior during pregnancy affect the infant's birth weight?

* These data come from Appendix 1 of Hosmer and Lemeshow (1989), and were collected at Baystate Medical Center, Springfield MA, during 1986.

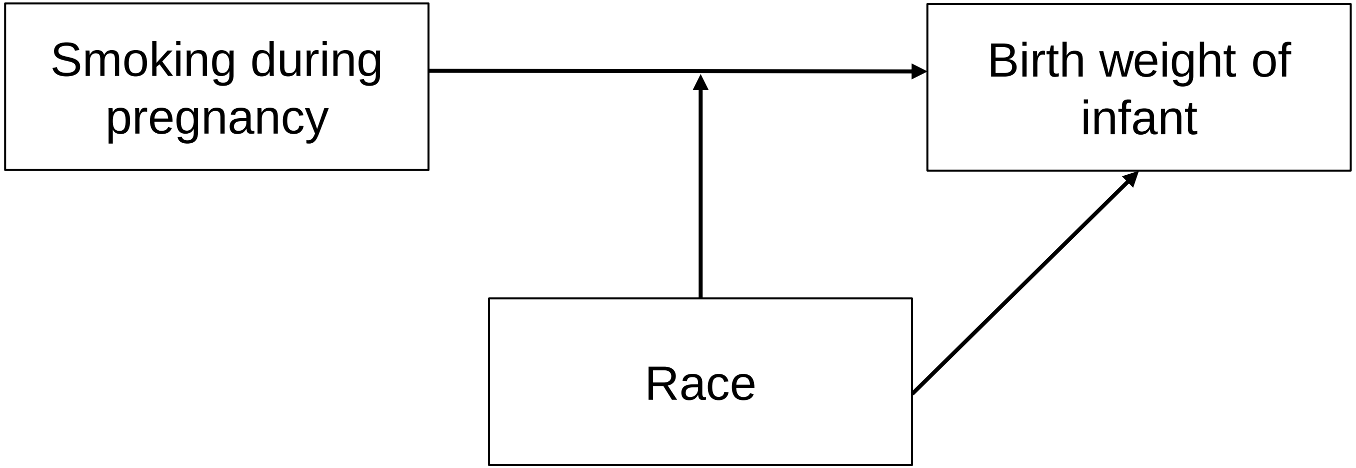
* Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

Columns	Variable	Abbreviation
2-4	Identification Code	ID
10	Low Birth Weight (0 = Birth Weight ge 2500g, 1 = Birth Weight < 2500g)	LBW
17-18	Age of the Mother in Years	AGE
23-25	Weight in Pounds at the Last Menstrual Period	LWT
32	Race (1 = White, 2 = Black, 3 = Other)	RACE
40	Smoking Status During Pregnancy (1 = Yes, 0 = No)	SMOKE
48	History of Premature Labor (0 = None, 1 = One, etc.)	PTL
55	History of Hypertension (1 = Yes, 0 = No)	HYPER
61	Presence of Uterine Irritability (1 = Yes, 0 = No)	URIRR
67	Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	PVFT
73-76	Birth Weight in Grams	BWT

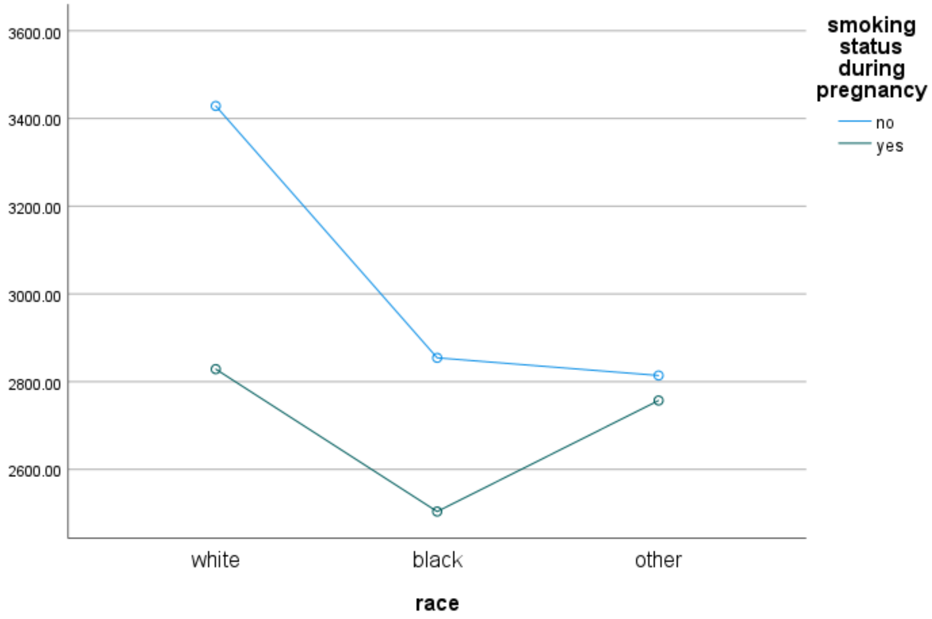
university of groningen

University Medical Center Groningen

Conceptual model



Interaction plot



In linear regression, the infant's mean birth weight is expressed as a linear function of the independent variables (regression equation)

Interaction plot: graphical display of the means for each combination of the levels of two categorical variables

ANOVA table

Tests of Between-Subjects Effects

Dependent Variable: birth weight in grams

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	14439242.57[a]	5	2887848.514	6.183	<.001
Intercept	965464191.55	1	965464191.55	2066.969	<.001
race	5818289.482	2	2909144.741	6.228	.002
smoke	3318053.571	1	3318053.571	7.104	.008
race * smoke	2097537.495	2	1048768.747	2.245	.109
Error	85477810.075	183	467091.858		
Total	1738735950.0	189			
Corrected Total	99917052.646	188			

a. R Squared = .145 (Adjusted R Squared = .121)

Post hoc tests for race

Multiple Comparisons

Dependent Variable: birth weight in grams

Bonferroni

(I) race	(J) race	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
white	black	384.0473*	151.09802	.036	18.9661	749.1285
	other	299.7247*	108.79826	.019	36.8476	562.6017
black	white	-384.0473*	151.09802	.036	-749.1285	-18.9661
	other	-84.3226	157.91324	1.000	-465.8707	297.2254
other	white	-299.7247*	108.79826	.019	-562.6017	-36.8476
	black	84.3226	157.91324	1.000	-297.2254	465.8707

Based on observed means.
The error term is Mean Square(Error) = 467091.858.

*. The mean difference is significant at the 0.05 level.

university of groningen

University Medical Center Groningen

Example 2: predicting the 10-year risk of coronary heart disease (CHD)

556 Biometrical Journal **57** (2015) 4, 556–570 DOI: 10.1002/bimj.201300260

Graphical assessment of incremental value of novel markers in prediction models: From statistical to decision analytical perspectives

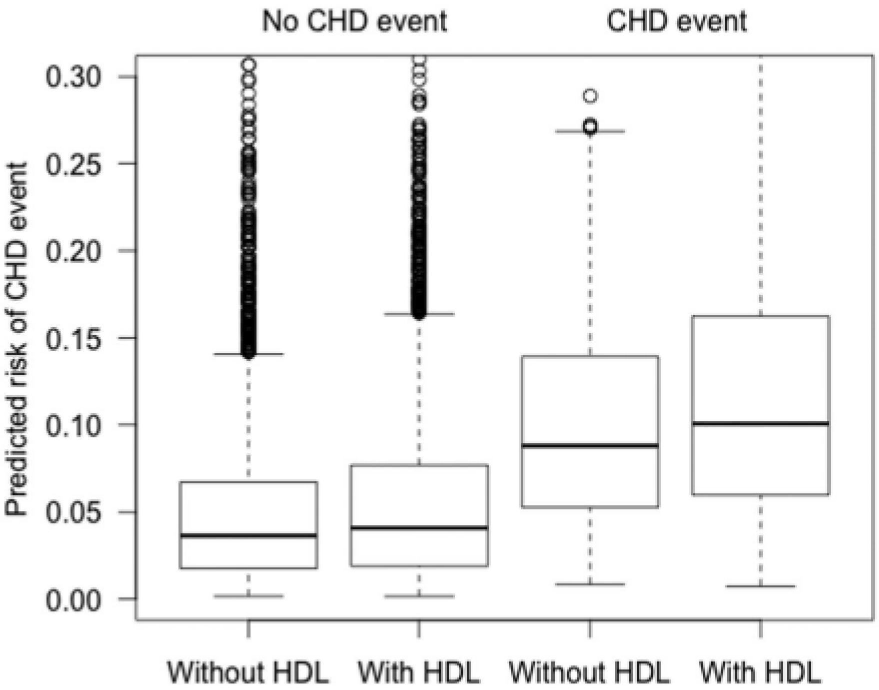
Ewout W. Steyerberg[*,1], **Moniek M. Vedder**[1], **Maarten J. G. Leening**[2,3], **Douwe Postmus**[4], **Ralph B. D'Agostino Sr.**[5], **Ben Van Calster**[1,6], and **Michael J. Pencina**[7]

[1] Department of Public Health, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[2] Department of Epidemiology, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[3] Department of Cardiology, Erasmus MC: University Medical Center Rotterdam, Rotterdam, The Netherlands
[4] Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[5] Framingham Heart Study, Framingham, MA, USA
[6] Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[7] Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, NC, USA

Received 5 November 2013; revised 24 April 2014; accepted 25 May 2014

- Reference model: multivariable logistic model with sex, diabetes, and smoking as dichotomous predictors and age, systolic blood pressure, and total cholesterol as continuous predictors
- Does adding HDL cholesterol to an existing model improve risk prediction?
- Analysis based on 3264 participants from the Framingham Heart Study aged 30 – 74 years
- A total of 183 individuals developed CHD (5.6% 10-year cumulative incidence)

Box plots stratified by CHD status - IDI

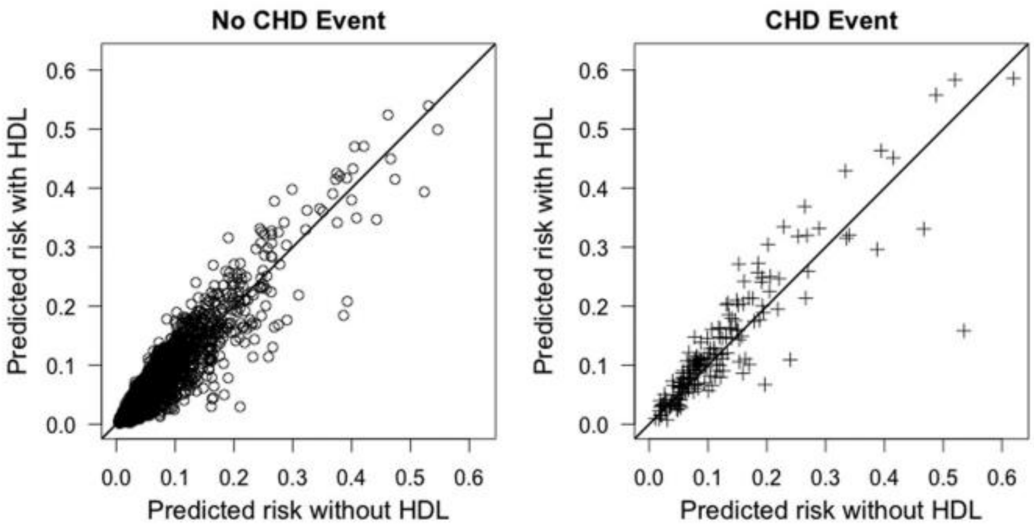


Discrimination slope = difference in mean predicted risks for those with and without the event

- Without HDL: 6.29%
- With HDL: 7.14%

Integrated discrimination index (IDI) = difference in discrimination slope
7.14 – 6.29 = 0.85%

Reclassification graphs - cNRI



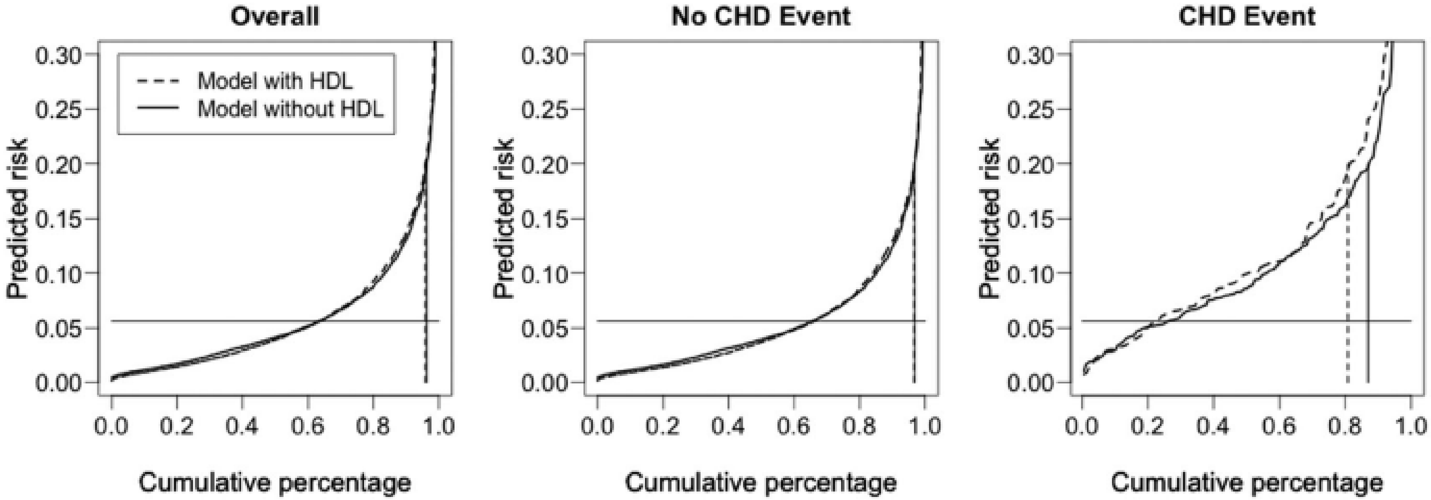
Continuous net reclassification improvement (cNRI)

cNRI nonevents: a net 5.5% of nonevents receive lower predicted risks

cNRI events: a net 24.6% of those with events receive higher predicted risk

cNRI = 5.5% + 24.6% = 30.1%

Predictiveness curves – link between threshold and sensitivity/specificity



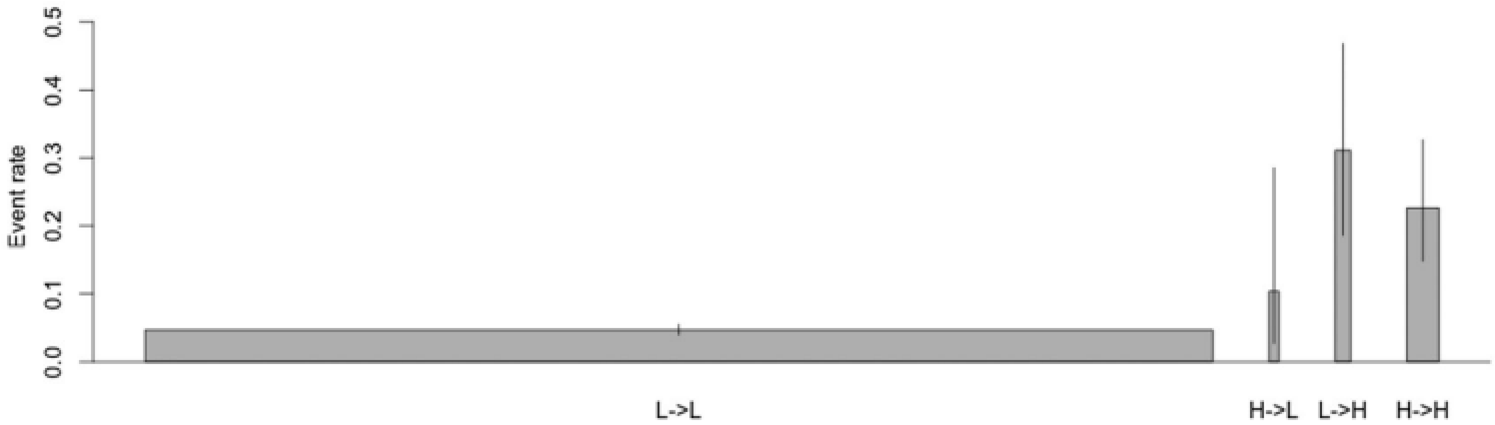
Specificity = P(- | no CHD event) = 96.82% model with HDL vs 96.66% model without HDL

Sensitivity = P(+ | no CHD event) = 13.1% model with HDL vs 19.1% model without HDL

university of groningen

University Medical Center Groningen

Net reclassification risk graph



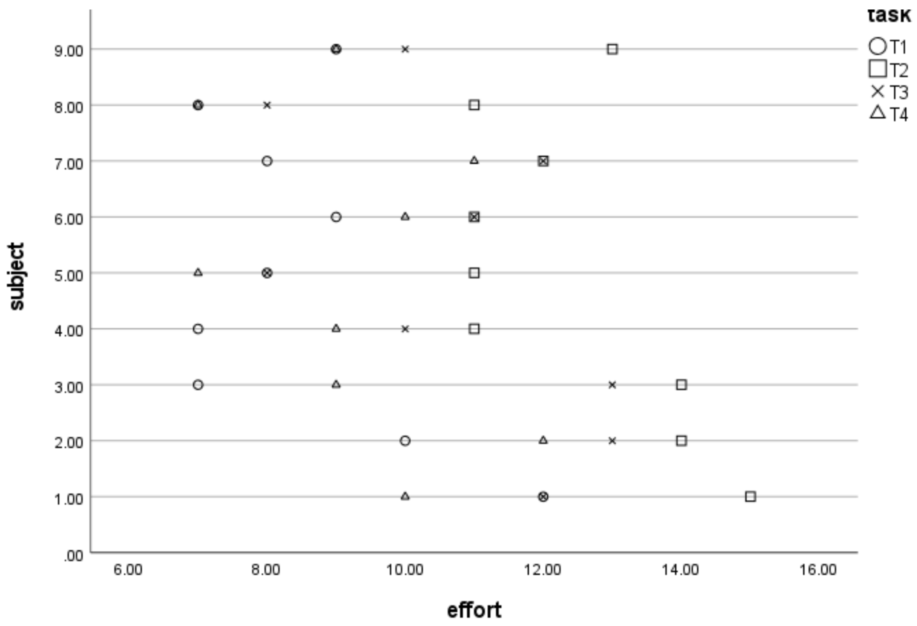
Example 3: randomized block design

- Nine different subjects are asked to complete four different tasks

- The objectives are
  - To assess whether there are systematic differences in the complexity of the four tasks
  - To estimate the between-subject variability in task proficiency

- This experiment is an example of a randomized block design with task as a fixed effect and subject as a random effect

Dot plot



Tasks 1 and 4 seem to take the least effort while task 2 seems to take the most effort

Moderate between-subject variability => interclass correlation (ICC)

Results

```
Linear mixed-effects model fit by REML
  Data: data
       AIC      BIC    logLik
  133.1308 141.9252 -60.56539

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    1.332465 1.100295

Fixed effects:  effort ~ Task
               Value Std.Error DF   t-value p-value
(Intercept) 8.555556 0.5760123 24 14.853079  0.0000
TaskT2      3.888889 0.5186838 24  7.497610  0.0000
TaskT3      2.222222 0.5186838 24  4.284348  0.0003
TaskT4      0.666667 0.5186838 24  1.285304  0.2110
 Correlation:
       (Intr) TaskT2 TaskT3
TaskT2 -0.45
TaskT3 -0.45   0.50
TaskT4 -0.45   0.50   0.50

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.80200345 -0.64316591  0.05783115  0.70099706  1.63142054

Number of Observations: 36
Number of Groups: 9
```

```
            numDF denDF  F-value p-value
(Intercept)     1    24 455.0075  <.0001
Task            3    24  22.3556  <.0001
```

ICC = 1.332465^2 / (1.332465^2 + 1.100295^2) = 0.59

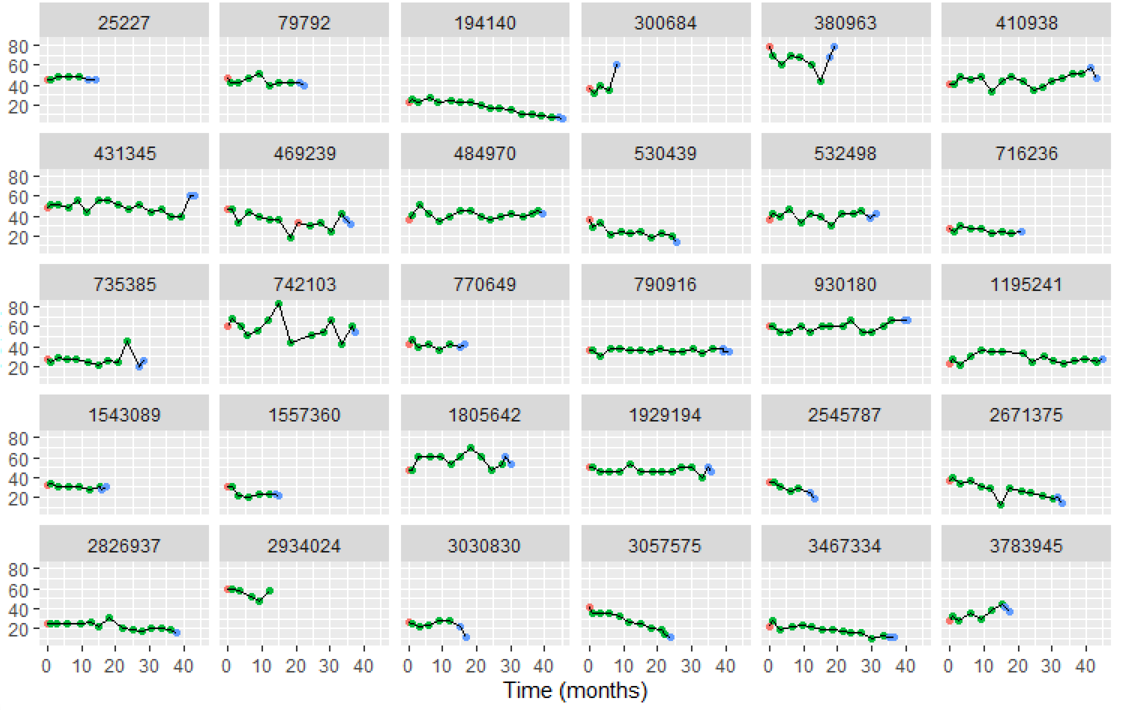
Example 4: longitudinal biomarker measurements

- Objective: model the serial trends in a biomarker through a linear mixed effects model

- The fixed effect structure models the average biomarker trajectory
  - Linear (one-slope), piece-wise linear (two or more slopes), quadratic, cubic spline?

- The random effects model the subject-level deviations from the average trajectory
  - Which terms are needed to appropriately model the subject-level deviations?

Individual profiles plot (trellis graph)



“An intelligent summary of data is often sufficient to fulfil the purposes for which the data were gathered, and more formal techniques such as confidence intervals and hypothesis tests sometimes add little to an investigator’s understanding”

John A. Rice, *mathematical statistics and data analysis, second edition*

www.umcg.nl

contact: d.postmus@umcg.nl

university of groningen

University Medical Center Groningen