Statistical Methodology for Very Small (and Very Large) Studies

Geert Molenberghs

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Universiteit Hasselt & KU Leuven, Belgium

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

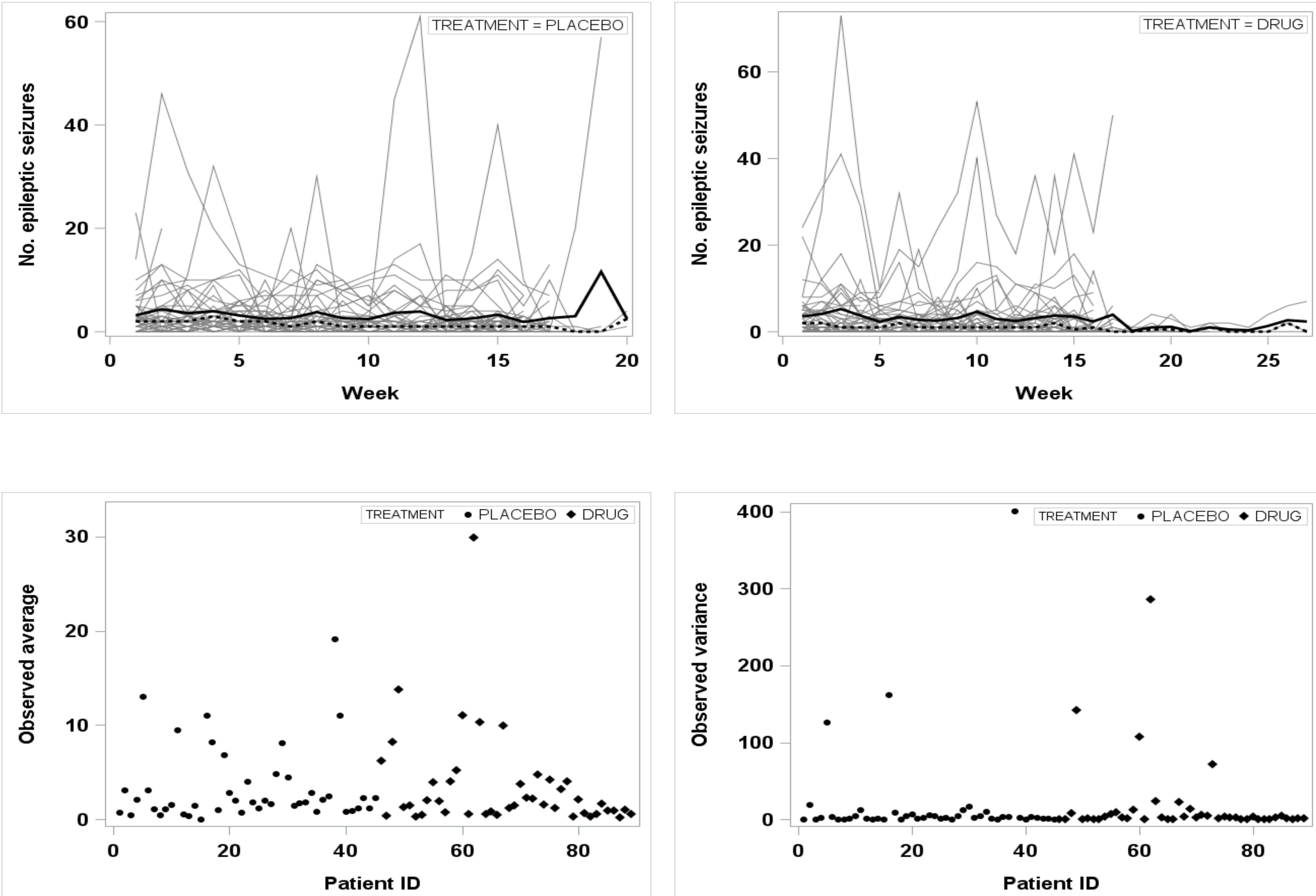
www.ibiostat.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics

EJP RD_WP20, 8 March 2024

The Epilepsy Data

- Randomized, double-blind, parallel group multi-center study
- placebo (45) $\longleftrightarrow$ new anti-epileptic drug (AED; 44)
- 12-week run-in period & 16 weeks of follow up (some until week 27)
- outcome: the number of epileptic seizures experienced during the last week
- research question: reduction in # seizures by new therapy


TREATMENT = PLACEBO
No. epileptic seizures
Week
TREATMENT = DRUG
No. epileptic seizures
Week
TREATMENT ● PLACEBO ◆ DRUG
Observed average
Patient ID
TREATMENT ● PLACEBO ◆ DRUG
Observed variance
Patient ID

The Standard Poisson-normal Model

- **The essence:**
  - ▷ Poisson regression model for epileptic seizures
  - ▷ Random effects to accommodate within-subject correlation

- **Poisson formulation:**

$$
\begin{aligned}
Y_{ij} &\sim \mathrm{Poi}(\lambda_{ij}) \\
\ln(\lambda_{ij}) &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i} \\
\boldsymbol{b_i} &\sim N(\boldsymbol{0}, D)
\end{aligned}
$$

Features Present

Count data	Poisson model
Correlation	Normal random effects
Overdispersion	Normal random effects

A Combined Model:
The Poisson-gamma-normal Model

- **Features Present:**

Count data	Poisson model
Correlation	Normal random effects
Overdispersion	Normal random effects Gamma random effects

The Poisson-gamma-normal Model

- Easy to fit in SAS procedure NLMIXED

- **Model for the epilepsy data:**

$$\ln(\lambda_{ij}) = \begin{cases} (\beta_{00} + b_i) + \beta_{01} t_{ij} & \text{if placebo} \\ (\beta_{10} + b_i) + \beta_{11} t_{ij} & \text{if treated,} \end{cases}$$

$$b_i \sim N(0, d)$$

Parameter Estimates

		Poisson	Negative-binomial
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	β_{00}	1.2662 (0.0424)	1.2594 (0.1119)
Slope placebo	β_{01}	-0.0134 (0.0043)	-0.0126 (0.0111)
Intercept treatment	β_{10}	1.4531 (0.0383)	1.4750 (0.1093)
Slope treatment	β_{11}	-0.0328 (0.0038)	-0.0352 (0.0101)
Negative-binomial parameter	α_1	—	0.5274 (0.0255)
Negative-binomial parameter	$\alpha_2 = 1/\alpha_1$	—	1.8961 (0.0918)
Variance of random intercepts	d	—	—

		Poisson-normal	Combined
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	β_0	0.8179 (0.1677)	0.9112 (0.1755)
Slope placebo	β_1	-0.0143 (0.0044)	-0.0248 (0.0077)
Intercept treatment	β_0	0.6475 (0.1701)	0.6555 (0.1782)
Slope treatment	β_2	-0.0120 (0.0043)	-0.0118 (0.0074)
Negative-binomial parameter	α_1	—	2.4640 (0.2113)
Negative-binomial parameter	$\alpha_2 = 1/\alpha_1$	—	0.4059 (0.0348)
Variance of random intercepts	d	1.1568 (0.1844)	1.1289 (0.1850)

Implications for Correlation Function

		Smallest value		Largest value	
Model	Arm	ρ	time pair	ρ	time pair
Poisson-normal	placebo	0.8577	26 & 27	0.8960	1 & 2
Poisson-normal	treatment	0.8438	26 & 27	0.8794	1 & 2
Combined	placebo	0.3041	26 & 27	0.3134	1 & 2
Combined	treatment	0.2234	1 & 2	0.3410	26 & 27

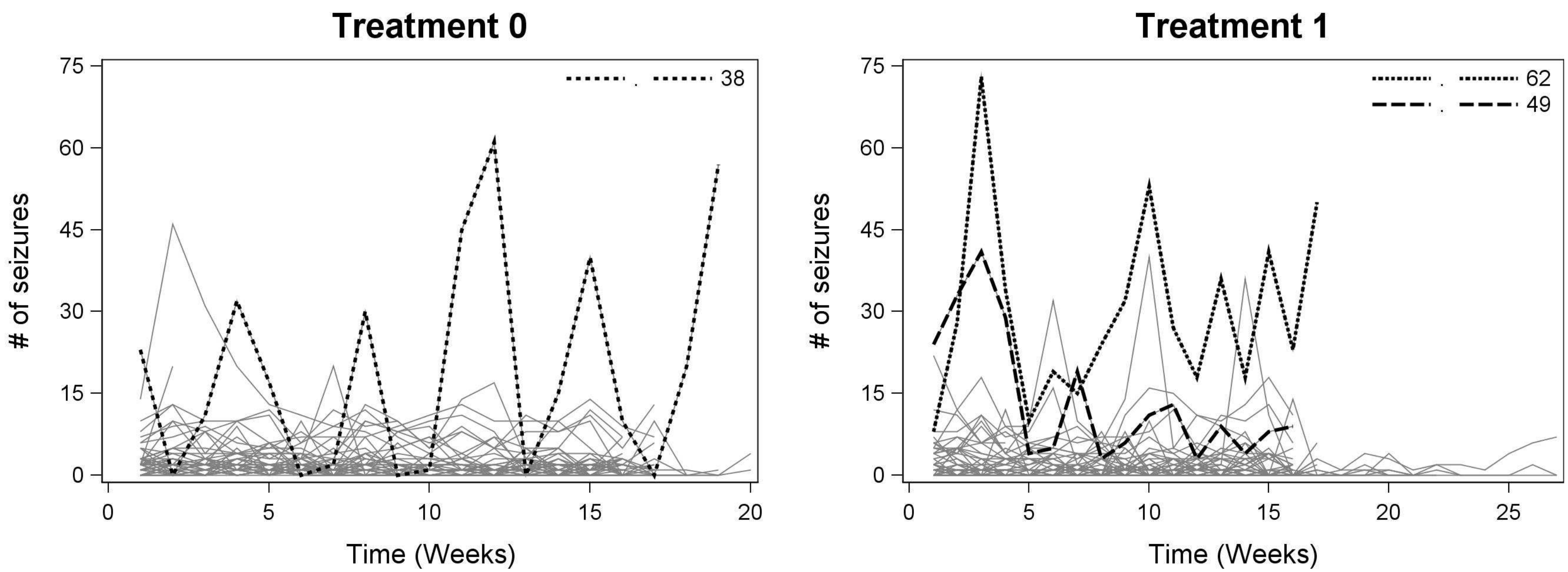
Implications for Hypothesis Testing

	p-values	
Model	$H_0 : \beta_{11} - \beta_{01} = 0$	$H_0 : \beta_{11}/\beta_{01} = 1$
Poisson	**0.0008**	**0.0038**
Poisson-normal	0.7115	**0.0376**
negative-binomial	**0.0131**	0.2815
combined	0.2260	0.1591

But: Aren't There Influential Subjects?

- For which subjects do small perturbations of ω_i generate large effects?

$$\ell = \sum_{i=1}^{N} \omega_i \ell_i$$



- Apart from **low profiles** and **high profiles**, there are **oscillators**
- Upon removal: treatment effect 0.013 (0.011) $\longrightarrow$ 0.022 (0.011)

(Rakhmawati, Molenberghs, Verbeke, and Faes 2016)

Features Present

Count data	Poisson model
Correlation	Normal random effects
Overdispersion	Normal random effects Gamma random effects
Diagnostic tool	Local influence

But: How Do We Get a Marginal Interpretation?

- $\surd$ **First:** Generalized Linear Mixed Model (GLMM) *and its Combined Model (CM)*

- **Second:** Marginalized Multilevel Model (MMM) *and its Combined Model (COMMM)*

- **Third:** Bridge Distributions

Features Present

Count data	Poisson model
Correlation	Normal random effects
Overdispersion	Normal random effects Gamma random effects
Diagnostic tool	Local influence
Marginal mean function	MMM & COMMM & bridge

Effect	Par.	Par. estimates and standard errors	
(a) Models without overdispersion random effects			
		(1a) GLMM & (3a) bridge	(2a) MMM
Intercept placebo	β_{00}	0.8179 (0.1677)	1.3960 (0.1887)
Slope placebo	β_{01}	-0.0143 (0.0044)	-0.0143 (0.0044)
Intercept treatment	β_{10}	0.6475 (0.1701)	1.2256 (0.1901)
Slope treatment	β_{11}	-0.0120 (0.0043)	-0.0120 (0.0043)
Std. dev. R.I.	$\sqrt{d}$	1.0755 (0.0857)	1.0755 (0.0857)
(c) Models with overdispersion random effects)			
		(1b) CM & (3b) c-bridge	(2b) COMMM
Intercept placebo	β_{00}	0.9112 (0.1755)	1.4757 (0.1962)
Slope placebo	β_{01}	-0.0248 (0.0077)	-0.0248 (0.0077)
Intercept treatment	β_{10}	0.6555 (0.1782)	1.2200 (0.1970)
Slope treatment	β_{11}	-0.0118 (0.0075)	-0.0118 (0.0075)
Std. dev. R.I.	$\sqrt{d}$	1.0625 (0.0871)	1.0625 (0.0871)
Overdispersion	α	2.4640 (0.2113)	2.4640 (0.2113)

But: What About Excess Zeroes?

- Features Present:

Count data	Poisson model
Correlation	Normal random effects
Overdispersion	Normal random effects Gamma random effects
Diagnostic tool	Local influence
Marginal mean function	MMM & COMMM & bridge
Excess zeros	ZI— & H—

		Poisson	Zero-Inflated Poisson	Negative Binomial	Zero-Inflated Negative Binomial
Poisson Part					
Slope difference	$\beta_{01} - \beta_{11}$	**-0.0195(0.0058)**	**-0.0214(0.0061)**	**-0.0227(0.0150)**	**-0.0147(0.0153)**
Zero-Inflated Part					
Intercept	γ_0		-1.2879(0.1203)		-7.1064(1.3344)
Slope	γ_1		0.0593(0.0109)		0.2921(0.0655)
Overdispersion	$v = \frac{1}{u}$			0.5274(0.02553)	0.5595(0.03142)

		MMM	Zero-Inflated MMM	Combined MMM	Zero-Inflated Combined MMM
Poisson Part					
Slope diff.	$\beta_{01} - \beta_{11}$	**0.0023(0.0062)**	**-0.0031(0.0065)**	**0.0130(0.0107)**	**0.0080(0.0096)**
Zero-Inflated Part					
Intercept	γ_0		-2.2957(0.2963)		-2.4278(0.3206)
Slope	γ_1		0.0657(0.0166)		0.0662(0.0183)
Overdispersion	$v = \frac{1}{u}$			0.4059(0.03481)	0.1792(0.0175)
Correlation	ρ		-0.1382(0.1601)		-0.0795(0.1669)

Features Present & Others

Count data	**Poisson model**
Semi-continuous data	
Correlation	**Normal random effects** Mixtures of normals
Overdispersion / Underdispersion	**Normal random effects** **Gamma random effects**
Diagnostic tool	**Local influence**
Marginal mean function	**MMM & COMMM & bridge**
Excess zeros	**ZI— & H—**
Inference paradigm	Likelihood / Bayes / moment-based
. . .	. . .

References

Iddi, S. and Molenberghs, G. (2013). A marginalized model for zero-inflated, overdispersed and correlated count data. *Electronic Journal of Applied Statistical Analysis*, **6**, 149–165.

Kassahun, W., Neyens, T., Molenberghs, G., Faes C., and Verbeke, G. (2015). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation*, **85**, 552–571.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G., and Verbeke, G. (2014). A Zero-inflated overdispersed hierarchical Poisson model. *Statistical Modeling*, **14**, 439–456.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke, G. (2014). Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine*, **33**, 4402–4419.

Molenberghs, G. and Verbeke, G. (2011). On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. *Journal of Statistical Planning and Inference*, **141**, 861–868.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R., and Demétrio, C.G.B. (2015). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Statistical Methods in Medical Research*, **24**, 434–452.

Rakhmawati, T., Molenberghs, G., Verbeke, G., and Faes, C. (2016). Local Influence Diagnostics for Incomplete Overdispersed Longitudinal Counts. *Journal of Applied Statistics*, **43**, 1722–1737.

Rakhmawati, T., Molenberghs, G., Verbeke, G., and Faes, C. (2016). Local Influence Diagnostics for Hierarchical Count Data Models With Overdispersion and Excess Zeros. *Biometrical Journal*, **58**, 1390–1408.

Rakhmawati, T., Molenberghs, G., Verbeke, G., and Faes, C. (2017). Local Influence Diagnostics for Generalized Linear Mixed Models With Overdispersion. *Journal of Applied Statistics*, **44**, 620–641.

Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics*, **38**, 215–232.

Standard Inference Paradigm: Maximum Likelihood Estimation

- Random outcome data Y_i, $i = 1, \ldots, N$
- Possibly covariates $\boldsymbol{x}_i$
- Distribution described by density function $f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$
- $\boldsymbol{\theta}$ parameter to be estimated from the data
- **Log-likelihood function:**

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{N} \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$$

- Maximum likelihood estimator defined as the solution to the **score equations**:

$$S(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Solution:
  - ▷ Closed-form in a number of (simple) but often-used settings
  - ▷ In contemporary problems numerical solution is needed

- Second derivative (**Hessian matrix**) used for:
  - ▷ Numerical optimization (Newton-Raphson,. . .)
  - ▷ Estimation of standard errors

$$H(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

- **Sometimes, MLE simply too cumbersome!**

Alternative Principle: Pseudo-likelihood

- *Arnold and Strauss (Indian J. Stat. 1991)*
- *Geys, Molenberghs, and Ryan (JASA 1999)*
- *Molenberghs and Verbeke (2005)*
- **Units:** clusters, repeated measures, spatial data, microarrays,...

$$\boxed{f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1|y_2, y_3) \cdot f(y_2|y_1, y_3) \cdot f(y_3|y_1, y_2)}$$

$$\boxed{f(y_1, y_2, y_3) \quad \longleftrightarrow \quad f(y_1, y_2) \cdot f(y_1, y_3) \cdot f(y_2, y_3)}$$

$$f(y_{i1}, \ldots, y_{in_i})$$

replaced by a product of convenient factors

- The **wrong** likelihood used
- The **right** results obtained:
  - ▷ Consistent, asymptotically normal estimators
  - ▷ Often minor loss of statistical efficiency
  - ▷ Often major gain of computational efficiency

Specific Use 1:
Pseudo-likelihood for HD Multivariate Longitudinal Data

- *Fieuws and Verbeke (Biometrics 2006); Fieuws* et al *(JRSS-C 2006)*

- M sequences of repeated measures

- **Example:** 44 sequences of hearing variables

- Data for patient i:

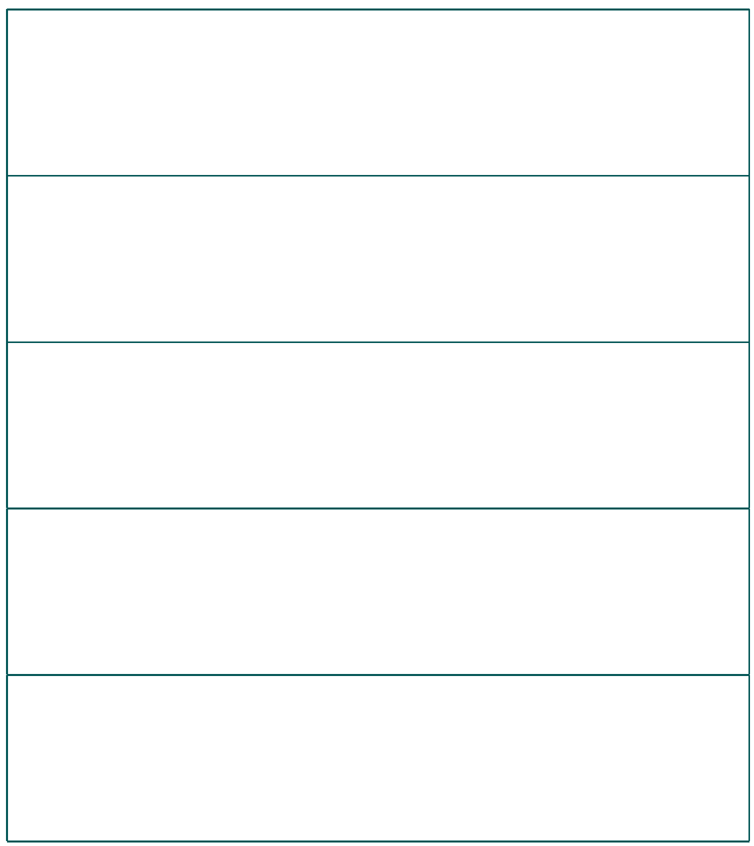
Y_{i11}	Y_{i12}	Y_{i13}	$\ldots$	Y_{i1n_i}
Y_{i21}	Y_{i22}	Y_{i23}	$\ldots$	Y_{i2n_i}
Y_{i31}	Y_{i32}	Y_{i33}	$\ldots$	Y_{i3n_i}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Y_{i,44,1}$	$Y_{i,44,2}$	$Y_{i,44,3}$	$\ldots$	$Y_{i,44,n_i}$

- Fit model to each of the $M(M-1)/2$ pairs

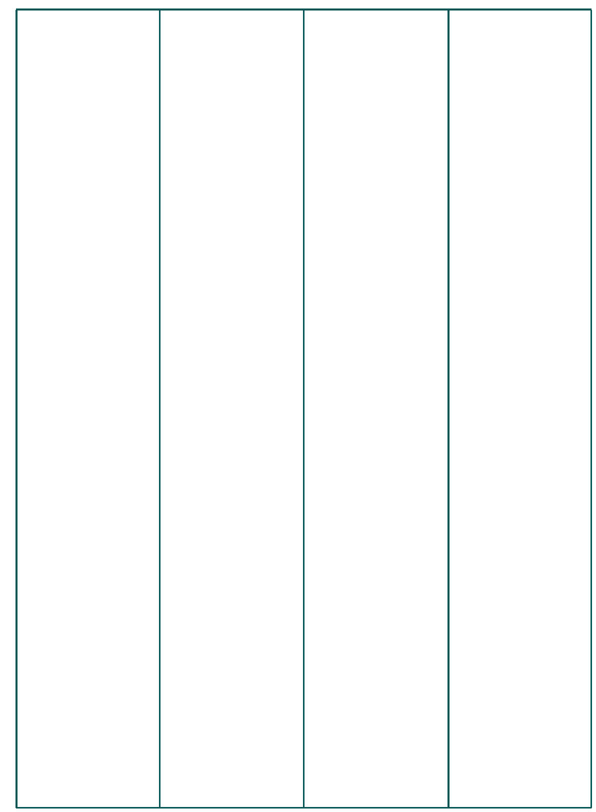
- Use PL to reach valid conclusions

Specific Use 2:

Split Sample Method: (In)dependent Subsamples



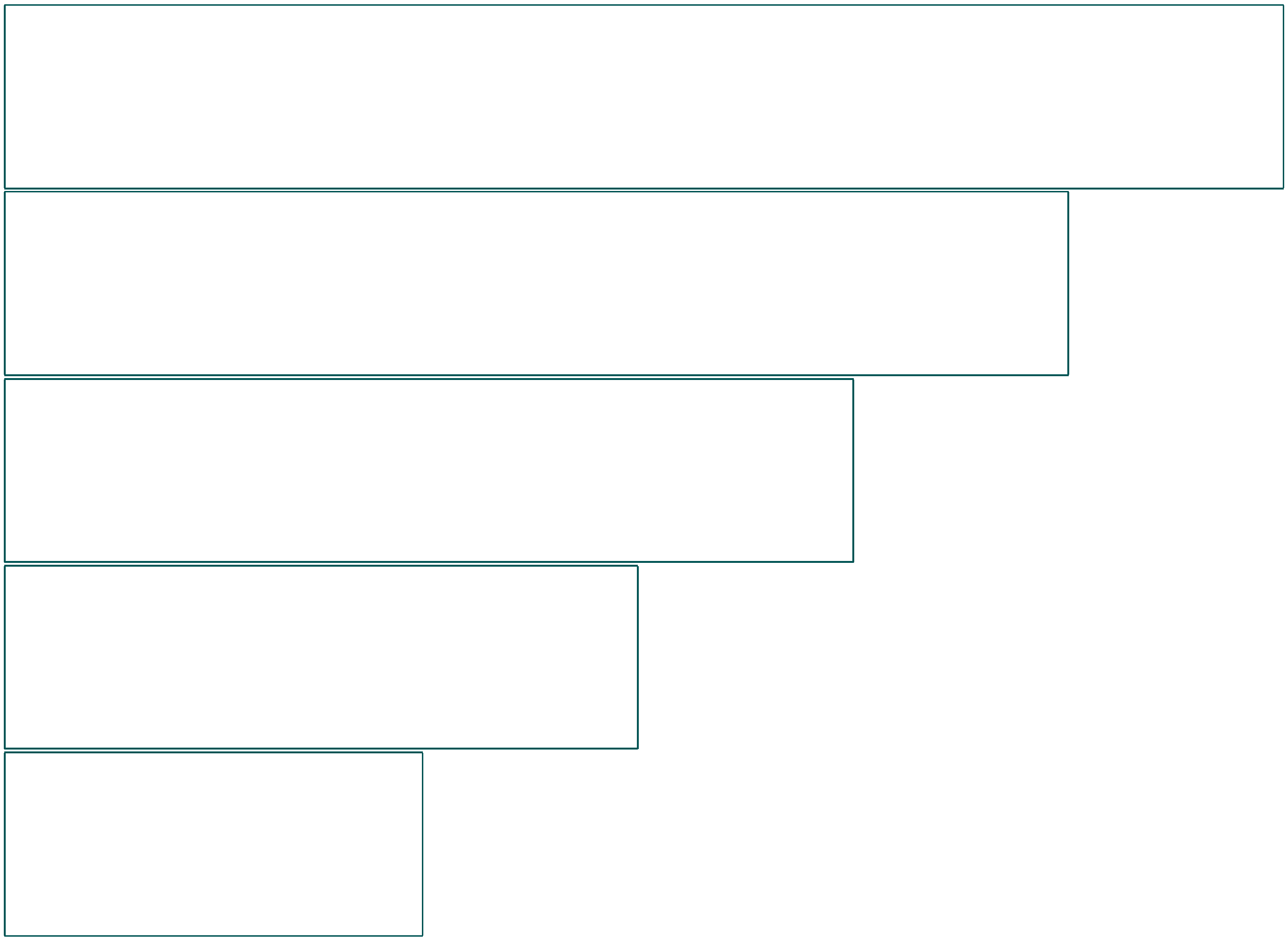
or



Behavior

- *Molenberghs, Verbeke, and Iddi (Stat. & Prob. Letters 2011)*
- **Univariate normal: equivalent**
- **Univariate Bernoulli (probability): equivalent**
- **Univariate Bernoulli (logit): different estimator, same precision**
- **Compound symmetry: different estimator, mild precision loss**

Specific Use 3:
Per Cluster Size



Fixed Cluster Size $\longleftrightarrow$ Variable Cluster Size

- **Fixed cluster size:** closed-form maximum likelihood estimator: **easy**

- **Variable cluster size:**

  ▷ **Estimate parameters per cluster size**

  ▷ **Average these**

  ▷ **But: Now weighted average needed (several weights possible)**

Specific Use 4:
Surrogate Markers

- **Model:**

$$S_{ij} = \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij}$$
$$T_{ij} = \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij}$$

- **Error structure:**

  ▷ **Individual level:**
  - ∗ Deviations ε_{Sij} and ε_{Tij} are correlated

  ▷ **Trial level:**
  - ∗ Treatment effects α_i and β_i are correlated
  - ∗ (Information from intercepts μ_{Si} and μ_{Ti} can be used as well)

- Estimation can be problematic:
  - ▷ especially in small studies
  - ▷ especialy when studies are of differing sizes
- **Solution 1:** Use multiple imputation to make all studies equally large
- **Solution 2:**
  - ▷ Analyze trial-by-trial: it can be shown that this is valid
  - ▷ Combine results across trials using weighted averages
  - ▷ When some (or all) trials are very large: sub-sampling is allowable
- **Solution 2-advantages:**
  - ▷ **:** Very stable $\longleftarrow$ small trials
  - ▷ **:** Very fast $\longleftarrow$ very large trials
- *Van der Elst, Hermans, Verbeke, Kenward, Nassiri, and Molenberghs (CSDA 2016)*

Statistics

Applied Surrogate Endpoint Evaluation Methods with SAS and R provides an overview of contemporary meta-analytic and information-theoretic methodology to evaluate candidate surrogate endpoints from clinical trials and beyond. The book strongly focuses on user-friendly software in both SAS and R for a variety of outcome types.

The book is aimed at researchers and practitioners who want to study and apply methodology for surrogate endpoint and biomarker evaluation. Methodology is described while keeping mathematical detail to a minimum. Throughout the book, a suite of generic case studies is used to illustrate the concepts and methodology. A large part of the book is devoted to the description and illustration of SAS macros, R language libraries, and R Shiny Apps. The software tools can be downloaded from the authors' web pages. Methodology, applications, and software encompass continuous, binary, categorical, time-to-event, and longitudinal outcomes.

The University of Hasselt and KU Leuven-based editor team, supplemented by a fine group of chapter authors, has over twenty years of experience in the field of surrogate marker evaluation in clinical and other studies. The book is rooted at the same time in methodological research, regular and short courses taught on the topic, as well as in vast experience with the design and conduct of clinical trials. The team's prolific contributions have led to numerous papers, chapters, and books on this topic. This book was written in a coherent fashion, with common notation, conventions, and case studies throughout all chapters.

CRC Press
Taylor & Francis Group
an informa business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

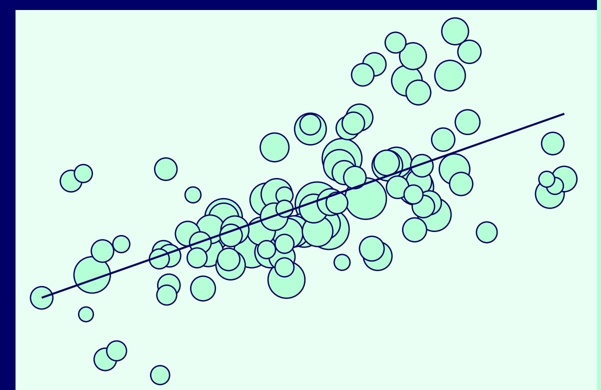
K23717
ISBN: 978-1-4822-4936-1
90000
9 781482 249361
www.crcpress.com

Applied Surrogate Endpoint Evaluation Methods with SAS and R

Alonso • Bigirumurame
Burzykowski • Buyse
Molenberghs • Muchene
Perualila • Shkedy • Van der Elst

Chapman & Hall/CRC Biostatistics Series

Applied Surrogate Endpoint Evaluation Methods with SAS and R



Ariel Alonso
Theophile Bigirumurame
Tomasz Burzykowski
Marc Buyse
Geert Molenberghs
Leacky Muchene
Nolen Joy Perualila
Ziv Shkedy
Wim Van der Elst

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Surrogate Markers and Beyond: Trial-by-trial estimator

- *Poveda, Molenberghs, Verbeke, Alonso (J. Biopharmaceutical Stat. 2019)*
- Very general **multivariate linear mixed model** can be used
- **Closed-form estimators per trial**
- **Weighting to combine across trials**
- Involves considerable matrix algebra – but computationally feasible
- **Simulations: 10 to 100 times faster & very efficient**

Meta-analysis in Schizophrenia

- 2128 patients treated by 198 psychiatrists
- From 6 to 52 patients per psychiatrist
- Psychiatrists with 1 or 2 patients excluded (1392 patients remaining)
- Three outcomes:
  - ▷ **PANSS:** Positive and Negative Syndrome Scale
  - ▷ **BPRS:** Brief Psychiatric Rating Scale
  - ▷ **CGI:** Clinician's Global Impression

Parameter Estimates for a Joint Model

	Trial-by-trial		REML	
Parameter	Estimate	Std. error	Estimate	Std. error
$\beta_{0,\text{BPRS}}$	-8.15	0.863	-7.85	0.519
$\beta_{1,\text{BPRS}}$	-1.49	0.408	-1.26	0.332
$\beta_{0,\text{CGI}}$	3.28	0.097	3.32	0.054
$\beta_{1,\text{CGI}}$	-0.16	0.046	-0.12	0.038
$\beta_{0,\text{PANSS}}$	-14.59	1.53	-13.87	0.911
$\beta_{1,\text{PANSS}}$	-2.74	0.707	-2.41	0.582

Specific Use 5: Leuven Diabetes Study

- *Ivanova, Molenberghs, and Verbeke (SMMR 2017)*
- 120 general practitioners — 2495 patients
- **Outcomes**
  - ▷ **LDL: low-density lipoprotein cholestrol**
  - ▷ **HbA1C: glycosylated hemoglobin**
  - ▷ **SBP: systolic blood pressure**
- **Ordinal targets**
- Multiple outcomes & measured repeatedly & ordinal

$\Longrightarrow$ **joint modeling**

Leuven Diabetes Study: Targets

		# Observations	
LDL targets		T_0	T_1
1:	$<$ 100 mg/dl	819	1106
2:	$\geq$ 100 mg/dl & $<$ 115 mg/dl	381	312
3:	$\geq$ 115 mg/dl & $<$ 130 mg/dl	287	220
4:	$\geq$ 130 mg/dl	485	250
missing		287	371
HbA1C targets		T_0	T_1
1:	$<$ 7 %	1201	1357
2:	$\geq$ 7 % & $<$ 8 %	604	474
3:	$\geq$ 8 %	413	176
missing		41	252
SBP targets		T_0	T_1
1:	$\leq$ 130 mmHg	1103	1152
2:	$>$ 130 mmHg & $\leq$ 140 mmHg	551	469
3:	$>$ 140 mmHg & $\leq$ 160 mmHg	466	324
4:	$>$ 160 mmHg	136	75
missing		3	239

Method	3 sequences	Partitioning	CPU
1 $\equiv$ ML	(123)		7'13"
2 $\equiv$ PLp	(12)(13)(23)		1'23"
3 $\equiv$ PLs	(123)		1'21"
4 $\equiv$ PLps	(12)(13)(23)		0'20"

Some Parameter Estimates (LDL)

Effect	1 ≡ ML	2 ≡ PLp	3 ≡ PLs	4 ≡ PLps
intercept 1	−1.076 (0.108)	−1.073 (0.107)	−1.063 (0.109)	−1.061 (0.110)
intercept 2	0.155 (0.105)	1.157 (0.106)	0.183 (0.107)	0.185 (0.109)
intercept 3	1.257 (0.110)	1.258 (0.115)	1.291 (0.112)	1.292 (0.118)
time	1.025 (0.076)	1.025 (0.071)	1.025 (0.077)	1.025 (0.072)
diabetes duration $T_0/10$	0.213 (0.088)	0.216 (0.090)	0.198 (0.090)	0.201 (0.091)
gender	0.497 (0.110)	0.497 (0.110)	0.497 (0.111)	0.497 (0.112)
insuline	0.853 (0.150)	0.829 (0.153)	0.877 (0.153)	0.852 (0.156)
random int. standard dev.	1.852 (0.089)	1.849 (0.085)	1.853 (0.090)	1.849 (0.087)

CPU Gain / Efficiency Loss

- Subsamples can be analyzed in parallel
- Base model above, with numerical integration over $Q = 3$ quadrature points:

7'13" $\longrightarrow$ 0'20"

- More demanding integration: $Q = 15$

10h02'42" $\longrightarrow$ 0h4'17"

- Statistical efficiency: almost always $\geq 95\%$
- For PLps occasionally $85\% - 87\%$

Conclusions

- Broad framework based on:
  - ▷ pseudo-likelihood
  - ▷ pairwise modeling
  - ▷ split sample
- Statistically valid procedures: consistent, asymptotically normal
- Can lead to tremendous CPU gain
- Statistical efficiency loss mostly acceptable

Incomplete Data

Setting the Scene Using Examples

- ▷ Orthodontic growth data
- ▷ Age-related macular degeneration trial
- ▷ Notation
- ▷ Taxonomy

Growth Data

- Taken from Potthoff and Roy, Biometrika (1964)

- Research question:

Is dental growth related to gender ?

- The distance from the center of the pituitary to the maxillary fissure was recorded at ages 8, 10, 12, and 14, for 11 girls and 16 boys

- Individual profiles:
  - ▷ Much variability between girls / boys
  - ▷ Considerable variability within girls / boys
  - ▷ Fixed number of measurements per subject
  - ▷ Measurements taken at fixed time points

Age-related Macular Degeneration Trial

- Pharmacological Therapy for Macular Degeneration Study Group (1997)
- An occular pressure disease which makes patients progressively lose vision
- 240 patients enrolled in a multi-center trial (190 completers)
- **Treatment:** Interferon-α (6 million units) versus placebo
- **Visits:** baseline and follow-up at 4, 12, 24, and 52 weeks
- **Continuous outcome: visual acuity:** $\#$ letters correctly read on a vision chart
- **Binary outcome:** visual acuity versus baseline ≥ 0 or ≤ 0

- Missingness:

Measurement occasion					
4 wks	12 wks	24 wks	52 wks	Number	%
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

CRF	TRT	VISUAL0	VISUAL4	VISUAL12	VISUAL24	VISUAL52	lesion
1002	4	59	55	45	.	.	3
1003	4	65	70	65	65	55	1
1006	1	40	40	37	17	.	4
1007	1	67	64	64	64	68	2
1010	4	70	.	.	.	.	1
1110	4	59	53	52	53	42	3
1111	1	64	68	74	72	65	1
1112	1	39	37	43	37	37	3
1115	4	59	58	49	54	58	2
1803	1	49	51	71	71	.	1
1805	4	58	50	.	.	.	1
...							

Notation

- Subject i at occasion (time) $j = 1, \ldots, n_i$
- **Measurement** Y_{ij}
- **Missingness indicator** $R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$
- Group Y_{ij} into a vector $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})' = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$

$$\begin{cases} \boldsymbol{Y}_i^o & \text{contains } Y_{ij} \text{ for which } R_{ij} = 1, \\ \boldsymbol{Y}_i^m & \text{contains } Y_{ij} \text{ for which } R_{ij} = 0. \end{cases}$$

- Group R_{ij} into a vector $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$
- D_i: time of dropout: $D_i = 1 + \sum_{j=1}^{n_i} R_{ij}$

Notation: Example

```
CRF      TRT     VISUAL0     VISUAL4     VISUAL12     VISUAL24     VISUAL52

1002      4        59          55          45            .            .
R-vector                        1           1            0            0
D-value                                                  3

1003      4        65          70          65           65           55
R-vector                        1           1            1            1
D-value                                                                  --> 5

1006      1        40          40          37           17            .
R-vector                        1           1            1            0
D-value                                                               4

...
}
```

Players On The Field

Quantity	Notation
Covariates	$\boldsymbol{X}_i$
Outcomes	$\boldsymbol{Y}_i$
Observed part of the outcomes	$\boldsymbol{Y}_i^o$
Missing part of the outcomes	$\boldsymbol{Y}_i^m$
Missingness indicators	$\boldsymbol{R}_i$

R_i: The Party Crasher

- We are interested in the relationship between $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$
- We are interested in how $\boldsymbol{X}_i$ =**vaccination status** influences $\boldsymbol{Y}_i$ =**infection**
- But... $\boldsymbol{R}_i$ **(missingness)** is the uninvited guest

$$
\begin{tikzcd}
 & \boldsymbol{R}_i \arrow[dl, shift left] \arrow[dr, shift left] & \\
\boldsymbol{Y}_i \arrow[ur, shift left] & & \boldsymbol{X}_i \arrow[ll] \arrow[ul, shift left]
\end{tikzcd}
$$

The Model We Like and The Model We Need

- We would love to build a model for how $\boldsymbol{X}_i$ influences $\boldsymbol{Y}_i$

$$f(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})$$

- But because of the nuisance $\boldsymbol{R}_i$, we need:

$$f(\boldsymbol{Y}_i, \boldsymbol{R}_i|\boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi})$$

We tend to break it up:

Model	Notation
Model of scientific interest	$f(\boldsymbol{Y}_i\|\boldsymbol{X}_i, \boldsymbol{\theta})$
Missingness model	$f(\boldsymbol{R}_i\|\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\psi})$ $= f(\boldsymbol{R}_i\|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$

The Missingness Model

$$f(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$$

Missing Completely at Random (MCAR) $f(\boldsymbol{R}_i|\boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ Missingness depends on covariates only
- ▷ Missingness of seizures can depend on age, gender, treatment, but not on infection status itself
- ▷ Missingness on visual acuity can depend on treatment arm and on lesion type, but not on visual acuity itself
- ▷ Simplest mechanism
- ▷ But... usually too simple to be clinically or epidemiologically plausible

Missing at Random (MAR) $f(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ Missingness depends on covariates and on **observed** outcomes
- ▷ Missingness on seizures **now** can depend on covariates and on earlier seizures variables
- ▷ Given that information, it does not depend on today's, possibly missing seizures
- ▷ Much more plausible than MCAR
- ▷ Common misunderstanding is that MAR implies that everybody has the same probability of being missing at some point — NO! But it is permitted to depend only on **observed** information
- ▷ Under MAR, we have all the data in hand to build models, for outcomes and for the missingness mechanism

Missing Not at Random (MNAR) $f(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ The full menu
- ▷ Missingness can depend on covariates, **and** on observed outcomes, **and** on missing outcomes
- ▷ Missingness in seizures today can depend on age, gender, treatment, **and** on earlier seizures, **and** on today's, potentially missing seizures
- ▷ **Major problem:** "We do not have the missing outcomes"
- ▷ **Major problem:** "We do not have the missing infection status"
- ▷ This also means that MAR and MNAR cannot be distinguished from each other based on data alone!

Where Does That Leave Us Towards Analyzing Incomplete Data?

Missing Completely at Random (MCAR) $f(\boldsymbol{R}_i|\boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ Too simplistic $\longrightarrow$ forget about it
- ▷ Should it apply anyway, then an MAR approach would do the job anyhow

Missing at Random (MAR) $f(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ Very appealing place for our primary analysis:
  - ∗ Quite general mechanism
  - ∗ Yet, we do not need to bother with unobserved data
  - ∗ Likelihood and Bayesian approaches come with extra appeal: **ignorability**

Missing Not at Random (MNAR) $f(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$

- ▷ MNAR can never be ruled out
- ▷ It is the playground of **sensitivity analysis**

Direct Likelihood/Bayesian Inference: Ignorability

- Under MAR, it looks like we have to deal with two models:

The model of interest $f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{X}_i, \boldsymbol{\theta})$

The missingness model $f(\boldsymbol{R}_i | \boldsymbol{Y}_i^o, \boldsymbol{X}_i, \boldsymbol{\psi})$

- But when we use **maximum likelihood** or **Bayesian** estimation, there is more good news:

$$\boxed{\textbf{MAR}}\ \textbf{:}\ f(\boldsymbol{Y}_i^o | \boldsymbol{X}_i, \boldsymbol{\theta})\, f\cancel{(\boldsymbol{R}_i | \boldsymbol{Y}_i^o,, \boldsymbol{X}_i, \boldsymbol{\psi})}$$

- There is no need to model the missing data mechanism

- Only the observed outcomes and the covariates need to be modeled – i.e., the data that we happen to have

- Just make sure that the software can handle unbalanced data because not everyone has the same number of measurements

- Where would we use maximum likelihood or Bayes?
  - ▷ Linear mixed models
  - ▷ Generalized linear mixed models

- Where would we **not** use maximum likelihood or Bayes?
  - ▷ Generalized estimating equations $\longleftarrow$ **non-ignorable under MAR!**

Taxonomy

- **Missingness pattern:** complete — monotone — non-monotone
- Dropout pattern: complete — dropout — intermittent
- **Model framework:** SEM — PMM — SPM
- **Missingness mechanism:** MCAR — MAR — MNAR
- **Ignorability:** ignorable — non-ignorable
- **Inference paradigm:** frequentist — likelihood — Bayes

A Word About Modeling Frameworks

- We considered **selection models:** (but did not say that yet)

  The data model of interest $f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{X}_i, \boldsymbol{\theta})$

  The missingness model $f(\boldsymbol{R}_i | \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{X}_i, \boldsymbol{\psi})$

- An alternative framework: **pattern-mixture models:**

  The data model per pattern $f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{R}_i \boldsymbol{X}_i, \boldsymbol{\theta}^*)$

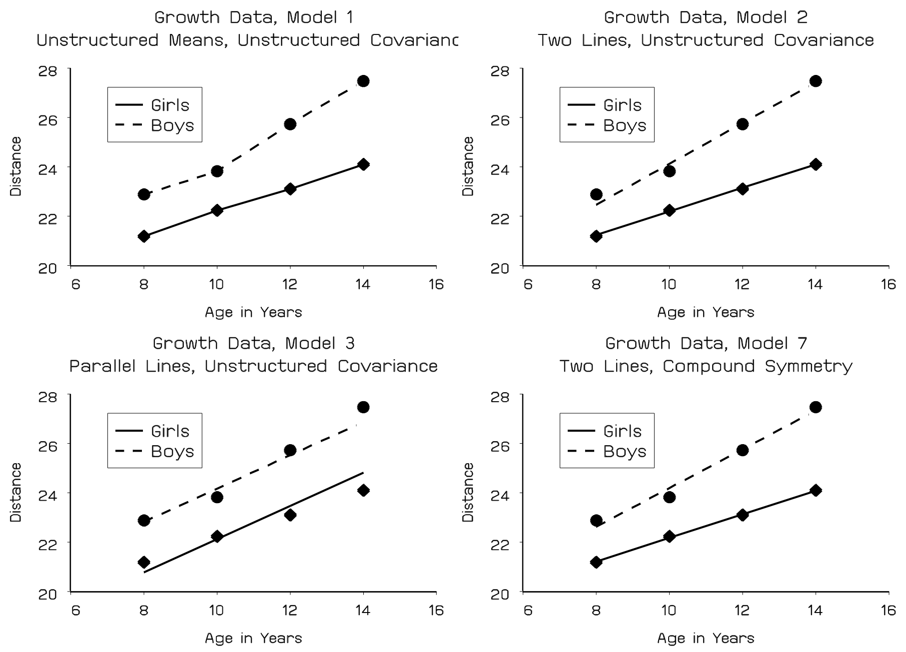
  The probability to belong to a pattern $f(\boldsymbol{R}_i | \boldsymbol{X}_i, \boldsymbol{\psi}^*)$

Frameworks and Their Methods

$\boxed{\text{MCAR/simple}}$	$\longrightarrow$	$\boxed{\text{MAR}}$	$\longrightarrow$	$\boxed{\text{MNAR}}$
CC?		**direct likelihood!**		joint model?
LOCF?		**direct Bayesian!**		**sensitivity analysis!**
single imputation?		**multiple imputation (MI)!**		
$\vdots$		**IPW $\supset$ W-GEE!**		
		d.l. + IPW = double robustness! (consensus)		

Original, Complete Orthodontic Growth Data

	Mean	Covar	# par
1	unstructured	unstructured	18
2	$\neq$ slopes	unstructured	14
3	$=$ slopes	unstructured	13
7	**$\neq$ slopes**	**CS**	**6**



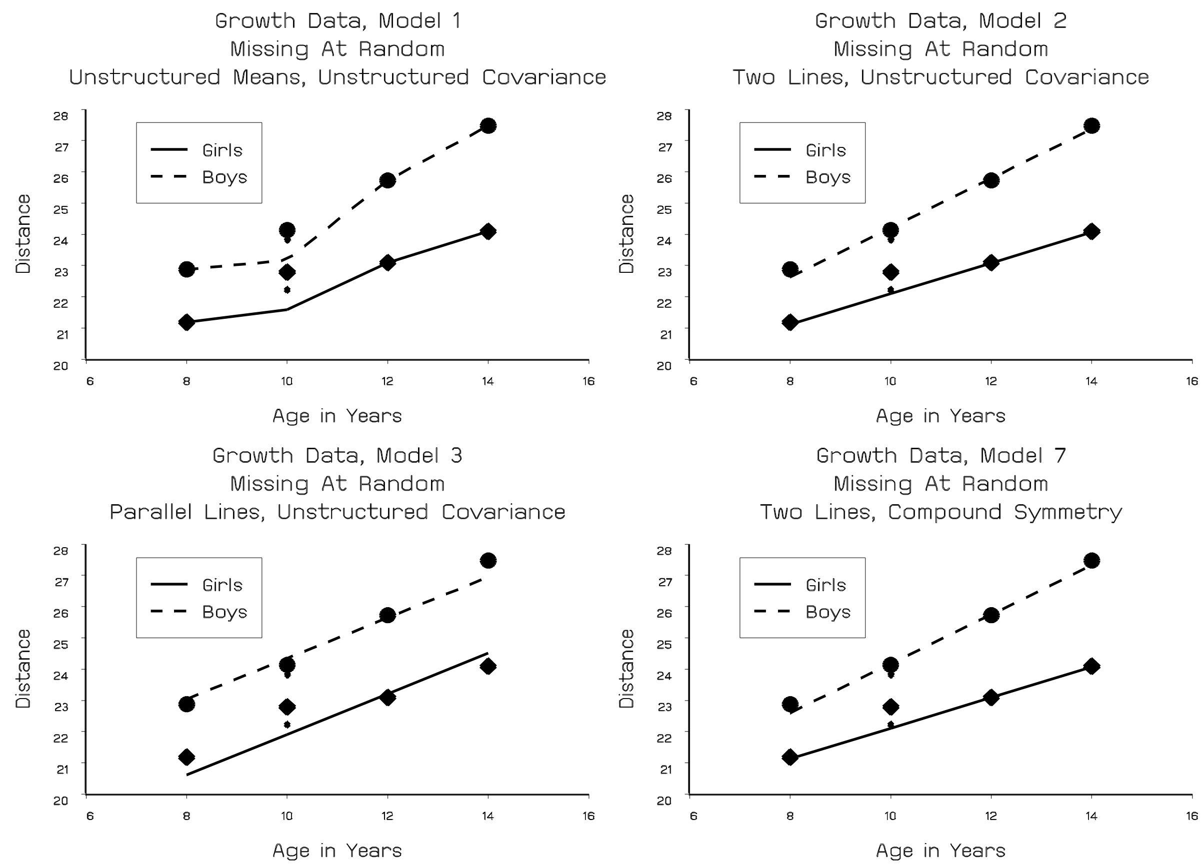
Incomplete Growth Data: Simple Methods

Method	Model	Mean	Covar	# par
Complete case	7a	$=$ slopes	CS	5
LOCF	2a	quadratic	unstructured	16
Unconditional mean	7a	$=$ slopes	CS	5
Conditional mean	1	unstructured	unstructured	18

distorting

Incomplete Growth Data: Direct Likelihood

	Mean	Covar	# par
7	$\neq$ slopes	CS	6



Analysis of the ARMD Trial

- Model for continuous outcomes:

$$Y_{ij} = \beta_{j1} + \beta_{j2} T_i + \varepsilon_{ij}$$

  with:

  ▷ $T_i = 0$ for placebo and $T_i = 1$ for interferon-α

  ▷ t_j $(j = 1, \ldots, 4)$ refers to the four follow-up measurements

  ▷ $\beta_{12}, \ldots, \beta_{42}$ are the treatment effects at the four follow-up times

  ▷ unstructured variance-covariance matrix

- Turning to the dichotomous outcome...

- Marginal mean for GEE:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j)] = \beta_{j1} + \beta_{j2}T_i$$

- Model for GLMM with random interecept:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j, b_i)] = \beta_{j1} + b_i + \beta_{j2}T_i$$

with

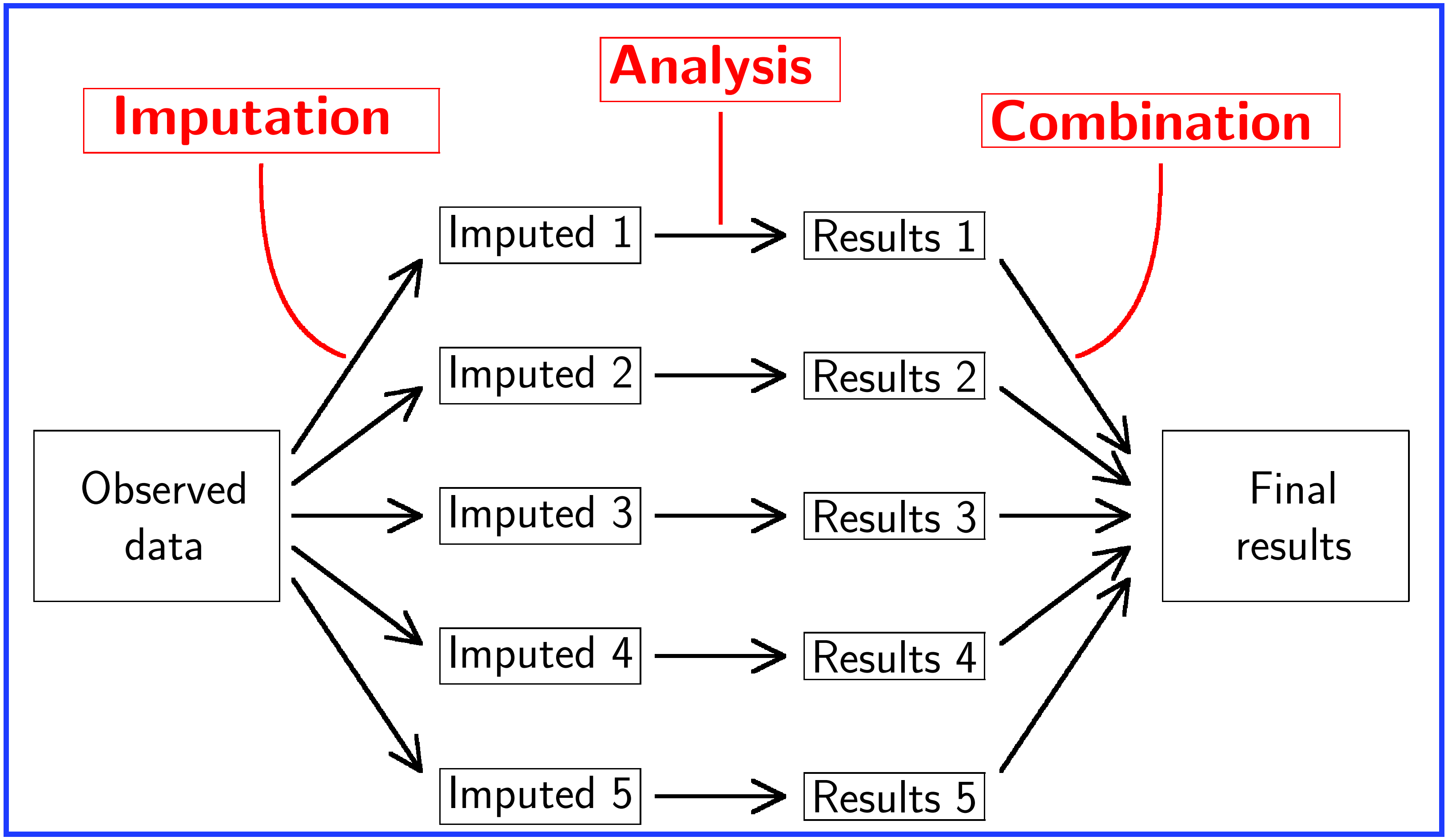
▷ $b_i \sim N(0, \tau^2)$

Effect	Parameter	CC	LOCF	direct lik.
Parameter estimates (standard errors) for linear mixed model				
Intercept 4	β_{11}	-3.24(0.77)	-3.48(0.77)	-3.48(0.77)
Intercept 12	β_{21}	-4.66(1.14)	-5.72(1.09)	-5.85(1.11)
Intercept 24	β_{31}	-8.33(1.39)	-8.34(1.30)	-9.05(1.36)
Intercept 52	β_{41}	-15.13(1.73)	-14.16(1.53)	-16.21(1.67)
Treatm. eff. 4	β_{12}	2.32(1.05)	2.20(1.08)	2.20(1.08)
Treatm. eff. 12	β_{22}	2.35(1.55)	3.38(1.53)	3.51(1.55)
Treatm. eff. 24	β_{32}	2.73(1.88)	2.41(1.83)	3.03(1.89)
Treatm. eff. 52	β_{42}	4.17(2.35)	3.43(2.15)	4.86(2.31)
p-values				
Treatm. eff. 4	β_{12}	0.0282	0.0432	0.0435
Treatm. eff. 12	β_{22}	0.1312	0.0287	0.0246
Treatm. eff. 24	β_{32}	0.1491	0.1891	0.1096
Treatm. eff. 52	β_{42}	0.0772	0.1119	0.0366
Treatm. eff. (overall)		0.1914	0.1699	0.1234

Effect	Parameter	CC	LOCF	direct lik.
Binary outcome: GLMM				
Int.4	β_{11}	-1.73(0.42)	-1.63(0.39)	-1.50(0.36)
Int.12	β_{21}	-1.53(0.41)	-1.80(0.39)	-1.73(0.37)
Int.24	β_{31}	-1.93(0.43)	-1.96(0.40)	-1.83(0.39)
Int.52	β_{41}	-2.74(0.48)	-2.76(0.44)	-2.85(0.47)
Trt.4	β_{12}	0.64(0.54)	0.38(0.52)	0.34(0.48)
Trt.12	β_{22}	0.81(0.53)	0.98(0.52)	1.00(0.49)
Trt.24	β_{32}	0.77(0.55)	0.74(0.52)	0.69(0.50)
Trt.52	β_{42}	0.60(0.59)	0.57(0.56)	0.64(0.58)
R.I. s.d.	τ	2.19(0.27)	2.47(0.27)	2.20(0.25)
R.I. var.	τ^2	4.80(1.17)	6.08(1.32)	4.83(1.11)

Multiple Imputation

- Multiple imputation ($M = 5$ imputations):



Use of MI in Practice

- Many analyses of the same incomplete set of data
- A combination of missing outcomes and missing covariates
- As an alternative to WGEE: MI can be combined with classical GEE
- Schematically:

$$\begin{array}{ll} \textbf{Imputation Task:} & \boxed{\textbf{Function to generate imputations}} \\ & \downarrow \\ \textbf{Analysis Task:} & \boxed{\textbf{Your favorite model function}} \\ & \downarrow \\ \textbf{Inference Task:} & \boxed{\textbf{Function for Rubin's combination rules}} \end{array}$$

MI Analysis of the ARMD Trial

- $M = 10$ imputations
- GEE:
$$\text{logit}[P(Y_{ij} = 1|T_i, t_j)] = \beta_{j1} + \beta_{j2} T_i$$
- GLMM:
$$\text{logit}[P(Y_{ij} = 1|T_i, t_j, b_i)] = \beta_{j1} + b_i + \beta_{j2} T_i, \qquad b_i \sim N(0, \tau^2)$$
- $T_i = 0$ for placebo and $T_i = 1$ for interferon-α
- t_j $(j = 1, \ldots, 4)$ refers to the four follow-up measurements
- Imputation based on the **continuous** outcome

- Results:

Effect	Par.	GEE	GLMM
Int.4	β_{11}	-0.84(0.20)	-1.46(0.36)
Int.12	β_{21}	-1.02(0.22)	-1.75(0.38)
Int.24	β_{31}	-1.07(0.23)	-1.83(0.38)
Int.52	β_{41}	-1.61(0.27)	-2.69(0.45)
Trt.4	β_{12}	0.21(0.28)	0.32(0.48)
Trt.12	β_{22}	0.60(0.29)	0.99(0.49)
Trt.24	β_{32}	0.43(0.30)	0.67(0.51)
Trt.52	β_{42}	0.37(0.35)	0.52(0.56)
R.I. s.d.	τ		2.20(0.26)
R.I. var.	τ^2		4.85(1.13)

When to Use Multiple Imputation?

- With missing outcomes (Y's) only, under MAR, and using likelihood/Bayes, ignorable likelihood/Bayes and MI are equivalent
- In that case, ignorable likelihood/Bayes is simpler
- But there are a number of settings where MI would be preferred:
  - ▷ When there are incomplete covariates X as well
  - ▷ When several researchers want to analyze the same incomplete set of data: MI will take care of the missingness for them all, in the same way
  - ▷ When using a non-likelihood/Bayes method, such as GEE
    - $*$ *MI-GEE generally tends to be more precise than WGEE*

- When a simple analysis is envisaged: e.g., a t test at a given time point in the study: direct likelihood would still force us to include **all** time points into the analysis. With MI, this 'multivariate aspect' is already taken care of at imputation time.

- For sensitivity analysis

Overview

MCAR/simple	CC LOCF	biased inefficient not simpler than MAR methods
MAR	**direct likelihood** **direct Bayes** **weighted GEE** **MI**	easy to conduct Gaussian & non-Gaussian
MNAR	variety of methods	strong, untestable assumptions most useful in **sensitivity analysis**