Dlaczego (czasem) R?
Why (sometimes) R?

Tomasz Burzykowski

I-BioStat, Hasselt University, Belgium, and

International Drug Development Institute (IDDI), Belgium

tomasz.burzykowski@uhasselt.be

© picture-alliance/dpa/B. Reisfeld

Appointments

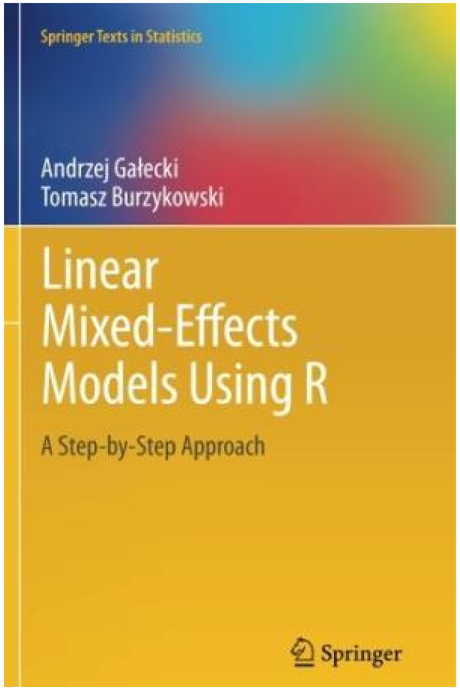
Academic

- 1992-1998 Biostatistician, Institute of Oncology, Warsaw
- 1998-2001 PhD researcher, Hasselt University, Belgium
- 2002-present Professor, Hasselt University, Belgium
- 2011-present Visiting Professor, Medical U of Bialystok
- 2011-present Guest Lecturer, Technical U of Warsaw

Professional

- 2006-2008 Senior Biostatistician, MSOURCE Medical Development, Warsaw
- 2009-present Vice-President of Research, International Drug Development Institute (IDDI), Belgium

R-related experience



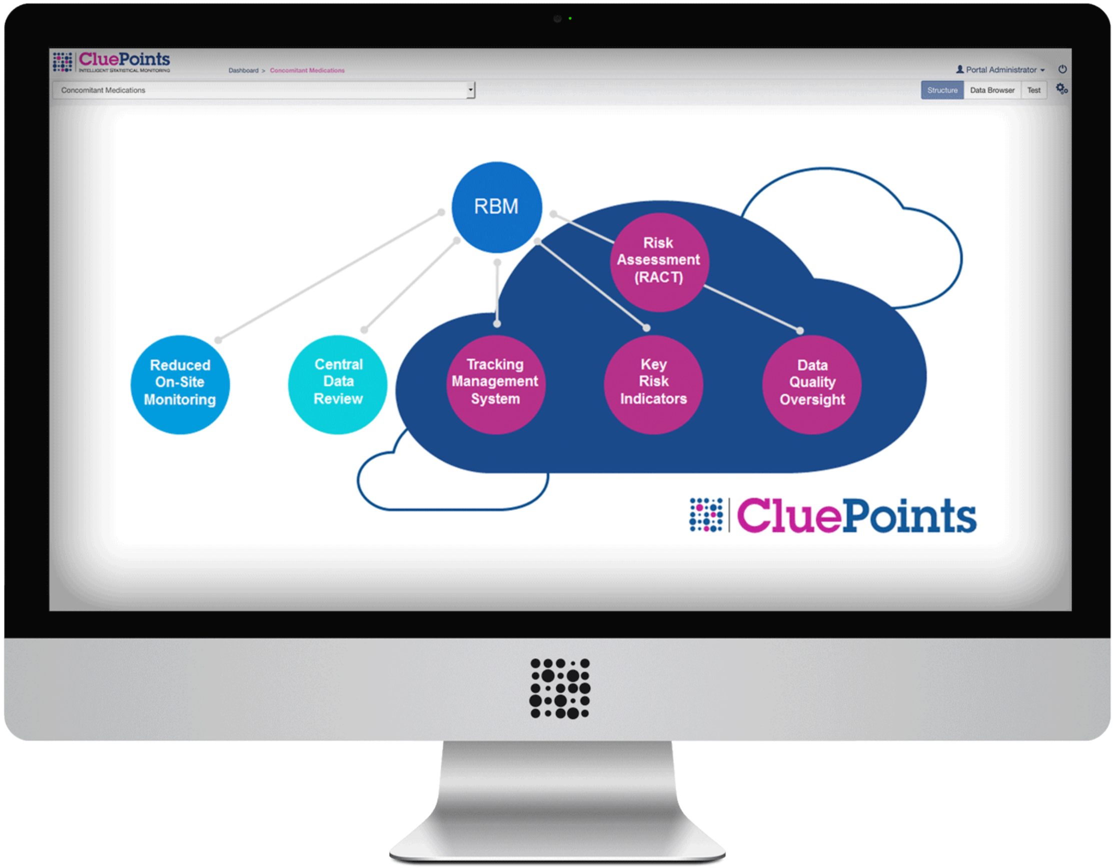
Springer Texts in Statistics

Andrzej Gałecki
Tomasz Burzykowski

Linear Mixed-Effects Models Using R

A Step-by-Step Approach

Springer



How Many Packages Do You Know That Fit Linear Mixed-effects Models?

How Many Packages Do You Know That Fit Linear Mixed-effects Models?

- R: arm, gamm, gamm4, GLMMarp, glmmAK, glmmBUGS, heavy, HGLMMM, lme4, lmec, lmm, longRPart, MASS, MCMCglmm, nlme, PSM, pedigreemm, ...
- SAS: PROC MIXED, NLMIXED, HPMIXED
- STATA: xtmixed
- SPSS: MIXED
- MLWin
- Mplus
- HLM
- ASREML
- ...

No Single Statistical Software Offers All Methods With All Required Features

Software That I Have Ever Used

- BMDP
- GLIM
- SPSS
- Statistica
- R
- SAS
- STATA

Software That I Have Ever Used

- ~~BMDP~~
- ~~GLIM~~
- ~~SPSS~~
- ~~Statistica~~
- R
- SAS
- STATA

Teaching

- Statisticians/stats students
  - focus on (advanced) methodology
  - reasonable implementation
  - documentation
  - stability of the code

- Biomed students/professionals
  - focus on applications
  - user-friendly implementation
  - error handling
  - stability of the code

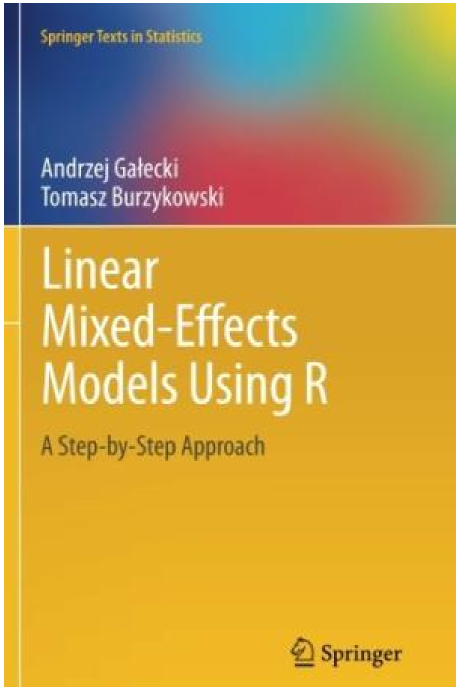
Teaching

- Statisticians/stats students:
  - R
    - free
    - availability of modern methods
    - programming/simulations
  - SAS
    - good coverage of methods
    - good documentation and error handling
    - preparation for working in the pharma industry
    - stability
- Biomed students
  - STATA (biostatistics)
    - user-friendly
    - good coverage of methods
    - very good documentation and error handling
    - stability
  - R (bioinformatics)

Research

- Used on a daily basis
  - STATA (good range of methods, user-written commands, very good documentation, error handling)
  - SAS (data management, LMMs, survival analysis, good documentation, error handling, IML programming)
  - R (new methodology, bioinformatics applications)

R-related Experience



Springer Texts in Statistics

Andrzej Gałecki
Tomasz Burzykowski

Linear Mixed-Effects Models Using R

A Step-by-Step Approach

Springer

- December 2011: final text ready
- Release of R 2.14.1: December 22
- Models give different results...

Consulting

- Academic: anything which works
- Pharma:
  - SAS is treated as the standard
  - other packages according to needs
    - sometimes methods “translated back” into SAS

Statistical Software in the Pharma Context

- Heavily regulated environment (e.g., FDA “guidelines”)
  - Applicable general documents:
    - 21 CFR Part 11 - Electronic Records; Electronic Signatures
    - Guidance for Industry: Part 11, Electronic Records
    - 21 CFR Part 58 - Good Laboratory Practice (GLP)
    - 21 CFR Part 312 - Good Clinical Practice (GCP)
    - 21 CFR Part 210 - Current Good Manufacturing Practice (cGMP)
    - ICH E6 - Good Clinical Practice Consolidated Guideline
  - Software guidance documents:
    - Guidance for Industry - Computerized Systems Used in Clinical Investigations (2007)
    - General Principles of Software Validation; Final Guidance for Industry and FDA Staff (2002)
  - Statistical guidance documents:
    - ICH E9 - Statistical Principles for Clinical Trials
    - Guidance for Industry and FDA Staff - Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (2010)

Statistical Software in the Pharma Context

"FDA's guidance documents do not establish legal enforceable responsibilities.

Instead, guidances describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited.

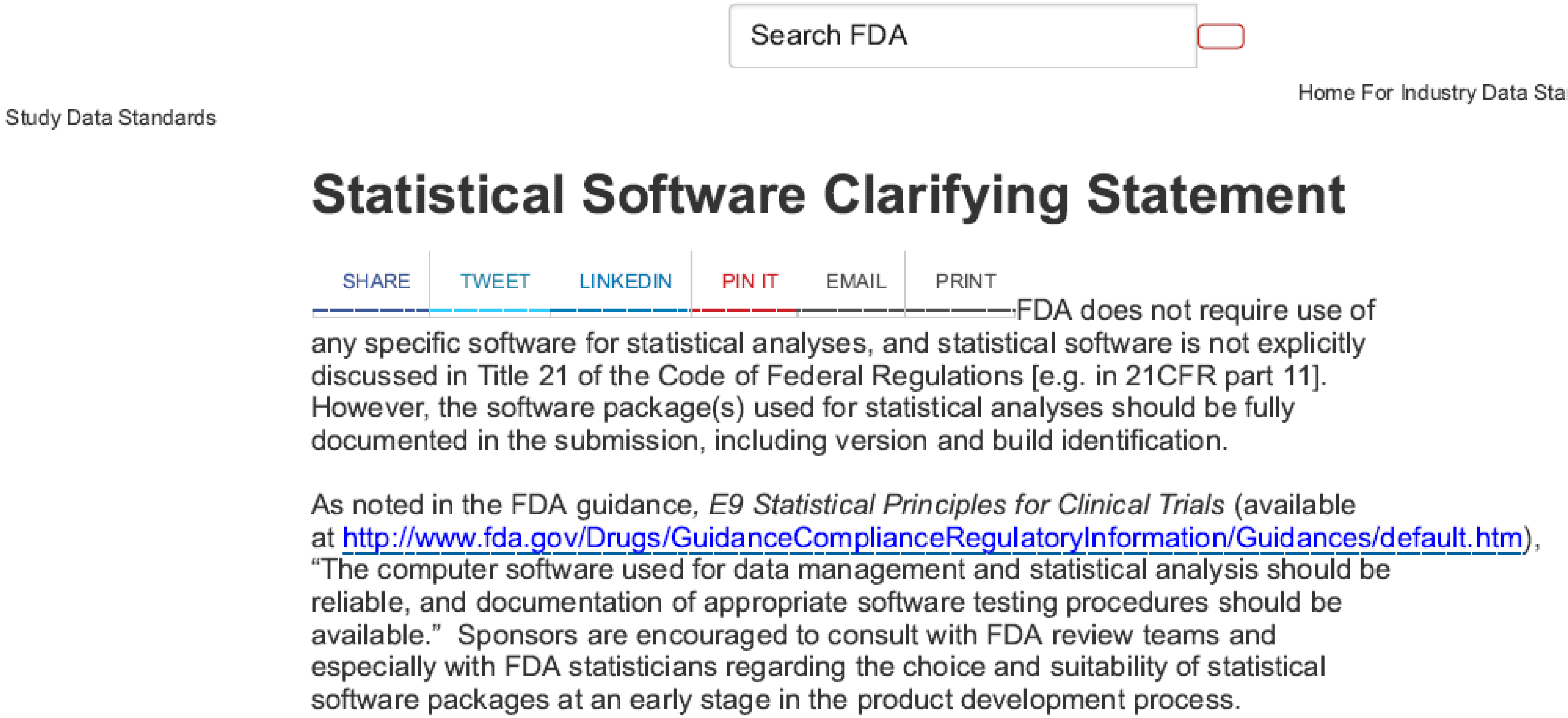
The use of the word should in Agency guidances means that something is suggested or recommended, not required."

Statistical Software in the Pharma Context

- FDA requests xpt files
  - The analyses are repeated!
- “SAS export files”
  - actually, open standard
- SAS assumed as a “standard”

Paul Schuetty’s presentation: https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/Using-R-in-a-regulatory-environment-FDA-experiences

Statistical Software in the Pharma Context



Statistical Software Clarifying Statement

FDA does not require use of any specific software for statistical analyses, and statistical software is not explicitly discussed in Title 21 of the Code of Federal Regulations [e.g. in 21CFR part 11]. However, the software package(s) used for statistical analyses should be fully documented in the submission, including version and build identification.

As noted in the FDA guidance, *E9 Statistical Principles for Clinical Trials* (available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm), "The computer software used for data management and statistical analysis should be reliable, and documentation of appropriate software testing procedures should be available." Sponsors are encouraged to consult with FDA review teams and especially with FDA statisticians regarding the choice and suitability of statistical software packages at an early stage in the product development process.

https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm445917.htm

Statistical Software in the Pharma Context

- FDA Stats Reviewers typically use SAS, but also R, Minitab, and STATA

- Internal use of R for
  - graphics
  - simulations
  - R & D

Paul Schuetty's presentation: https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/Using-R-in-a-regulatory-environment-FDA-experiences

R in the Pharma Context

- https://www.r-project.org/doc/R-FDA.pdf

- "... to demonstrate that R, when used in a qualified fashion, can support the appropriate regulatory requirements for validated systems, thus ensuring that resulting electronic records are 'trustworthy, reliable and generally equivalent to paper records.' "

- "Base R plus Recommended Packages" released in both source code and binary executable forms under the Free Software Foundation's GNU Public License

IDDI

ADDING VALUE TO CLINICAL DATA

TRIAL DESIGN
INTEGRATED IWRS / EDC
eCRF DESIGN DATA MANAGEMENT
STATISTICAL ANALYSES / KRI's
IDMC SUPPORT
BIOMARKERS VALIDATION
MEDICAL WRITING
APPROVED
ASSISTANCE TO FDA/EMA

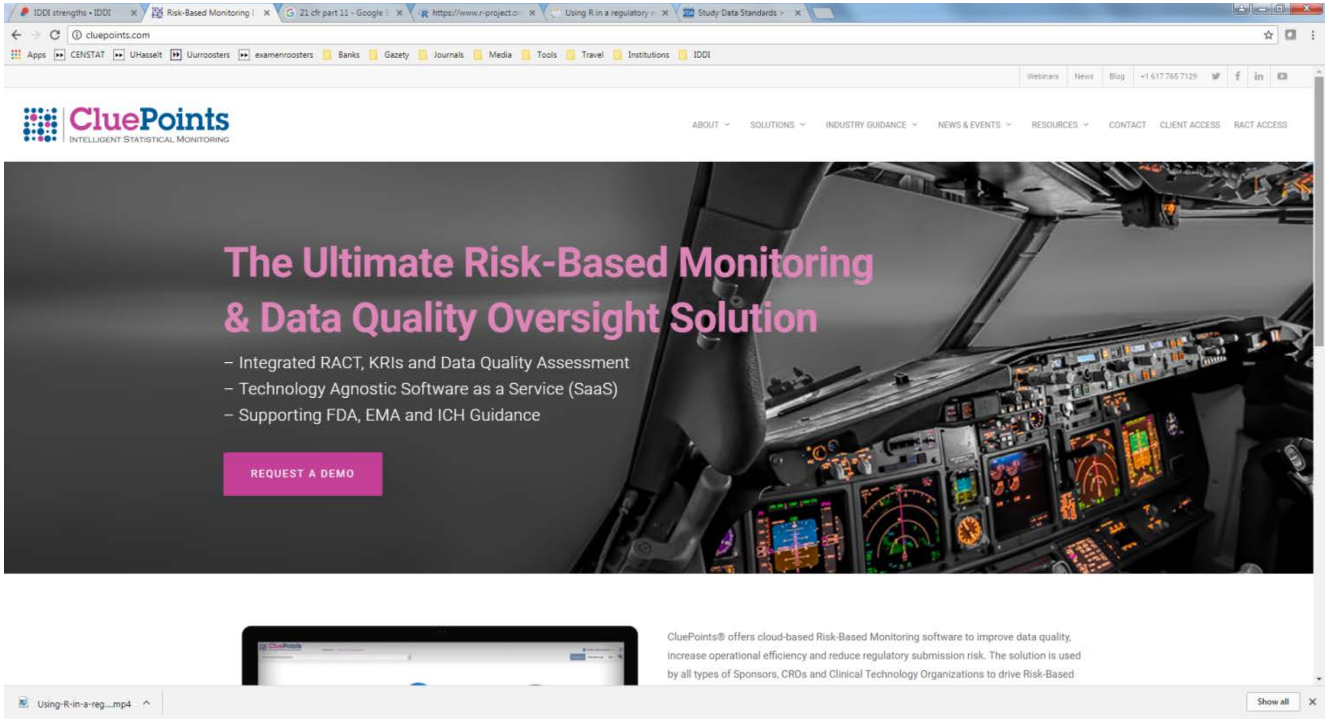
IDDI

- SAS for the majority of the (standard) analyses

- R used mainly for graphics
  - sometimes for more non-standard approaches such as multiple imputation of missing data

- Sample-size calculations
  - specialized software
  - R packages (gsDesign)

ClueTPoints
INTELLIGENT STATISTICAL MONITORING



The Ultimate Risk-Based Monitoring & Data Quality Oversight Solution

– Integrated RACT, KRIs and Data Quality Assessment
– Technology Agnostic Software as a Service (SaaS)
– Supporting FDA, EMA and ICH Guidance

CluePoints® offers cloud-based Risk-Based Monitoring software to improve data quality, increase operational efficiency and reduce regulatory submission risk. The solution is used by all types of Sponsors, CROs and Clinical Technology Organizations to drive Risk-Based

Background

Typical Clinical Trial Budget

> 100,000 trials worldwide

Typical registration trial

Duration: 2-3 years

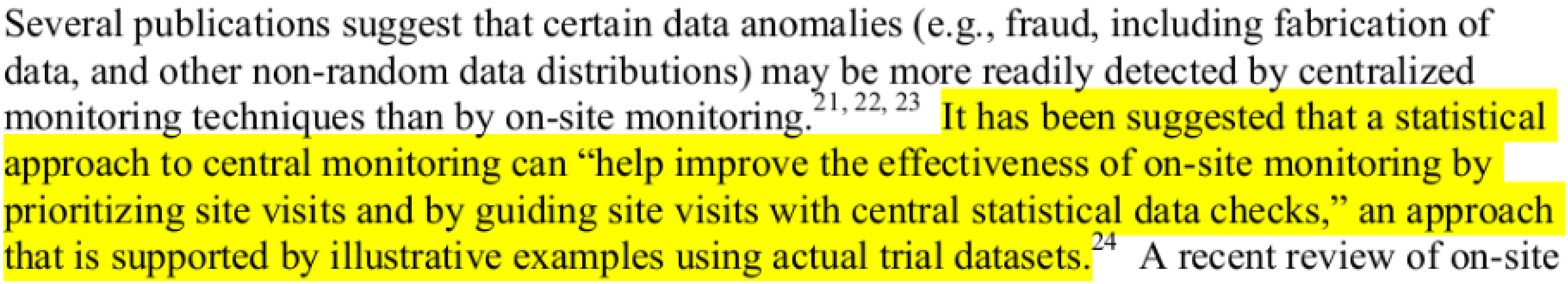
Trial budget: 100M$

Of which, on-site monitoring: 35M$

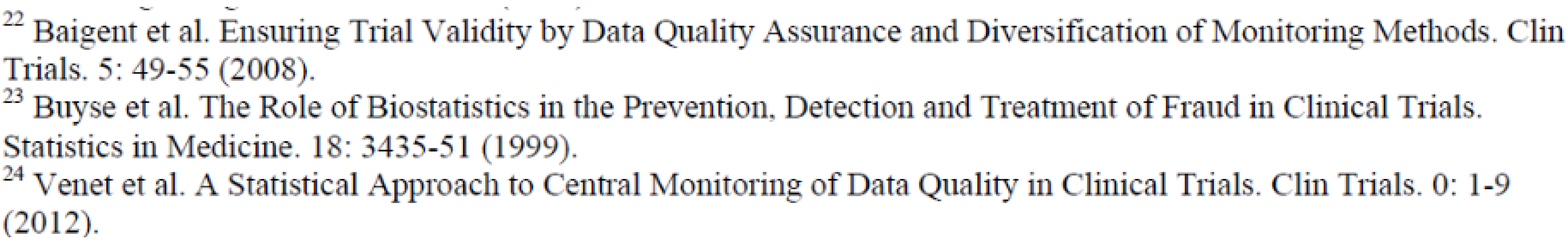
Of which, source data verification (SDV): 20M$

Background

FDA Guidance for industry



Several publications suggest that certain data anomalies (e.g., fraud, including fabrication of data, and other non-random data distributions) may be more readily detected by centralized monitoring techniques than by on-site monitoring.[21, 22, 23] It has been suggested that a statistical approach to central monitoring can "help improve the effectiveness of on-site monitoring by prioritizing site visits and by guiding site visits with central statistical data checks," an approach that is supported by illustrative examples using actual trial datasets.[24] A recent review of on-site



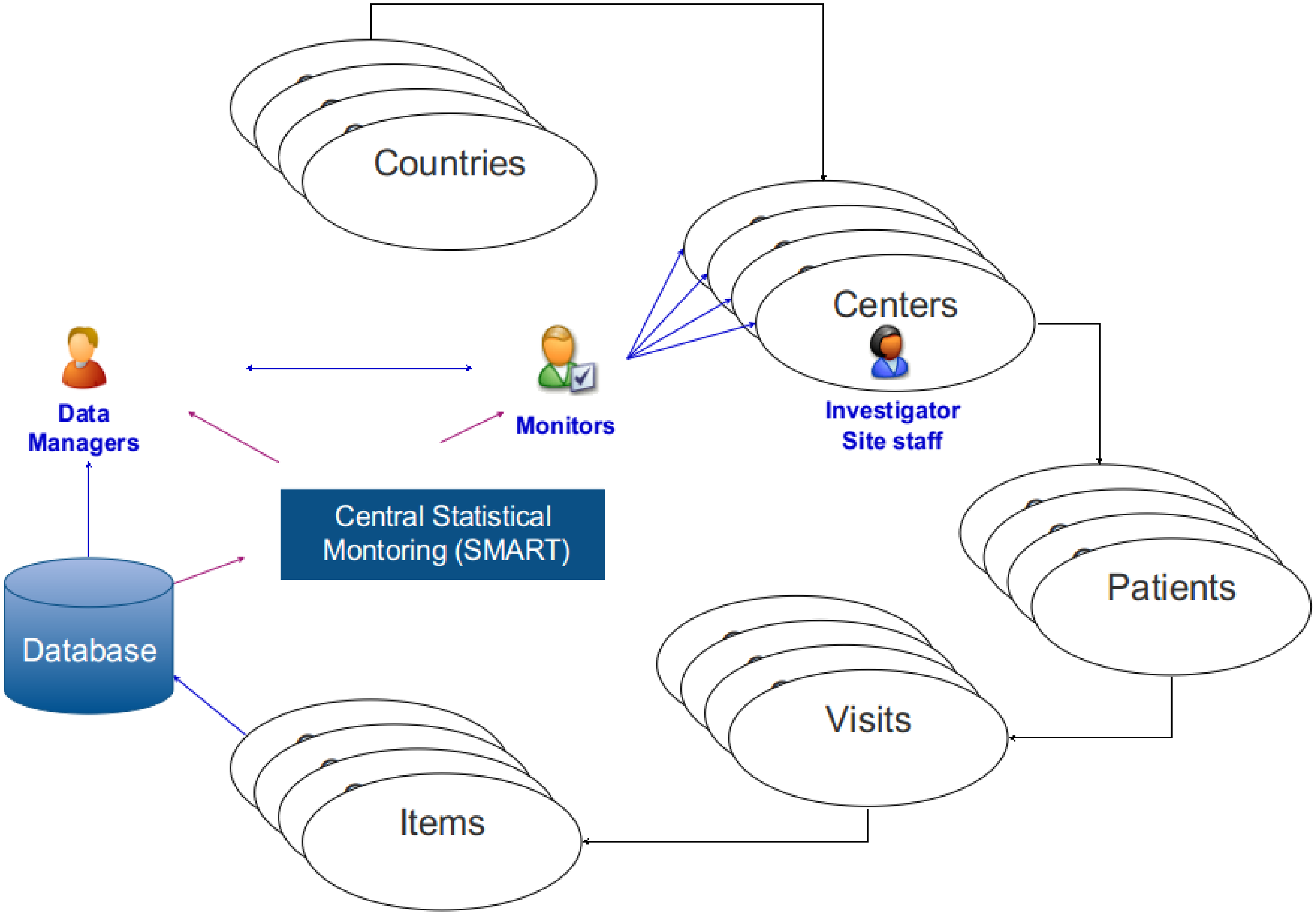
[22] Baigent et al. Ensuring Trial Validity by Data Quality Assurance and Diversification of Monitoring Methods. Clin Trials. 5: 49-55 (2008).

[23] Buyse et al. The Role of Biostatistics in the Prevention, Detection and Treatment of Fraud in Clinical Trials. Statistics in Medicine. 18: 3435-51 (1999).

[24] Venet et al. A Statistical Approach to Central Monitoring of Data Quality in Clinical Trials. Clin Trials. 0: 1-9 (2012).

Background

Central Statistical Monitoring



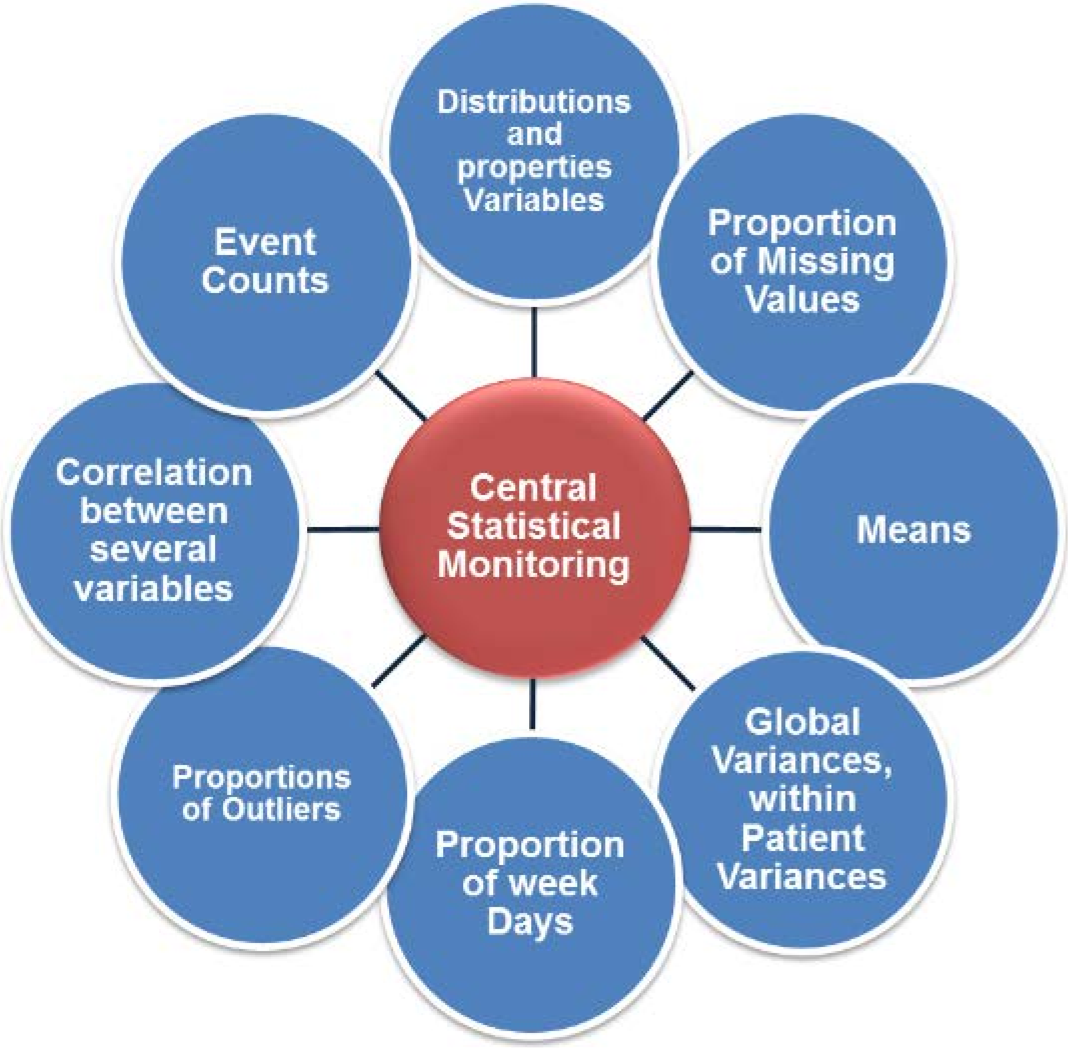
Introduction to CSM

Statistical principles

- Exploit the structure of data, which results from the use of the same protocol in all participating centers
- Every variable in the clinical database is potentially indicative of data quality
- Atypical trends/patterns in data can be detected by comparing each center against all other centers

Introduction to CSM

Statistical Tests


Distributions and properties Variables
Event Counts
Proportion of Missing Values
Correlation between several variables
Central Statistical Monitoring
Means
Proportions of Outliers
Proportion of week Days
Global Variances, within Patient Variances

R at CluePoints

SMART (Software for Monitoring and Analysis of Research Trials)

- Research version developed fully in R
  - Existing packages, as well self-written functions
  - “Automated” analyses

- Commercialization based on the open-source license?
- Protection of the code when sharing the software?
- Stability with updated versions of the R engine?

- Eventually
  - R for development
  - Production: web application based on C++ re-coding

Conclusions?

- In the (academic) research environment, knowledge of R is (probably) sufficient

- Outside it, you may have to deal with other software
  - teaching
  - industry standards
  - worth preparing (yourself/your students) for it ?

- If you develop a package, write a decent vignette !

WYDZIAŁ NAUK O ZDROWIU
UNIWERSYTETU MEDYCZNEGO W BIAŁYMSTOKU

UNIKATOWE STUDIA PODYPLOMOWE

Biostatystyka – zastosowania statystyki w medycynie klinicznej, biologii i naukach o zdrowiu

1 rok akademicki – 2 semestry
154 godziny dydaktyczne
Koszt – 3900 zł za rok
tryb niestacjonarny, b-learning, zajęcia w j. angielskim

Studia powstały i realizowane są z myślą o osobach chcących nabyć lub podwyższyć swoje kwalifikacje w obszarach wysoce pożądanych na rynku pracy. Biostatystyka jest dziedziną o szerokim zastosowaniu w badaniach biomedycznych, a anglojęzyczne studia podyplomowe w zakresie biostatystyki oferowane przez UMB są ofertą unikatową na polskim rynku. Uczestnicy studiów mogą podnieść swoje kwalifikacje jako biostatystyk, lub też uzyskać nowe kwalifikacje.

Studia realizowane są w języku angielskim przy użyciu technik „blended learning”. Materiały wykładowe w formie elektronicznej, sesje konsultacyjne i seminaria prowadzone przy użyciu komunikatorów internetowych; konieczność przyjazdu do Białegostoku tylko na egzaminy.

Kandydaci: absolwenci studiów przynajmniej pierwszego stopnia w matematyce, statystyce, informatyce, bioinformatyce, bioinżynierii, biotechnologii; ewentualnie (w przypadku ukończenia innych studiów minimum I stopnia), udokumentowane doświadczenie zawodowe w zakresie zastosowań statystyki; udokumentowana znajomość języka angielskiego na poziomie minimum B2

Szczegółowe informacje
http://www.umb.edu.pl/wnoz