Complex Structured Data: Statistical Modeling

Edwin van den Heuvel
Professor in Statistics
December 2015

Technische Universiteit
Eindhoven
University of Technology

Where innovation starts

Introduction

- Mathematical Statistics
  - 1987-1991: MSc at UvA
  - 1991-1996: PhD at UvA
- Industrial Statistics
  - 1996-2002: Consultant at IBIS UvA
- Pharmaceutical Statistics
  - 2002-2010: Manager Organon/MSD
- Medical Statistics
  - 2010-2014: Professor at UMCG/RUG
- Statistics
  - 2014-…..: Professor at Mathematics

General Description Research

Compare and develop statistical models and techniques for the analysis of complex structured and incomplete data sets from observational and experimental studies

Application Areas:

- Measurement reliability
- Meta-analysis & harmonization
- Clinical Trials
- Life course epidemiology & Health Monitoring

Statistical Areas:

- Mixed Models
- Survival/Reliability Analysis
- Statistical Intervals
- Missing Data

General Description Research

Mixed Models: *Formulation*

- ***Linear***: $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i$
  - With $\boldsymbol{y}_i$ a vector of all outcomes on subject i
  - With $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ known design matrices
  - With $\boldsymbol{\beta}$ the vector of fixed effects
  - With $\boldsymbol{u}_i$ random effects $\mathbb{E}(\boldsymbol{u}_i) = \boldsymbol{0}$, $\mathrm{VAR}(\boldsymbol{u}_i) = \boldsymbol{G}(\boldsymbol{\theta})$
  - With $\boldsymbol{e}_i$ a vector of residuals $\mathbb{E}(\boldsymbol{e}_i) = \boldsymbol{0}$, $\mathrm{VAR}(\boldsymbol{e}_i) = \boldsymbol{R}(\boldsymbol{\eta})$
- ***Non-linear***: $\boldsymbol{y}_i = f(\boldsymbol{X}_i, \boldsymbol{\beta}, \boldsymbol{u}_i) + \boldsymbol{e}_i$
  - With f a general non-linear function
- ***Generalized linear***: $\mathbb{E}(\boldsymbol{y}_i|\boldsymbol{u}_i) = g^{-1}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i)$
  - With g the link function

General Description Research

Mixed Models: *Formulation*

- ***Conditional models***:

$$y_{i,t} = \sum_{k=1}^{p} [\beta_{ik} + u_{ik}(t)] y_{i,t-k} + e_{it}$$

  - With $y_{i,t}$ an outcome at time t for subject i
  - With $\boldsymbol{\beta}_i$ a vector of fixed effects (possibly as function of several baseline covariates)
  - With $\boldsymbol{u}_i(t)$ a stochastic process with $\mathbb{E}(\boldsymbol{u}_i(t)) = \boldsymbol{0}$, $\mathrm{VAR}(\boldsymbol{u}_i(t)) = \boldsymbol{G}(\boldsymbol{\theta})$
  - With $\boldsymbol{e}_i$ a vector of residuals $\mathbb{E}(\boldsymbol{e}_i) = \boldsymbol{0}$, $\mathrm{VAR}(\boldsymbol{e}_i) = \boldsymbol{R}(\boldsymbol{\eta})$
  - With $u_i(t)$ and e_{it} independent for each t
- They are often applied in econometrics and multivariate forms

General Description Research

Mixed Models: *Formulation*

- ***Joint models***:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i$$
$$P(T_i > t|\boldsymbol{u}_i) = S(t, \boldsymbol{u}_i)$$

  - With $\boldsymbol{y}_i$ a vector of longitudinal outcomes on subject i
  - With $T_i = \min(T_i^*, C_i)$ an observed survival time, T_i^* the true survival and C_i an independent censoring time
  - With $\boldsymbol{y}_i$ and T_i independently distributed, conditionally on the random effects $\boldsymbol{u}_i$
  - With S a survival function and all other terms as before
- The survival part is often modeled with a proportional hazard model

General Description Research

Mixed Models: *Parameter Estimators*

- Through (Restricted) Maximum likelihood or Generalized Estimating Equations
- Are biased in many settings, in particular for variance components
- Are difficult to obtain for large data sets
- Have unknown finite distributions, which complicates construction of intervals
- Are seriously affected by missing data issues, which requires special modeling

Measurement Reliability

- A measurement system is a systematic and replicable set of steps to quantify or classify objects with respect to a certain dimension or unit by assignment of (a set of) numbers
- Types of measurements:
  - Engineering & Physics (dimensions, strength)
  - Chemical (concentrations)
  - Socio-Economic (income, status)
  - Psychology (memory, IQ, cognition)
  - Medical (physical activity, frailty)
  - Biological (bioassay, microbiology)

Measurement Reliability

Historical:

- Implicitly developed by Johann Carl Friedrich Gauss (1777-1855) for calculations of orbits of planets – discovery of Ceres
- He realized that observations are not without error and assumed the following statistical model

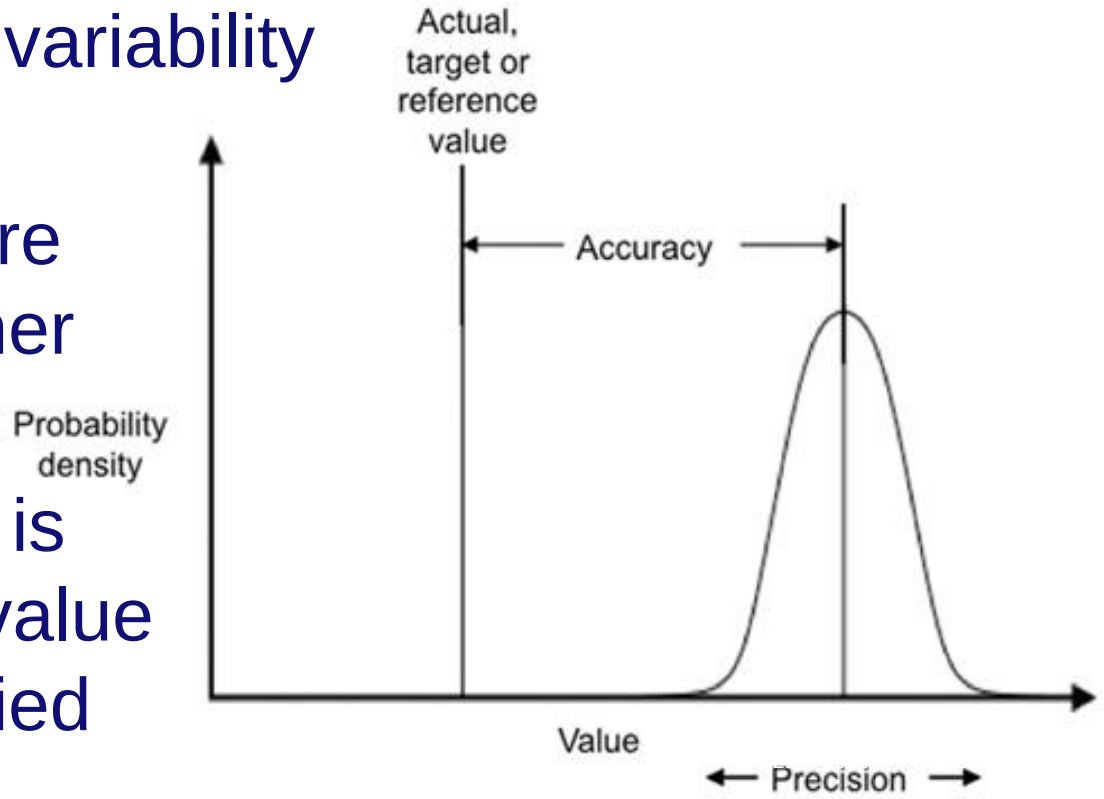
Observed value	=	True value	+	Error

- This model was later extended to allow for possible systematic differences

Measurement Reliability

Characteristics:

- ***Accuracy & linearity*** discuss how well the true value can be recovered from reference material
- ***Precision*** discuss variability in error term
- ***Specificity*** measure the influence of other constituents
- ***Quantitation limit*** is the minimum true value that can be quantified



Measurement Reliability

Research Results:

- Estimation of measurement variability
  - *Quality Engineering*, 2002, **15**(2):323-331.
  - *Quality Engineering*, 2005, **17**(4):495-507.
  - *Quality and Reliability Engineering International*, 2005, **21**:491-508.
  - *Quality Engineering*, 2015, Accepted.
- Confidence intervals on precision
  - *Journal of Biopharmaceutical Statistics*, 2007, **17**:1-20.
  - *Communications in Statistics - Simulation and Computation*, 2010, **39**(4):777-794.
- Validation of (micro)biological methods
  - *Vaccine*, 2012, **30**(2):201-209.
  - *Journal of Microbiological Methods*, 2010, **82**(3):193-197.
  - *Pharmaceutical Statistics*, 2011, **10**(3):203-212.
  - *Pharmaceutical Statistics*, 2013, **12**(5):291-299.
  - *Pharmaceutical Statistics*, 2015, **44**(2):120-128.

Measurement Reliability

Research Results:

- Data handling approaches:
  - *Pharmaceutical Statistics*, 2013, **12**(6):375-384.
  - *Analytical Chemistry*, 2015: DOI: 10.1021/acs.analchem.5b02832
- Agreement in medical diagnosis
  - *Radiation Oncology*, 2012, **7**(32):1-9.
  - *Developmental Medicine & Child Neurology*, 2013, **55**(6):539-545.
  - *Statistical Methods in Medical Research*, 2014: 096228021452278 7.
  - *Journal of Clinical Oncology*, 2015, **33**(4):349-356.
  - *Manual Therapy*, 2015, **20**(4):580-586.
- Validation of questionnaires
  - *Schizophrenia Research*, 2013, **147**(1):175-180.
  - *Schizophrenia Research*, 2013, **150**(2-3):410-415.
  - *ACTA Dermato-Venereologica*, 2014, **94**(4):442-447.

Measurement Reliability

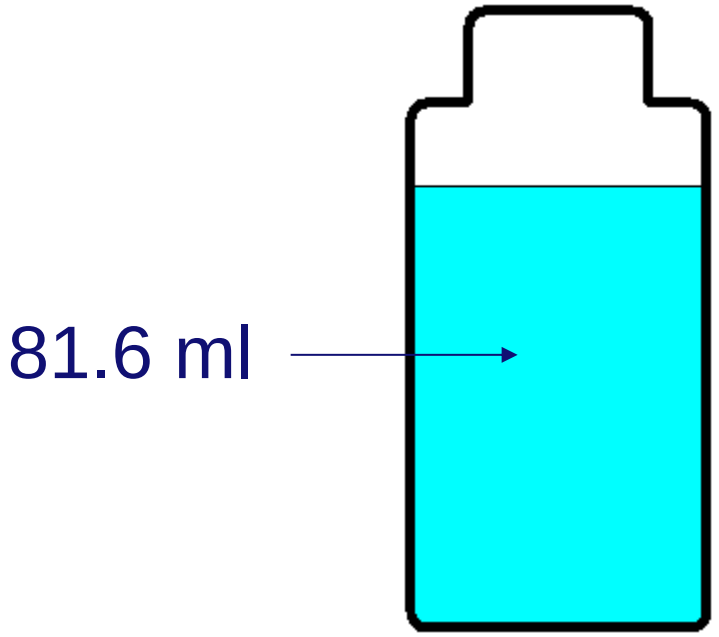
Validation of Microbiology:

- Medicinal products should be bacterial free
- Classical methods: growth-based and slow
- New and rapid methods are being developed and implemented
- The performance must be tested, but spiking low and precise numbers of organisms is impossible
- Need clever designs and statistics to estimate ***limit of detection***

Measurement Reliability

Validation of Microbiology:

Volume V_t

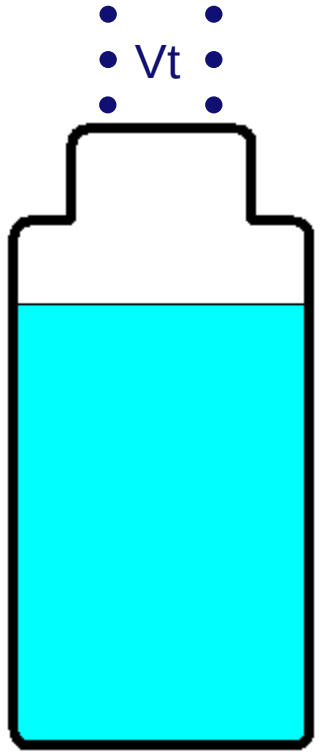


Measurement Reliability

Validation of Microbiology:

Spike:

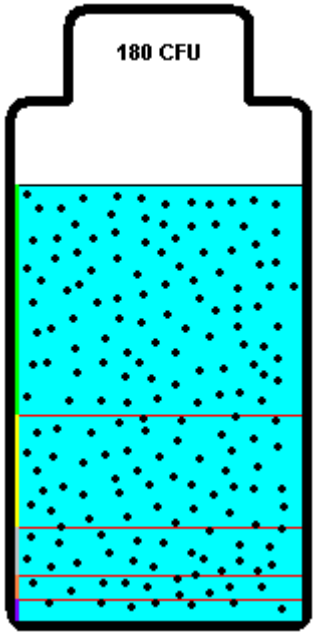
- 6 Bioballs of
- ± 30 colony forming units



Measurement Reliability

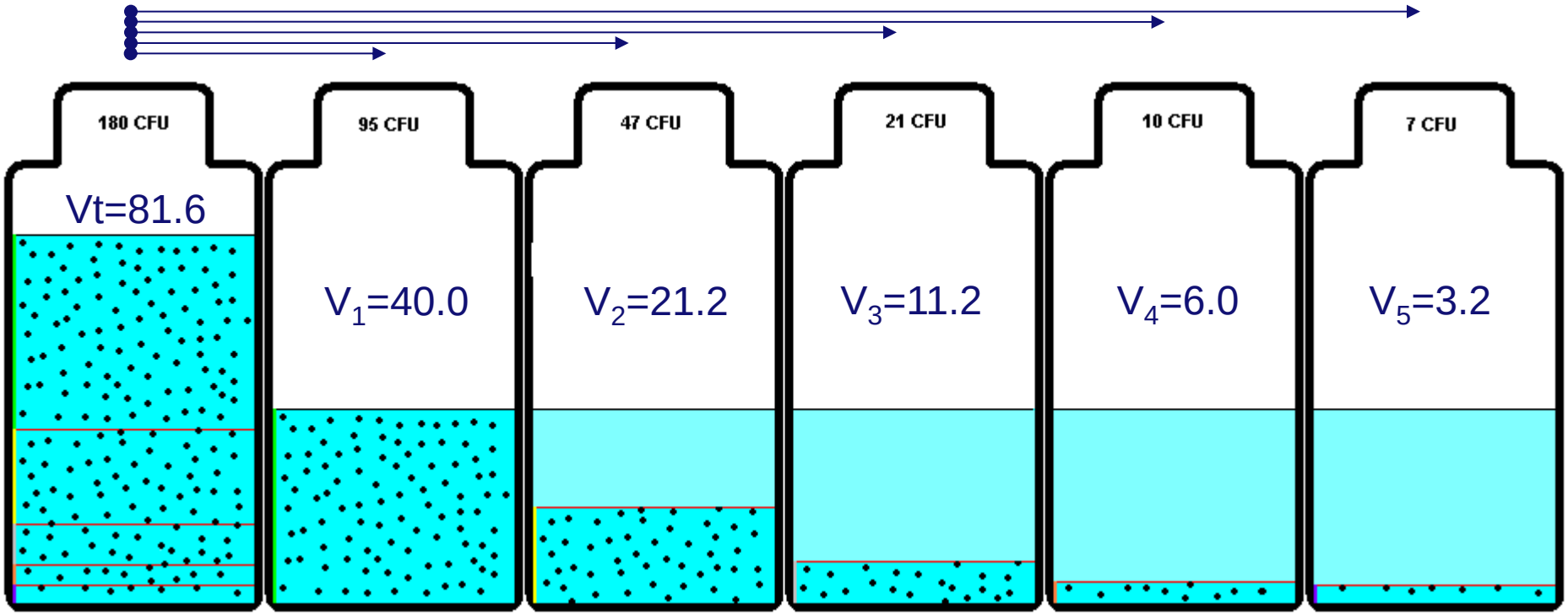
Validation of Microbiology:

- Solution is firmly shaking not stirred
- Homogeneous solution of ± 180 CFU



Measurement Reliability

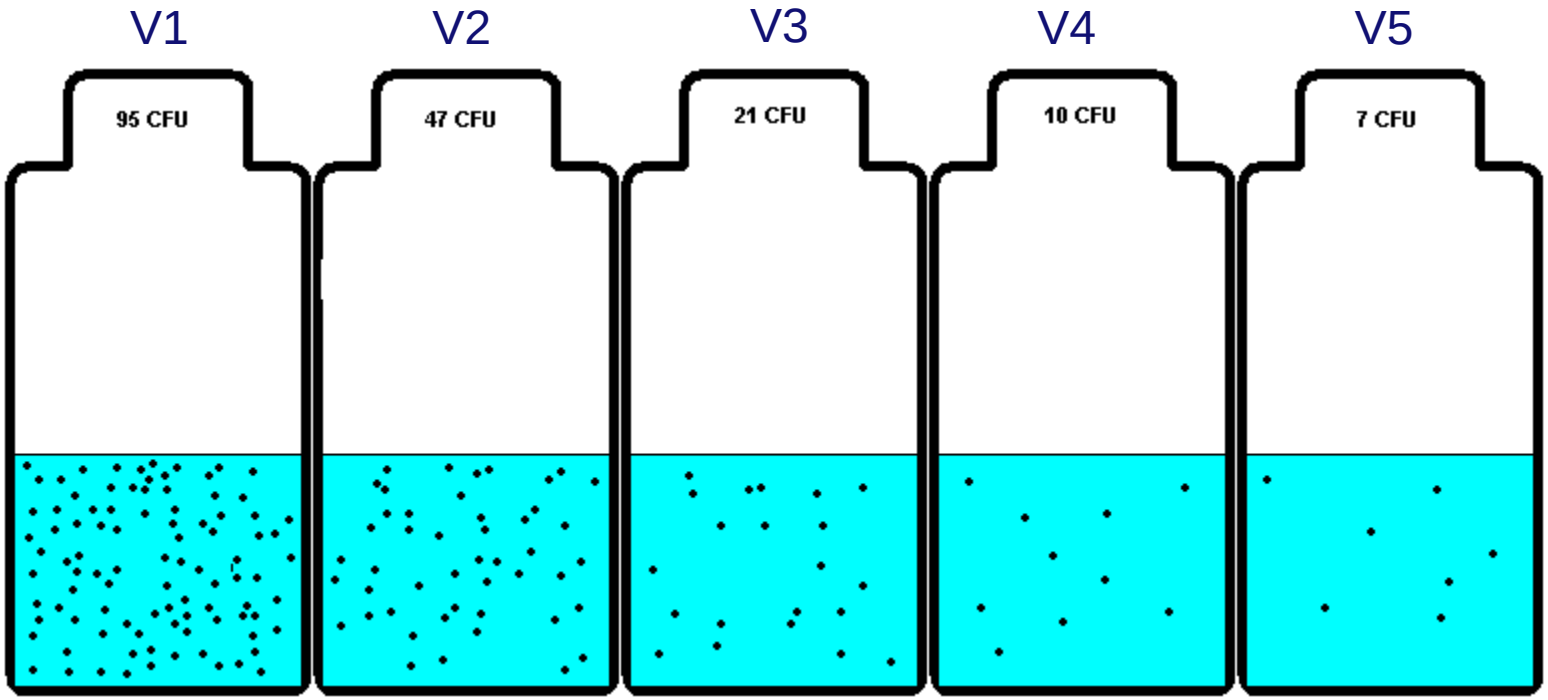
Validation of Microbiology:



Dilution ratio R=0.53

Measurement Reliability

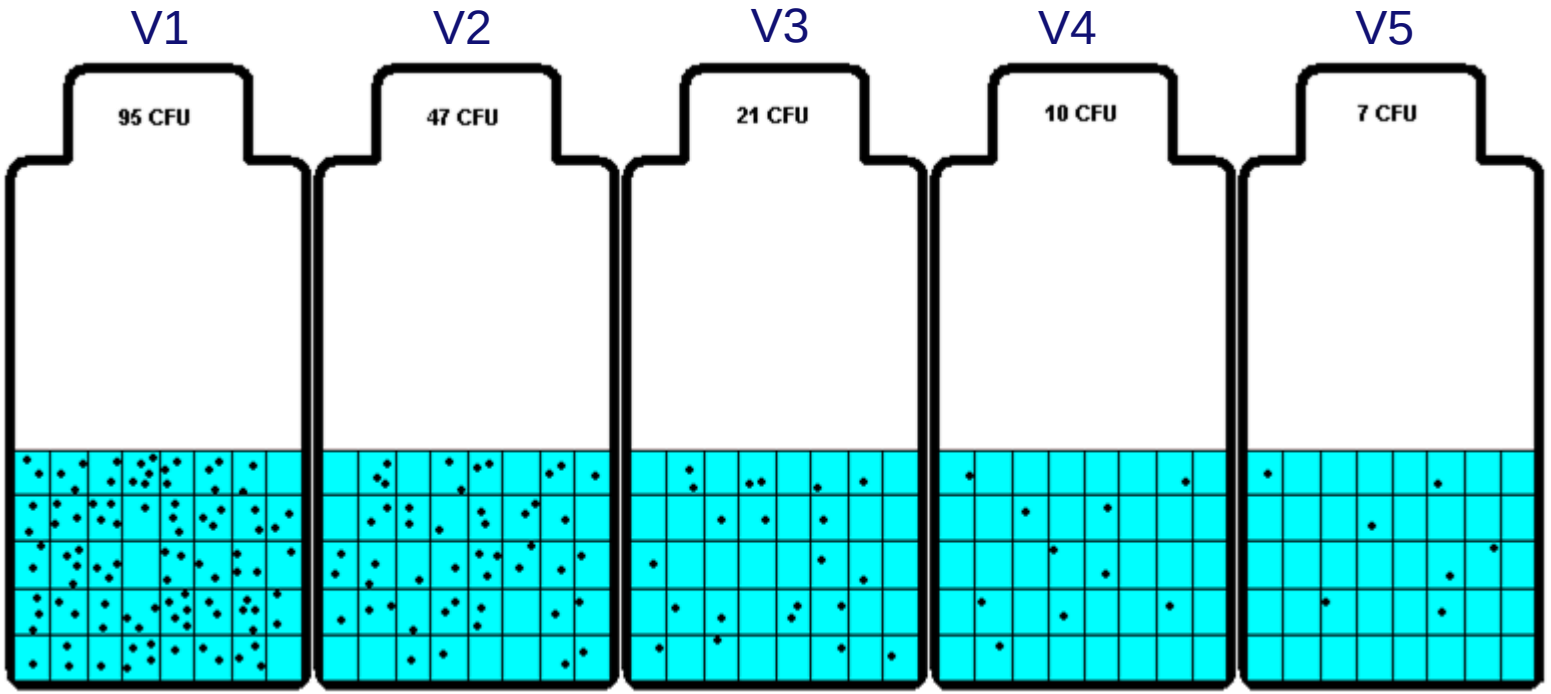
Validation of Microbiology:



Solutions are firmly shaken not stirred

Measurement Reliability

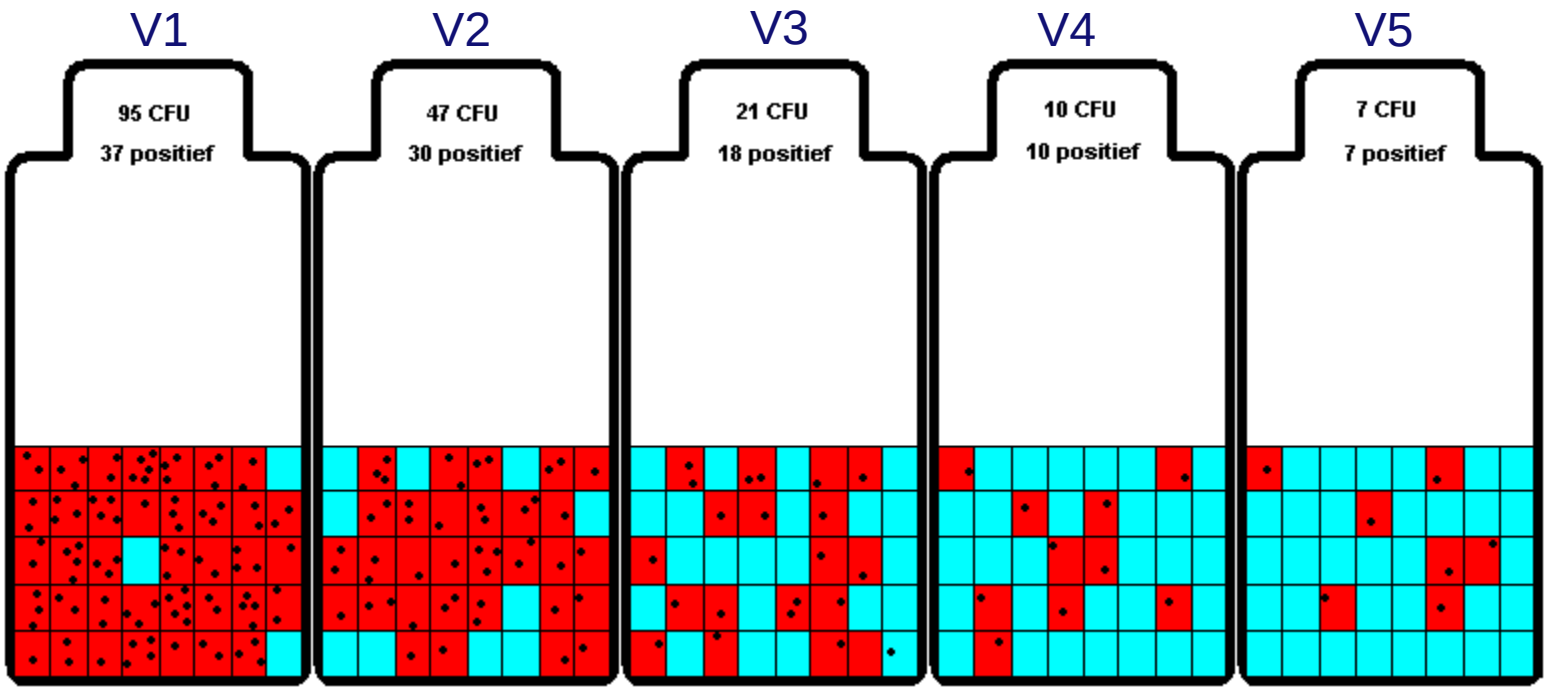
Validation of Microbiology:



Divide each dilution over 40 incubation bottles

Measurement Reliability

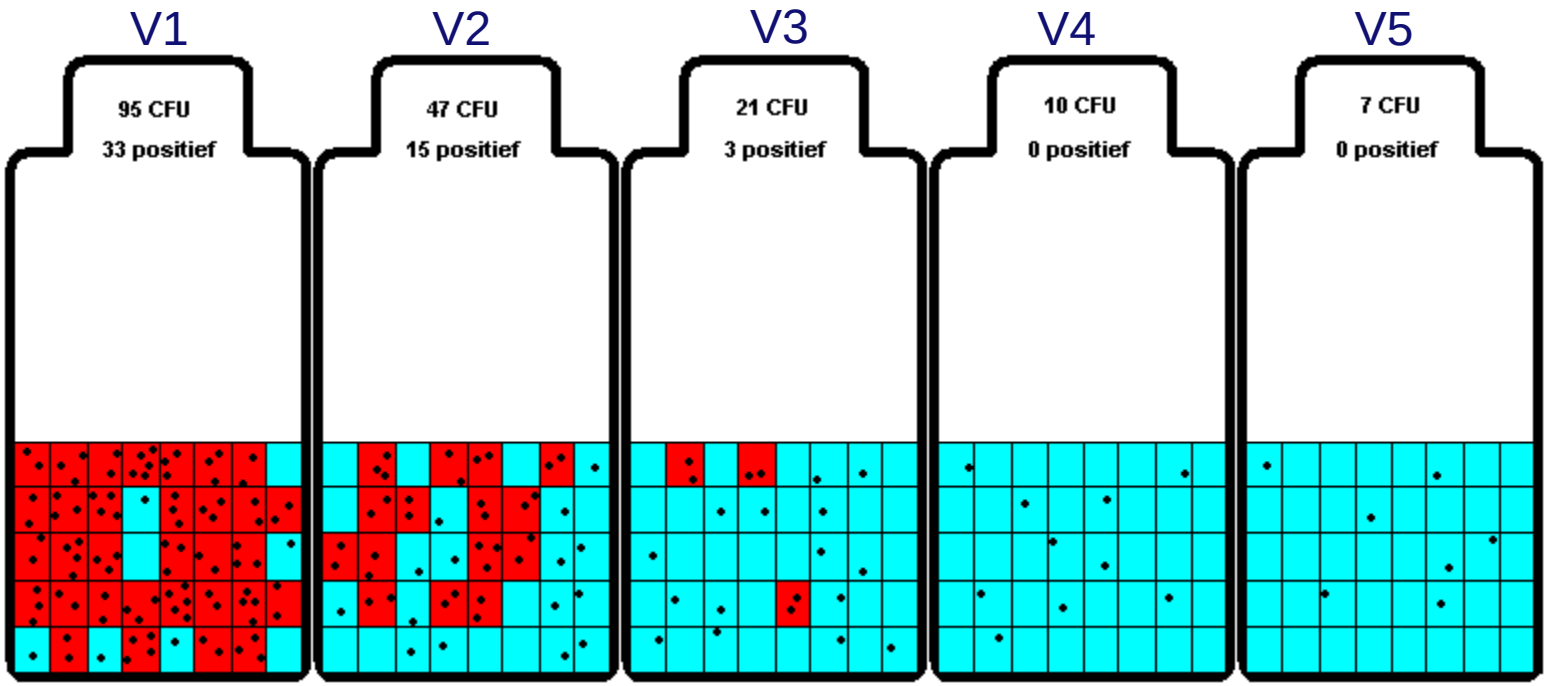
Validation of Microbiology:



Limit of detection = 1 ⇒ Red samples positive

Measurement Reliability

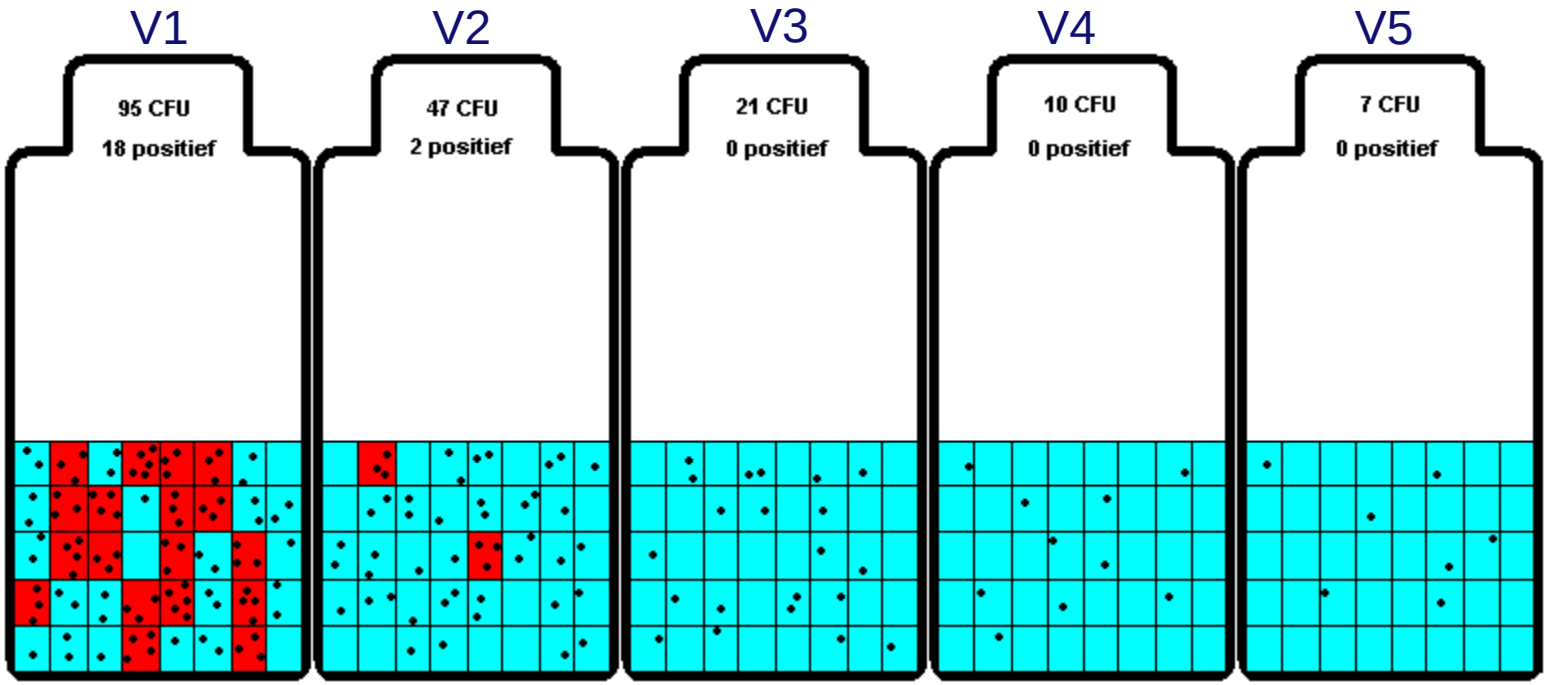
Validation of Microbiology:



Limit of detection = 2 $\Rightarrow$ Red samples positive

Measurement Reliability

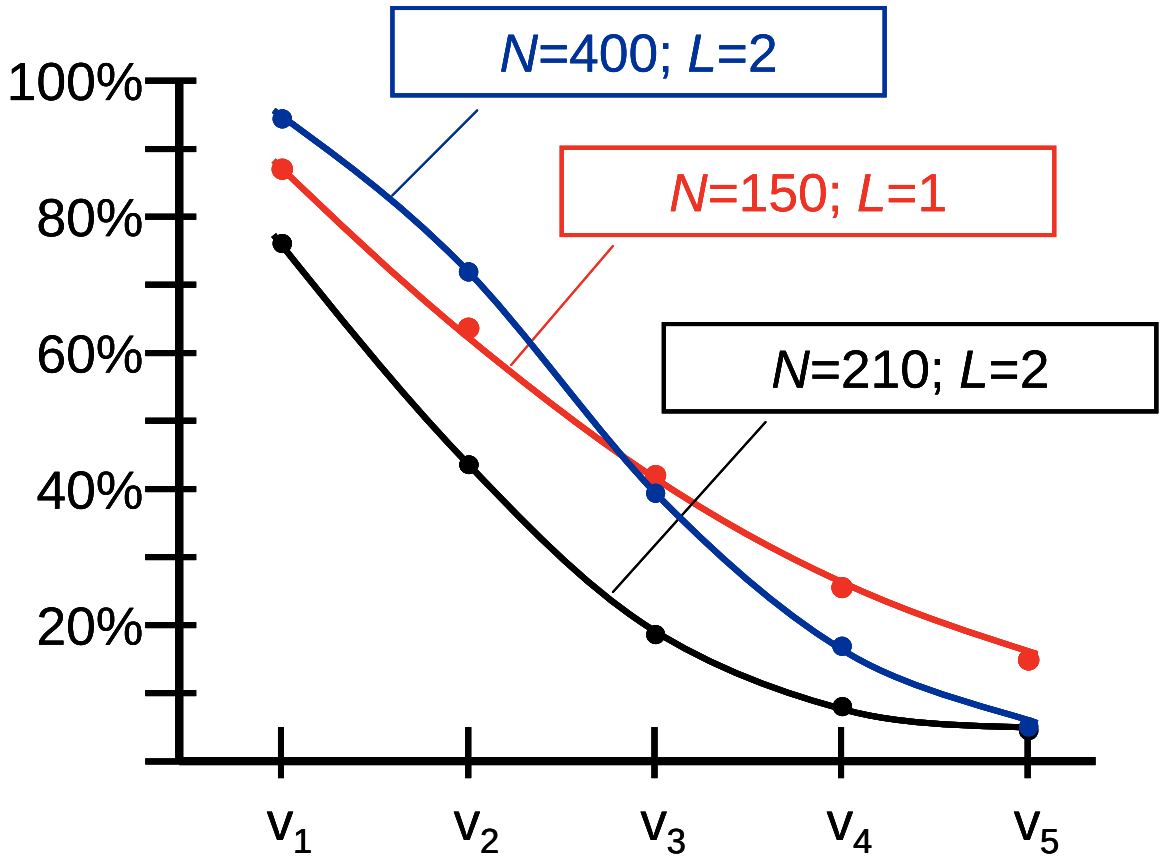
Validation of Microbiology:



Limit of detection = 3 $\Rightarrow$ Red samples positive

Measurement Reliability

Validation of Microbiology:



Results make it possible to estimate both the spike *N* and limit of detection *L*

Measurement Reliability

Validation of Microbiology:

- We applied Pearson's minimum chi-square statistic (an not maximum likelihood)

$$\chi^2(L,N) = \sum_{i=1}^{K} \frac{(U_i - \mu_i(L,N))^2}{n_i P_i(L,N)\left(1 - P_i(L,N)\right)}$$

  - With U_i number of successes (out of n_i)
  - With $\mu_i(L,N) = n_i\left(1 - P_i(L,N)\right)$
  - With $P_i(L,N) = \sum_{x=0}^{L-1} \binom{N}{x} \left(\frac{V_i}{n_i V_t}\right)^x \left(1 - \frac{V_i}{n_i V_t}\right)^{N-x}$
- Alternative statistical detection models have been investigated and published

Measurement Reliability

Research Plans:

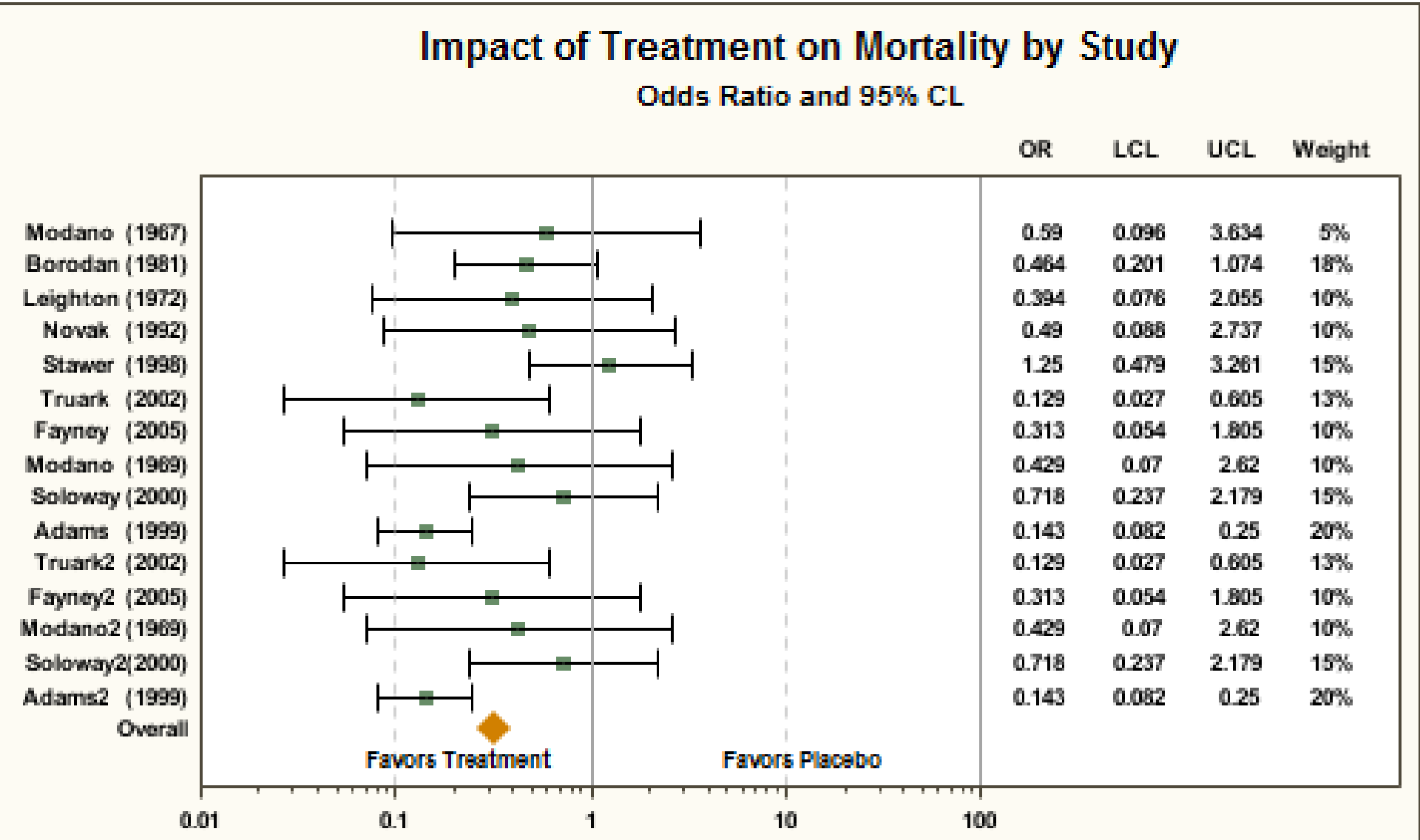
- Validation of qualitative and quantitative microbiological methods:
  - Maximum likelihood issues
  - Deconvolution issues
  - Optimal designs for robustness
  - STW project proposal submitted (2 PhD's & Software Engineer)
- Validation of high-complex chemical analysis
  - Missing data analysis (Post-doc)
  - Bootstrap & cross-validation (DSM)

Meta Analysis & Harmonization

"Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature"

Glass (1976)

Meta Analysis & Harmonization



Meta Analysis & Harmonization

"…popular practice of analysing summary measures from selected publications is a poor man's solution."

"…I hope that we will have full multi-center multi-study databases that can be analysed by appropriate random effects models considering both random variation within and between studies and/or centres."

"…there is no 'meta'-aspect on the analysis anymore and the term 'meta-analysis' can be skipped from the dictionary."

Van Houwelingen (1997)

Meta Analysis & Harmonization

Research Result:

Lancet Infectious Disease, 2014, **14**(12):1228-1239.

Effectiveness of seasonal influenza vaccine in community-dwelling elderly people: a meta-analysis of test-negative design case-control studies

Maryam Darvishian, Maarten J Bijlsma, Eelko Hak, Edwin R van den Heuvel

Journal of Clinical Oncology, 2015, **33**(4):349-356.

Magnetic Resonance Imaging Improves Breast Screening Sensitivity in *BRCA* Mutation Carriers Age ≥ 50 Years: Evidence From an Individual Patient Data Meta-Analysis

Xuan-Anh Phi, Nehmat Houssami, Inge-Marie Obdeijn, Ellen Warner, Francesco Sardanelli, Martin O. Leach, Christopher C. Riedl, Isabelle Trop, Madeleine M.A. Tilanus-Linthorst, Rodica Mandel, Filippo Santoro, Gek Kwan-Lim, Thomas H. Helbich, Harry J. de Koning, Edwin R. Van den Heuvel and Geertruida H. de Bock

Journal of Clinical Psychopharmacology, 2013, **33**(5):675-681.

Estimating Dopamine D_2 Receptor Occupancy for Doses of 8 Antipsychotics

A Meta-Analysis

Irene M. Lako, PhD,† Edwin R. van den Heuvel, PhD,‡ Henrikus Knegtering, MD, PhD,†§ Richard Bruggeman, MD, PhD,*†// and Katja Taxis, PhD**

Meta Analysis & Harmonization

Challenges Individual Participant Data

- Harmonize data from different instruments?
  - *Journal of Clinical Epidemiology*, 2015, **68**(2), 154-162.
  - Two submitted papers using latent variable models and standardization
- Pooling longitudinal data from different studies?
- Analyze without sharing data at one location?
  - *International Journal of Epidemiology*, 2014, **43**(6):1929-1944.
  - *Biopreservation and Biobanking,* 2015, **13**(3):178-182.

Correct	CSHA (n=1730)			NuAge (n=432)	
# words	Rey	Free B	Cued B	Free B	Cued B
0	4.57	4.28	0.23	0.23	0
1	7.28	2.31	0.29	0	0
2	13.1	3.70	0.58	0.69	0
3	21.6	5.09	0.40	3.24	0
4	17.7	8.61	0.58	6.48	0.23
5	13.1	10.9	0.81	13.4	0
6	5.32	14.0	1.27	14.1	0.69
7	2.60	17.4	2.08	16.0	0.93
8	1.05	14.3	2.66	14.8	0.93
9	0.06	10.7	4.22	12.3	2.31
10	0.17	6.42	7.11	8.80	1.16
11	0	2.02	16.5	4.63	5.56
12	0	0.29	63.2	3.24	7.64
13	0	Na	Na	1.85	10.6
14	0	Na	Na	0.23	18.8
15	0	Na	Na	0	19.4
16	Na	Na	Na	0	31.7

Meta Analysis & Harmonization

Research Plans:

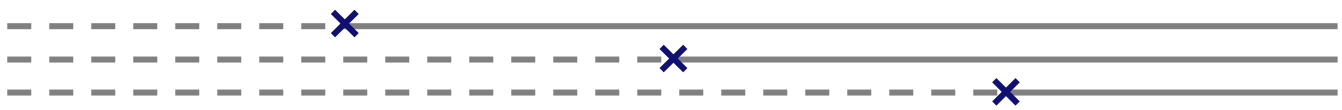
- Improve existing methods of meta-analysis
  - PhD Student – NWO Scholarship for teachers
- Confidence interval on heterogeneity (ICC's)
  - Based on Beta-distribution (*Biometrics*, 2015, **71**(2):548-555)
  - Submitted a grant – Develop theory and R-package
- Models for vaccine effectiveness
  - PhD Student – finishes 2016
- Harmonization
  - PhD Student in McMaster University
  - Master Student TU/e – missing data

Clinical Trials

Stepped Wedge Design:

- All patients are first treated with the control
- Groups of patients (or patients) are changing at different switch moments to the new intervention
  - Switch moments are determined upfront
- Example: Three groups with four periods

SWD-S3



  - Dotted lines indicate control treatment
  - Solid lines indicate new treatment
  - Crosses indicate switch moments

Clinical Trials

Stepped Wedge Design:

- Two types of stepped wedge designs (SWDs):
  - Cross-sectional: in each cluster-period combination different subjects are recruited
  - Longitudinal: recruitment of subjects starts at the beginning and are followed until the end
- Most research work has been conducted for cross-sectional SWD's
  - Statistical analyses: Hussey and Hughes (2007)
  - Calculation of sample sizes: Woertman *et al.* (2013)
  - Comparison with CRCT: Hemming *et al*. (2015)
  - Optimal stepped wedge design: Lawrie *et al*. (2015)

Clinical Trials

Hussey and Hughes (2007):

- Let Y_{ijk} be the response for subject $k(1,2,\dots,m_{ij})$ in period $j(1,2,\dots T)$ for cluster $i(1,2,\dots,C)$
- The proposed cross-sectional mixed model is

$$Y_{ijk} = \mu + a_i + \beta_j + \gamma \cdot x_{ij} + e_{ijk}$$

  - With μ the overall mean
  - With $a_i \sim N(0, \tau^2)$ an i.i.d. random cluster effect
  - With β_j a fixed time effect ($\beta_T = 0$ for identifiability)
  - With $x_{ij} \epsilon \{0,1\}$ a treatment indicator variable
  - With γ the treatment effect
  - With $e_{ijk} \sim N(0, \sigma^2)$ i.i.d. residuals

Clinical Trials

Van den Heuvel (2014): Comparing Designs

- For Y_{ij} response of subject i at time t_{ij} the longitudinal mixed model

$$Y_{ij} = Z_{i0} + Z_{i1} \cdot t_{ij} + \gamma\left(t_{ij} - x_{ij}\right) \cdot 1_{(x_{ij},\infty)}\left(t_{ij}\right) + e_{ij}$$

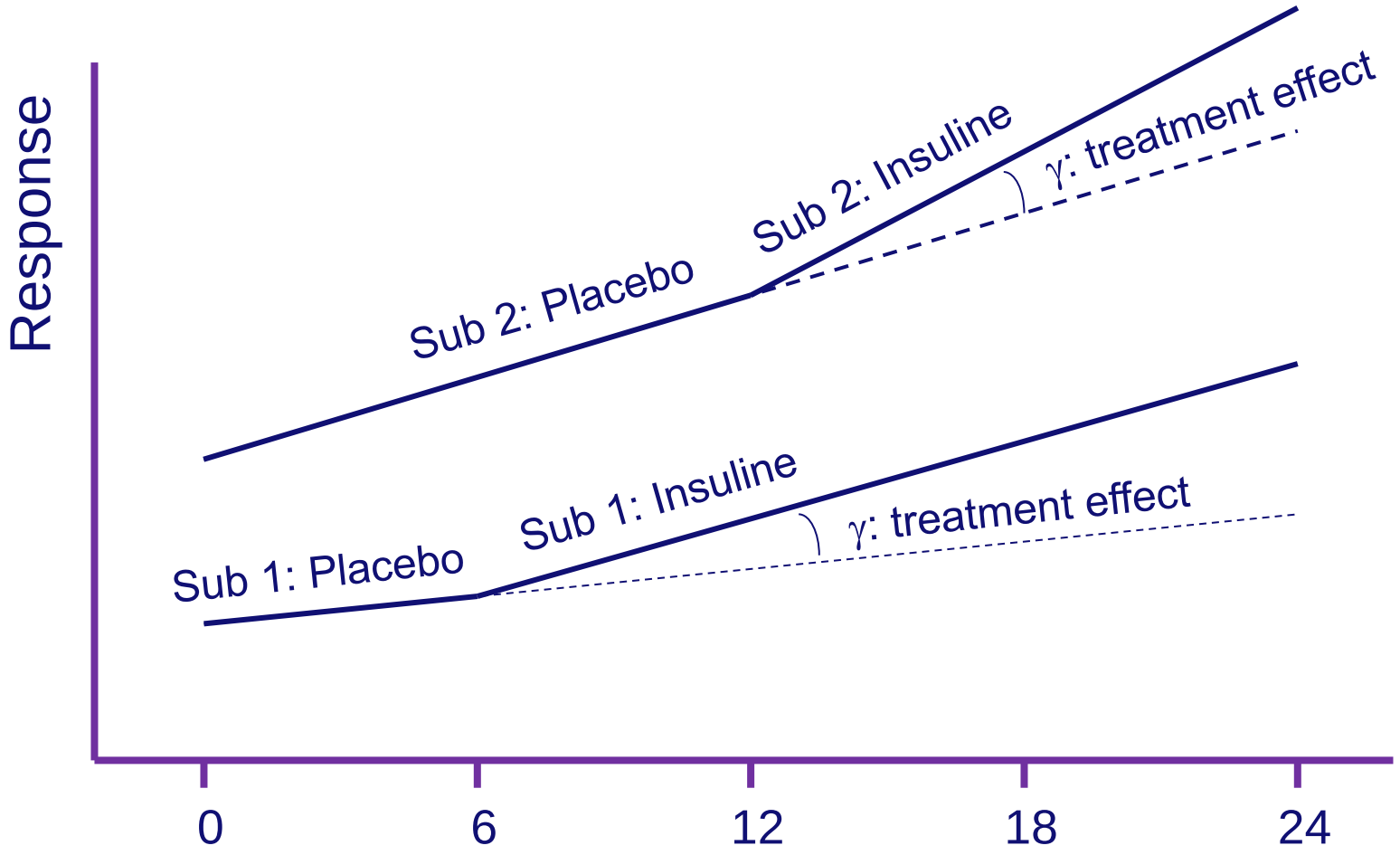
$$\begin{pmatrix} Z_{i0} \\ Z_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{pmatrix}\right)$$

  - With t_{ij} time points of measurements
  - With x_{ij} switch moments for treatment
  - With $e_{ij} \sim N(0, \sigma_0^2)$ i.i.d. residuals
  - With γ treatment effect on growth

Statistical Methods in Medical Research, 2014: 0962280214558864.

Clinical Trials

Random Coefficients Model:



Clinical Trials

Research Plans

- PhD Student – longitudinal SWD's
  - Overview paper (*Journal of Clinical Epidemiology*, 2014, 67(4):454-461)
  - Screening cancer (*European Journal of Surgery Oncology*, 2015, 41(9):1188-1196)
  - Survival analysis of terminal end-points – paper submitted
  - Optimal designs (numbers and switches) – work in progress
  - Growth mixture models – Collaboration Balakrishnan McMaster University
- Unsure how to follow-up research

Life Course Epidemiology

Aims to understand how risk factors, that operate across an individual's life course or across generations, affect each other and how they simultaneously affect the development of disease outcome

- Analysis of ***longitudinal data*** with
  - Focus on understanding temporal relationships
  - Individual changes – health monitoring
- Phenotype and genotype related issues
  - Multiple area's of disease
  - Multiple research disciplines (social, psychological, and biological information)

Life Course Epidemiology

Research Results:

- Analysis of longitudinal observational data
  - *Many publications in medical and psychiatric journals*
  - *Biometrics*, 2015, DOI: 10.1111/biom.12414
  - *Computational Statistics and Data Analysis*, 2014, 77:70-83.
  - *Three submitted papers to statistical journals*
- Analysis of longitudinal clinical trials
  - *Many publications in medical and psychiatric journals*
- Causal Inference
  - *Human Reproduction*, 2014, 29(3): 510-517.
  - *Journal of Clinical Epidemiology,* 2014, 67(2):190-198.
  - *Developmental Medicine & Child Neurology*, 2013, 55(11):976-976.
  - *Two submitted papers*

Life Course Epidemiology

Research Plans

- Generalizations of mixed models
  - Longitudinal and time-to-event data: joint modeling
  - Location and error variance both random
  - Clustering of mixed models (PhD Student Philips)
  - Conditional and non-parametric models for high-frequent data (PhD Student Philips)
  - Plans for ERC or TOP Grant
- Collaboration with Framingham Heart Study:
  - Observational longitudinal data: more than 30 years of follow-up