ARISTOTLE UNIVERSITY
OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE
MSc Health Statistics and Data Analytics

Analysis sets & handling of missing data

Anna-Bettina Haidich
Associate Professor of Medical Statistics –Epidemiology
haidich@auth.gr

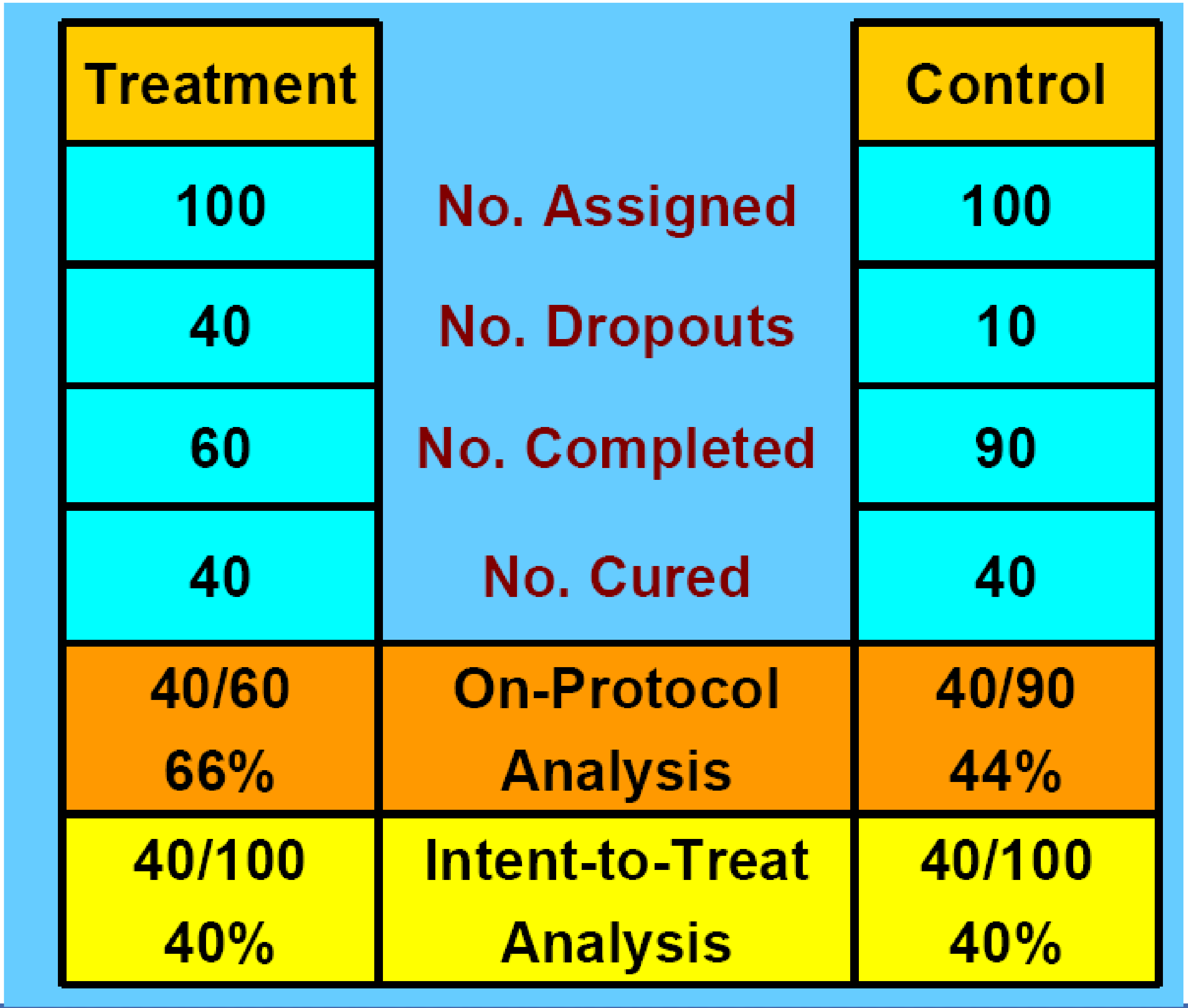
THESSALONIKI 2021-22

Planning Statistical Analysis

- **Analysis Sets**
  - Intent-to-treat (ITT)
  - Full analysis set (FAS)
  - Per-Protocol (PP)

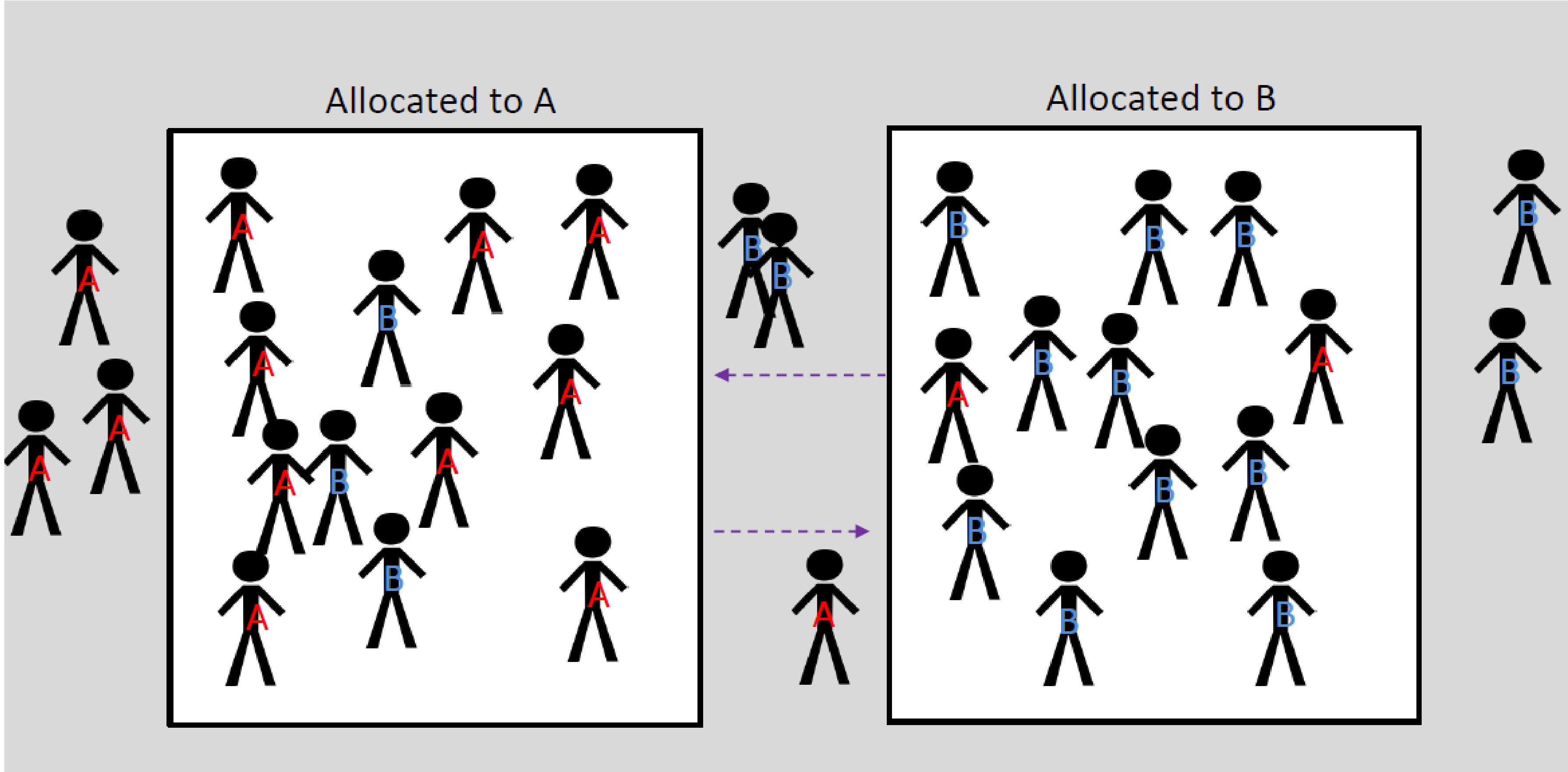
- **Pre-specification of analysis**
  - Primary analysis
  - Secondary analysis
  - Handling of missing data
  - Sensitivity analysis
  - Subgroup analysis
  - Post-hoc and exploratory analysis

Per Protocol or Intention to Treat Analysis?



Treatment		Control
100	No. Assigned	100
40	No. Dropouts	10
60	No. Completed	90
40	No. Cured	40
40/60 66%	On-Protocol Analysis	40/90 44%
40/100 40%	Intent-to-Treat Analysis	40/100 40%

The Analysis Set



Analysis Set

- Non-compliance: Participant allocated to receive the active treatment does not receive the treatment as defined in the protocol
  - Therapy trial – participant should complete a minimum number of sessions to count as compliant
  - Drug trial – participant does not take the drug as prescribed e.g. missed doses
  - Participant allocated to the intervention withdraws partway through

Analysis Set

- Contamination: A participant in the control arm receives some aspects of the active treatment
  - Therapy trial – participants in the active arm share some components of the active treatment
  - Trial of an over-the-counter medicine/supplement – participants purchase the active treatment outside the trial
  - Vaccine trial – person administering the vaccine used the wrong vial

- False inclusions: Participants who were randomized but later found to not be eligible for the trial based on the inclusion and exclusion criteria

Analysis Set

- Intention-to-treat
  - analysis as randomised
  - ideally no missing data, loss to follow-up
  - usually more conservative in superiority studies

- Full analysis set
  - analysis as randomised
  - include available data only

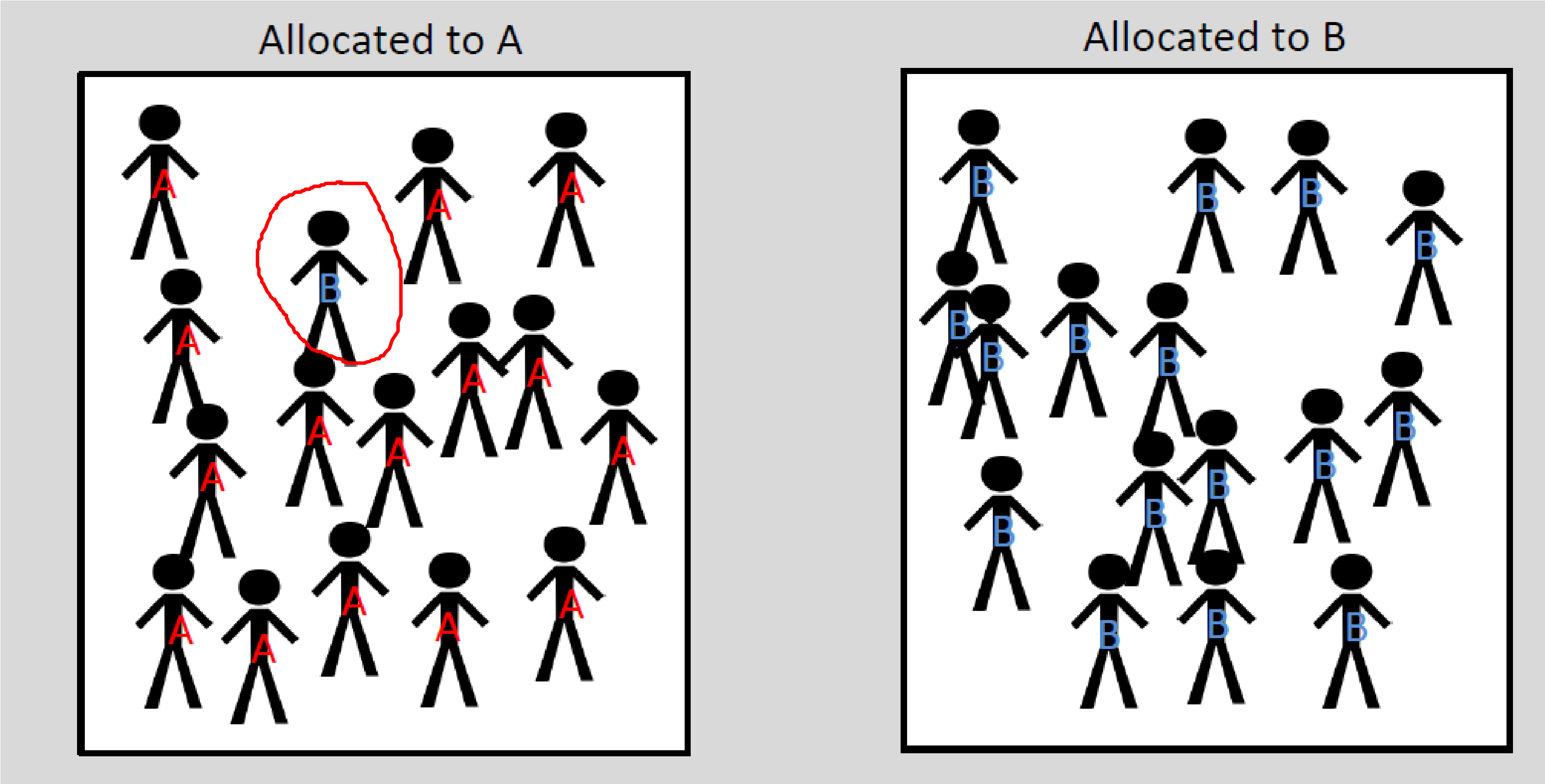
- Per protocol
  - analyse according to treatment received and includes only patients who satisfy entry criteria and properly follow the protocol
  - might be expected to enhance any treatment effect due to removal of “noise”, but can also work the other way

Full Analysis Dataset

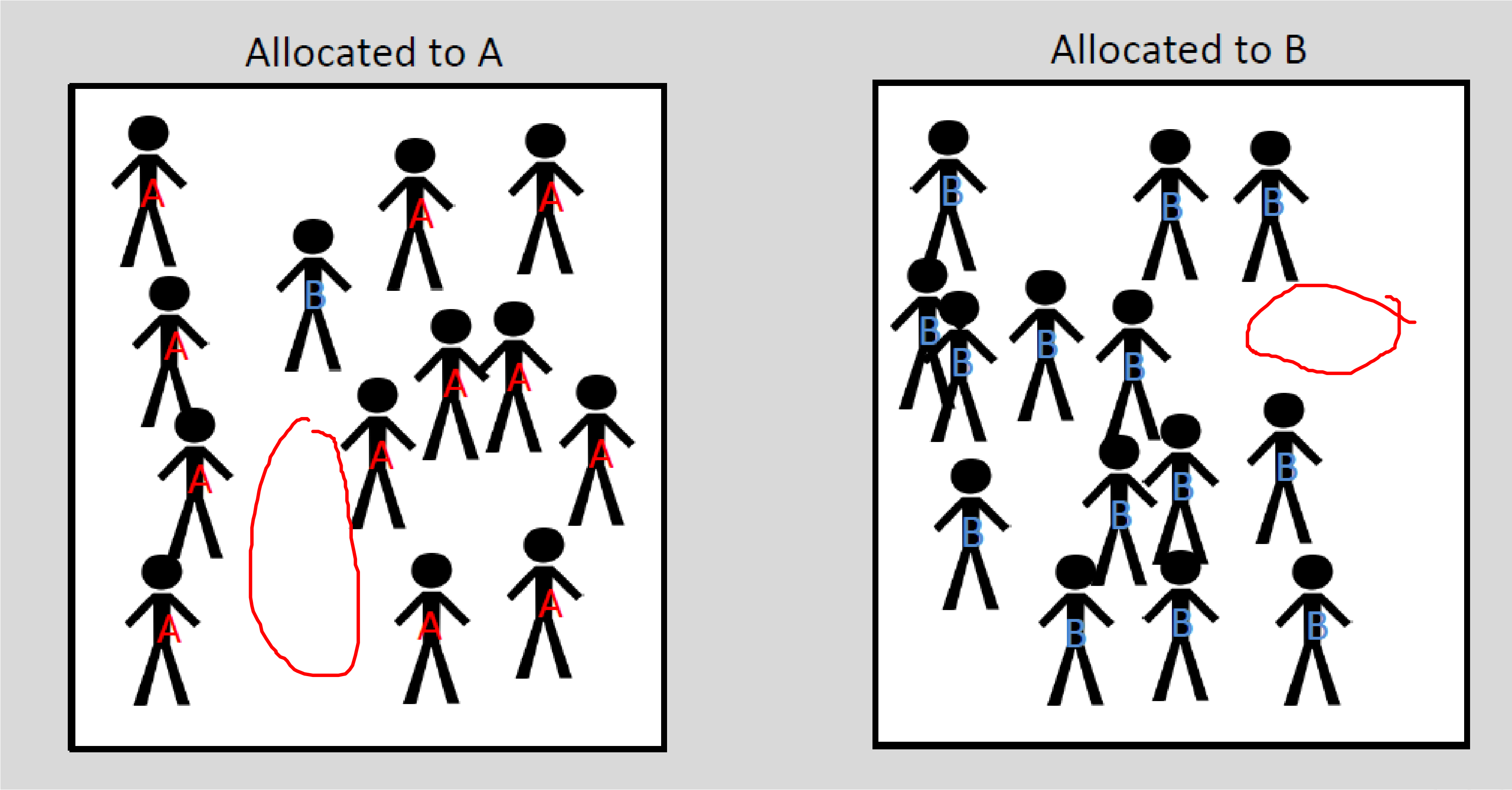
- Ideally all outcomes should be complete for ITT analysis

- In practice
  - participants may be lost to follow-up or have missing data
  - participants randomised in error (false inclusions)
  - some participants never start treatment

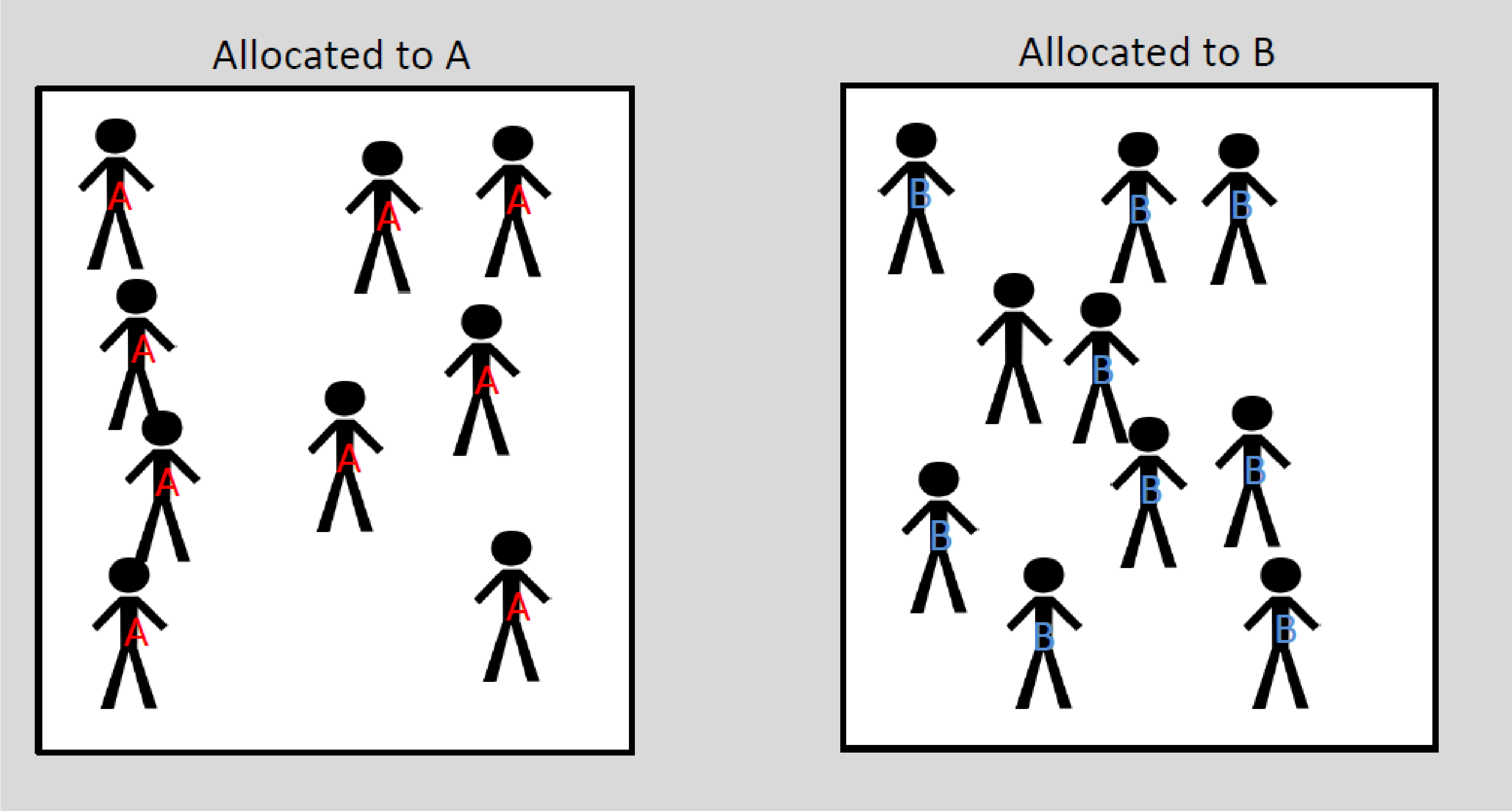
Intention-to-treat



Full analysis set



Per-Protocol

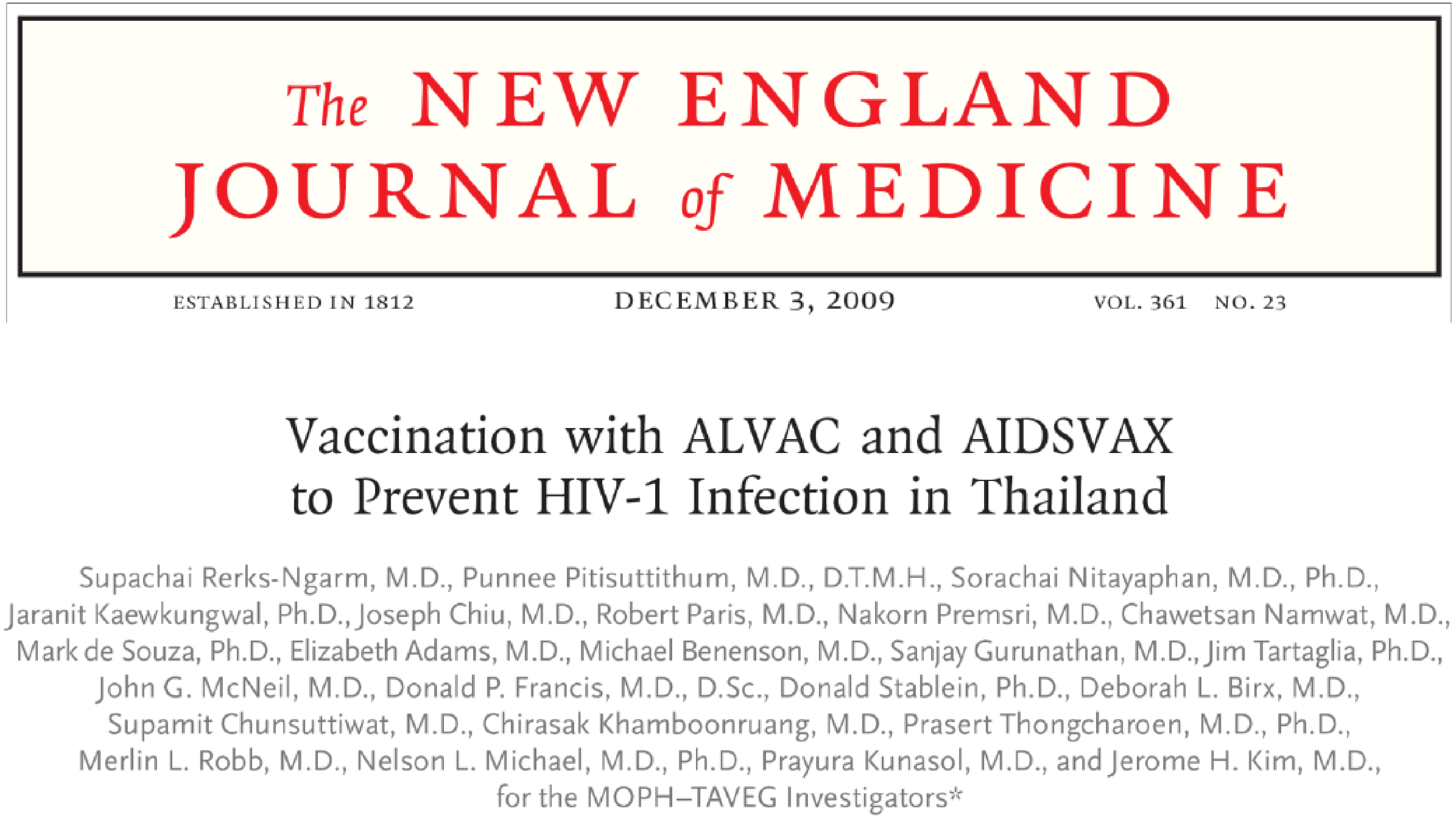


Other analysis set

- Modified intention-to-treat
  - Exclusion of some randomised subjects in a justified way (e.g. ineligible after randomisation or never start treatment)

- Treatment received
  - Analyse as treatment actually received
  - Usually to assess treatment safety

Example



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 3, 2009 VOL. 361 NO. 23

Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand

Supachai Rerks-Ngarm, M.D., Punnee Pitisuttithum, M.D., D.T.M.H., Sorachai Nitayaphan, M.D., Ph.D., Jaranit Kaewkungwal, Ph.D., Joseph Chiu, M.D., Robert Paris, M.D., Nakorn Premsri, M.D., Chawetsan Namwat, M.D., Mark de Souza, Ph.D., Elizabeth Adams, M.D., Michael Benenson, M.D., Sanjay Gurunathan, M.D., Jim Tartaglia, Ph.D., John G. McNeil, M.D., Donald P. Francis, M.D., D.Sc., Donald Stablein, Ph.D., Deborah L. Birx, M.D., Supamit Chunsuttiwat, M.D., Chirasak Khamboonruang, M.D., Prasert Thongcharoen, M.D., Ph.D., Merlin L. Robb, M.D., Nelson L. Michael, M.D., Ph.D., Prayura Kunasol, M.D., and Jerome H. Kim, M.D., for the MOPH–TAVEG Investigators*

Population for analysis

- Vaccine efficacy estimates (% reduction in HIV infection):

- Intention-to-treat
  - N=16402, 26.4% (95% CI: -4.0 to 4.79; P=0.08)

- Per-protocol
  - N=12452, 26.2% (95% CI: -13.3 to 51.9; P=0.16)

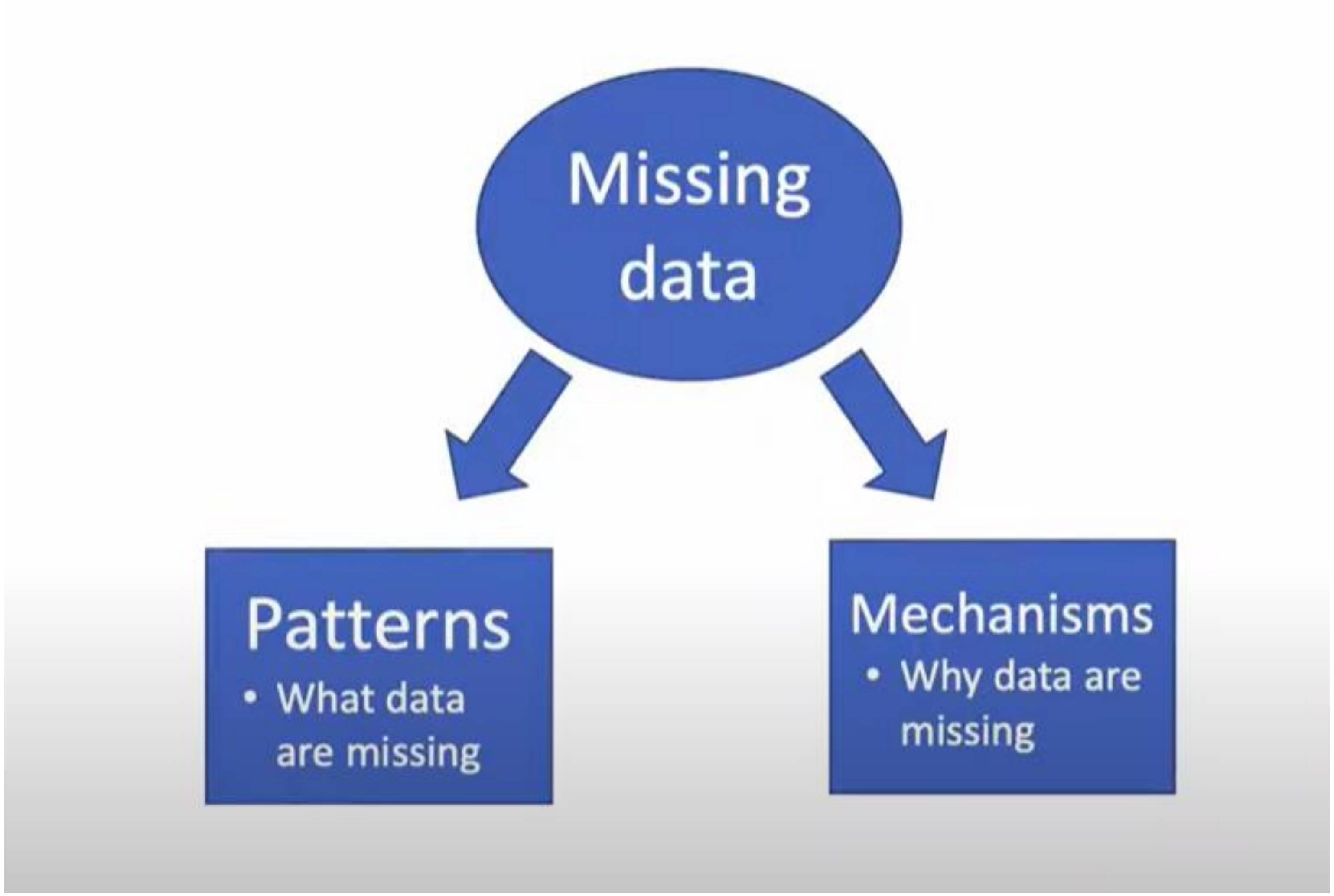
- Modified intention-to-treat
  - N=16395 (excluding 7 subjects who were found to have had HIV infection at baseline), 31.2% (95% CI: 1.1 to 52.1; P=0.04)

Missing data: general considerations

- Effect on sample size
- Effect on variability (completers might be more likely to have similar values) → artificial narrowing of the confidence interval

Handling of missing data

- Trial design
  - nature of the outcome variable (e.g. mortality vs sophisticated methods of diagnosis)
  - follow-up time
  - treatment modalities (e.g. medical vs surgical)
  - target population (e.g. psychiatric disorders)
- Trial management
  - number of visits and assessments
  - ease of data collection (e.g. electronic data capture)
  - continuing data collection after withdrawal
- Statistical method



Missing data

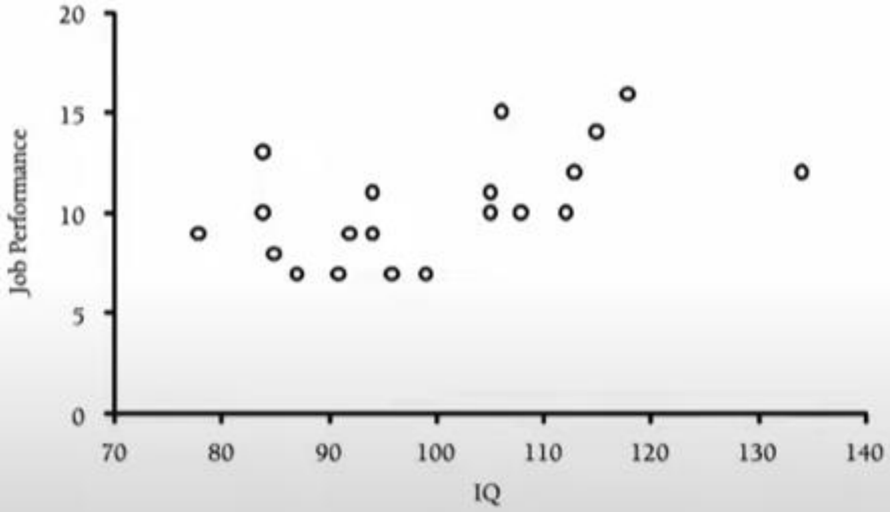
Patterns
- What data are missing

Mechanisms
- Why data are missing

Mechanisms of missing data

- Missing completely at random (MCAR)
  - Missingness of a variable is completely independent of itself and other variables
    - e.g. patients moving to another city for non-health reasons
- Missing at random (MAR)
  - Missingness of a variable is dependent on another variable, post drop-out observations can be predicted from the observed variables
  - e.g. patient drop-outs due to lack of efficacy
- Not missing at random (NMAR)
  - Missingness of a variable is related to itself, the unobserved responses depend on information not available for the analysis
    - e.g. after a series of visits with good outcomes a patient drops out due to lack of efficacy

Example dataset



1. Complete-data scatterplot of the IQ and job performance scores from Table 2.1.

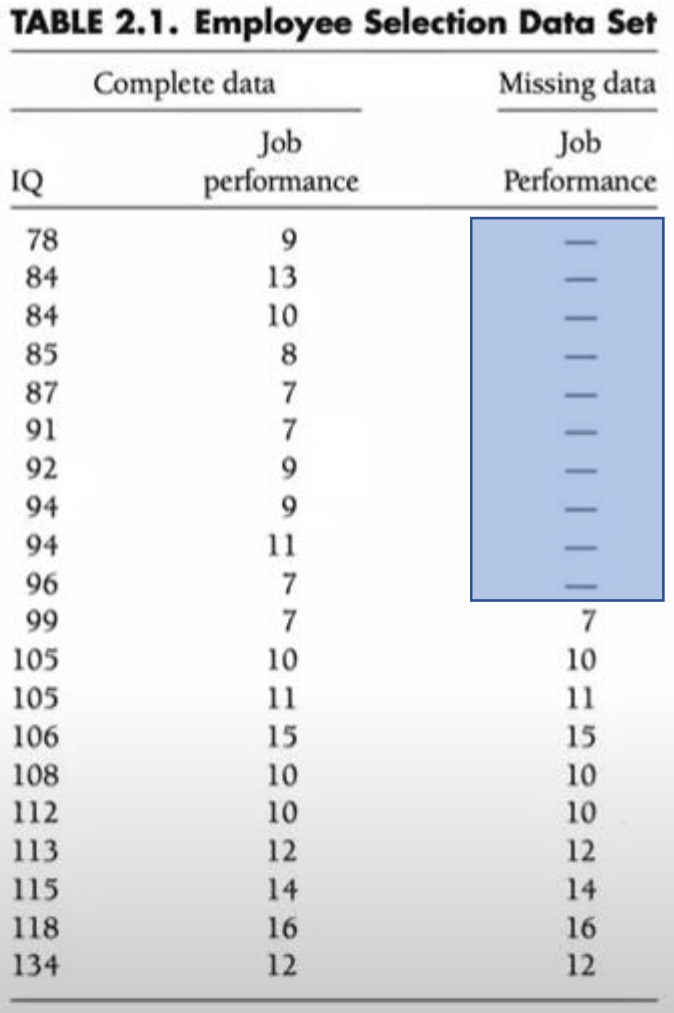


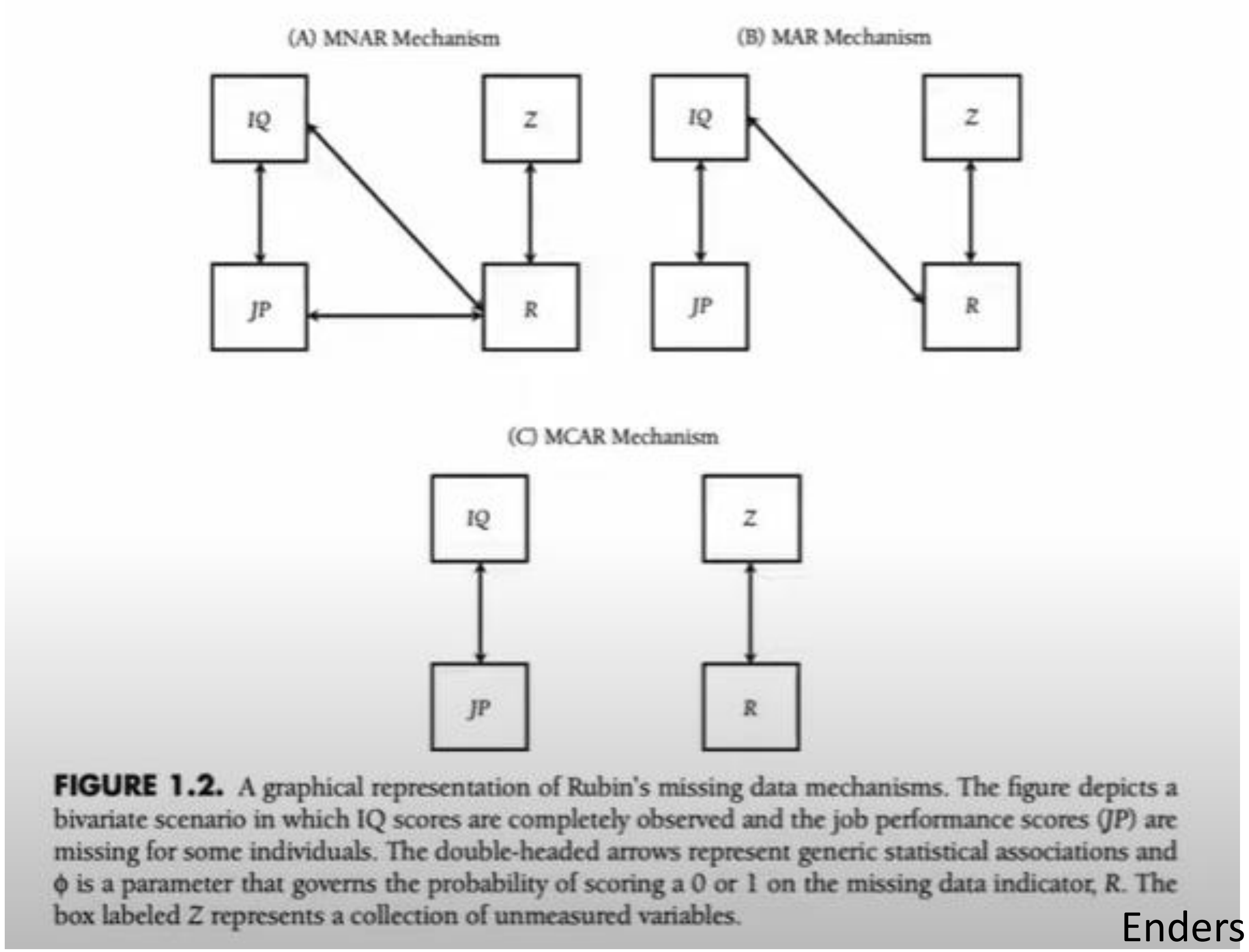
TABLE 2.1. Employee Selection Data Set

	Complete data	Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Is this MCAR, MAR or MNAR?

Enders C. (2010) Applied missing data analysis. Guilford publications

Missing data mechanism



(A) MNAR Mechanism

(B) MAR Mechanism

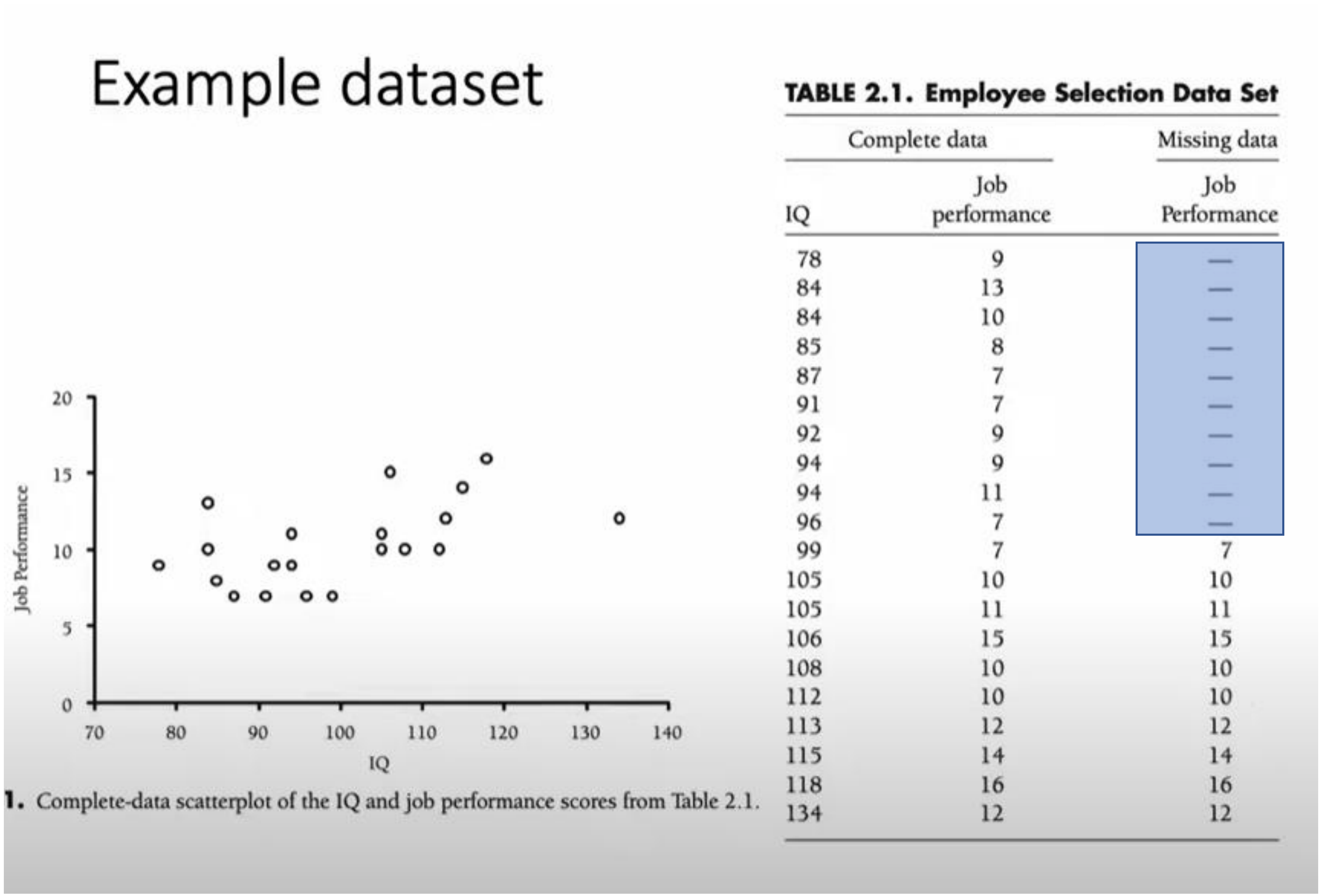
(C) MCAR Mechanism

FIGURE 1.2. A graphical representation of Rubin's missing data mechanisms. The figure depicts a bivariate scenario in which IQ scores are completely observed and the job performance scores (*JP*) are missing for some individuals. The double-headed arrows represent generic statistical associations and ϕ is a parameter that governs the probability of scoring a 0 or 1 on the missing data indicator, *R*. The box labeled *Z* represents a collection of unmeasured variables.

- Z = collection of unmeasured variables
- R= missing data indicator (0 = no, 1=yes)

Enders C. (2010) Applied missing data analysis. Guilford publications

Example dataset



1. Complete-data scatterplot of the IQ and job performance scores from Table 2.1.

TABLE 2.1. Employee Selection Data Set

Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Is this MCAR, MAR or MNAR?

Μάλλον είναι MAR

Enders C. (2010) Applied missing data analysis. Guilford publications
https://www.youtube.com/watch?v=TxcLeMsZ7Bk

Statistical methods

- Available case analysis (Full Analysis Set, FAS)
  - might be expected to enhance treatment effect by the removal of “noise” and can be used when data are MCAR
- Baseline observation carried forward (BOCF)
- Last observation carried forward (LOCF)
- Multiple imputation
- Mixed effects model
- General guidance
  - choice of method unlikely to be biased in favour of the experimental treatment
  - pre-specify method to be used for primary analysis
  - explore pattern of missing data and conduct sensitivity analysis

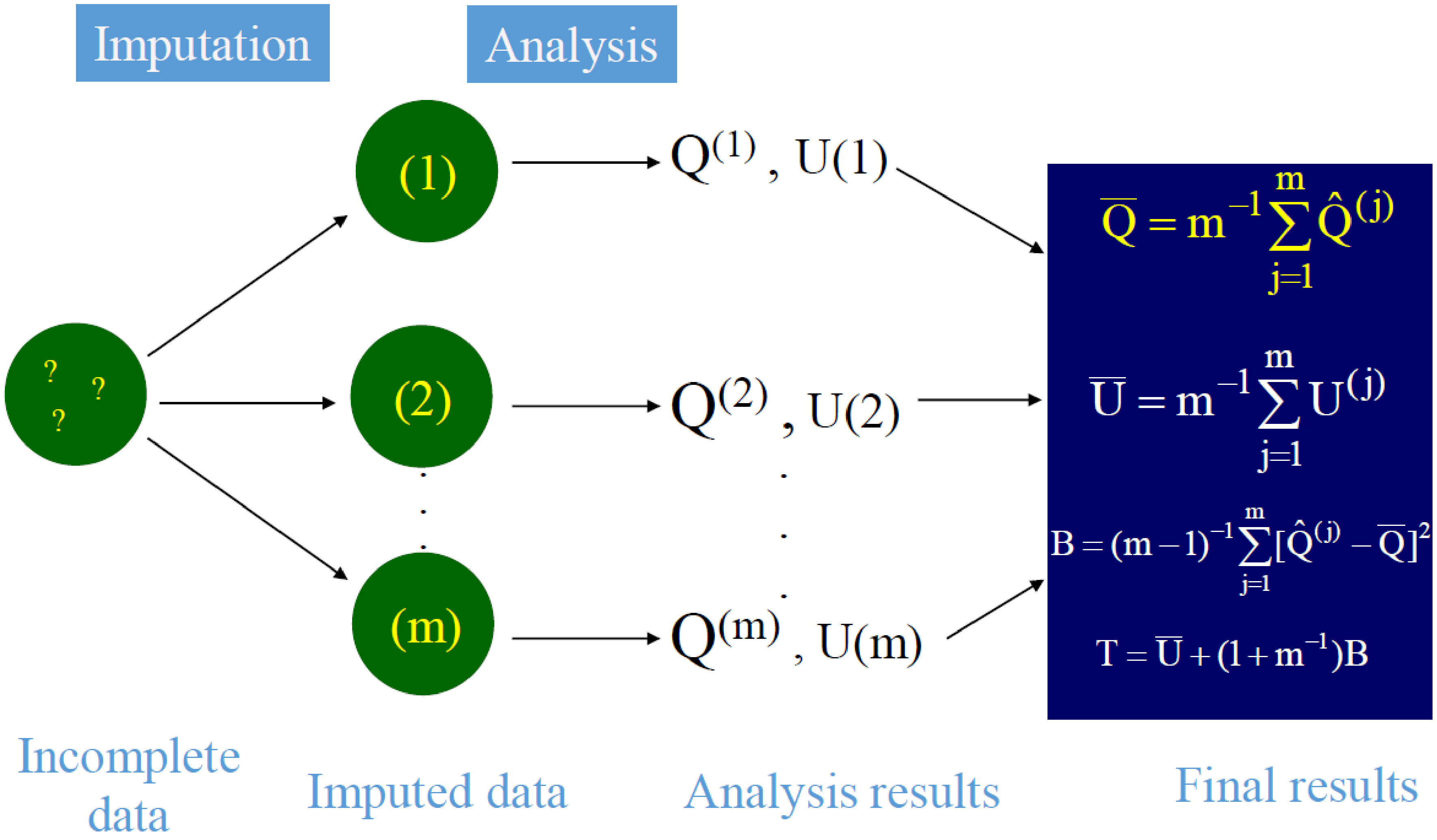
Single imputation methods

- Patient’s condition expected to deteriorate over time (e.g. Alzheimer’s disease)
  - LOCF likely to give overly optimistic results for both treatment groups
  - If withdrawals in the active treatment are earlier the treatment effect will be biased in favour of the test product
- Patient’s condition expected to improve spontaneously over time (e.g. depression)
  - LOCF conservative in case patients in the experimental group tend to withdraw earlier
- Chronic pain trial
  - BOCF may be appropriate (pain returns to its baseline level if patient withdraws from treatment)

Multiple imputation

- Generate $m > 1$ imputed datasets by filling in the missing values with plausible values
- Perform standard analysis on each of the m imputed datasets
- Combine results from the m analysis using Rubin rules

Multiple imputation



Imputation

Analysis

? ? ?

(1)

(2)

(m)

$Q^{(1)}$, U(1)

$Q^{(2)}$, U(2)

$Q^{(m)}$, U(m)

$$\bar{Q} = m^{-1}\sum_{j=1}^{m}\hat{Q}^{(j)}$$

$$\bar{U} = m^{-1}\sum_{j=1}^{m}U^{(j)}$$

$$B = (m-1)^{-1}\sum_{j=1}^{m}[\hat{Q}^{(j)} - \bar{Q}]^2$$

$$T = \bar{U} + (1+m^{-1})B$$

Incomplete data

Imputed data

Analysis results

Final results

Multivariable models

$$\text{outcome} \longrightarrow \underset{\text{Dependent variable}}{\boldsymbol{y}} = \overbrace{\underbrace{\boldsymbol{b_o + b_1 x_1 + \ldots. + b_p x_p + \varepsilon}}_{\text{Independent variables}}}^{\text{Exposure, risk factors, confounders and effect modifiers}}$$

Multivariable models

- Dependent variable
  - continuous → linear regression
  - binary → logistic regression
  - Time to event → Cox proportional hazards regression
  - Counts → Poisson regression
- Independent variable
  - continuous/categorical

Choosing a “Final Model”

- When faced with potentially many possible predictors, how does a researcher go about choosing a “best” model?

- Model building and selection is a combination of science, statistics, and the research goal(s)

Choosing a “Final Model”

- If goal is to maximize precision of adjusted estimates
  - Keep only those predictors that are statistically significant in final model
- If goal is to present results comparable to results of similar analyses presented by other researchers (on similar or different populations)
  - Present at least one model that includes the same predictor set as the other research

Choosing a "Final Model"

- If goal is Prediction
  - Here the primary issue is minimizing prediction error rather than causal interpretation of the predictors in the model

- If goal is to evaluate a predictor of primary interest
  - then eliminating variables based solely on statistical significance is not the best approach.

Different selection methods

- **“10% change in estimate” variable selection rule**
  - A potential confounder is included in the model if it changes the coefficient, or effect estimate, of the primary exposure variable by 10%.
  - More reliable models than variable selection methods based on statistical significance (Greenland, 1989).

- **Using Both 10% Rule and P Values**
  - “In the multiple regression models, confounders were included if they were significant at a 0.05 level or they altered the coefficient of the main variable by more than 10 percent in cases in which the main association was significant.”

Kulkarni et al (*N Engl J Med*, 2006)

Different selection methods

- **More Cautious Approach to Guard Against Confounding (10% Rule + conservative P value + a priori confounders)**
  - “We analyzed raw test scores adjusted for a priori confounders, including linear terms for age, family income, and score on the HOME scale[14,15] and dummy-coded variables for sex, HMO, maternal IQ, maternal education, single-parent status, and birth weight. Other covariates were included in the full model if the **P value was less than 0.20 or if their inclusion resulted in a change of 10%** or more in the estimate of the main effect of mercury exposure[19,20]…”

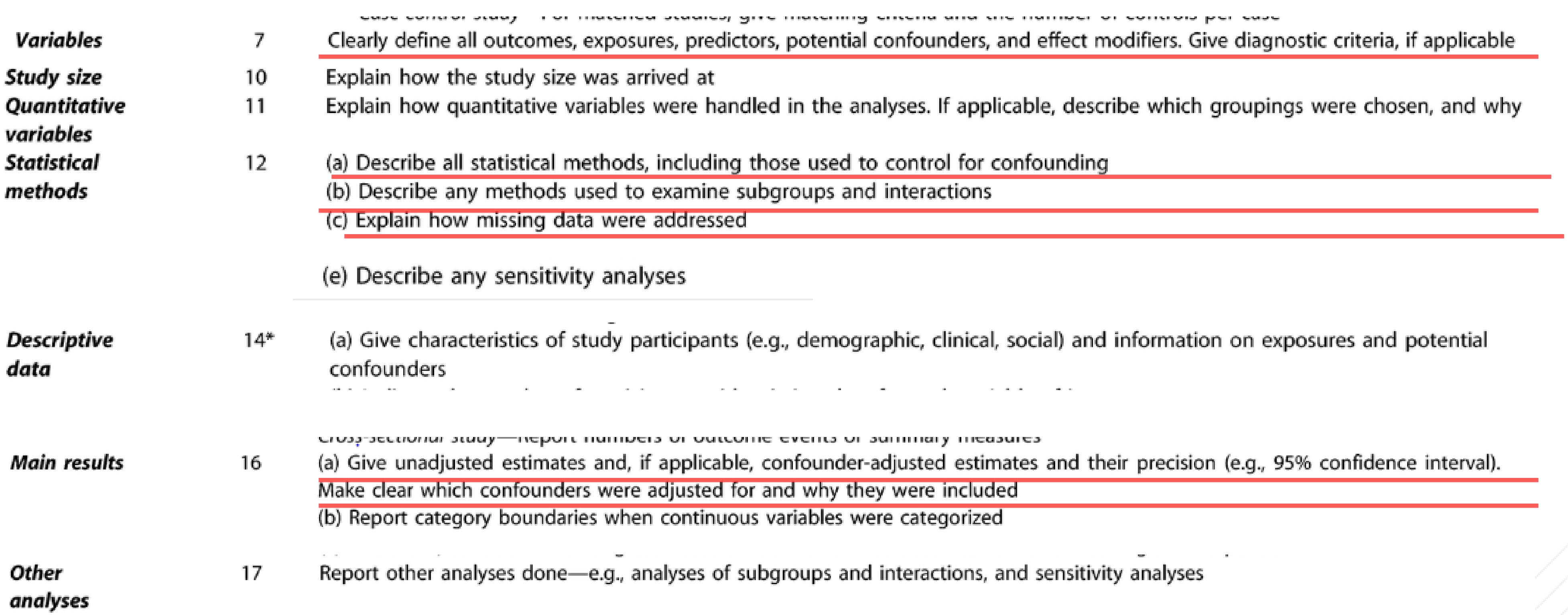
Thompson et al (*N Engl J Med*, 2007)

Different selection methods

- **Automated Variable Selection Procedures**
  - Backwards selection is considered superior to forwards selection
  - Although finding a significant set of predictors, have no way to make decisions about collinearity or confounding, and they can even produce nonsensical models
  - An interactive backwards elimination" is better, where the researcher, makes the decision at each step.

STROBE Statement

Strengthening the reporting of observational studies in epidemiology



		Case-control study—For matched studies, give matching criteria and the number of controls per case
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen, and why
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (e) Describe any sensitivity analyses
Descriptive data	14*	(a) Give characteristics of study participants (e.g., demographic, clinical, social) and information on exposures and potential confounders
		Cross-sectional study—Report numbers of outcome events or summary measures
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (e.g., 95% confidence interval). Make clear which confounders were adjusted for and why they were included (b) Report category boundaries when continuous variables were categorized
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses

https://www.strobe-statement.org/index.php?id=strobe-home

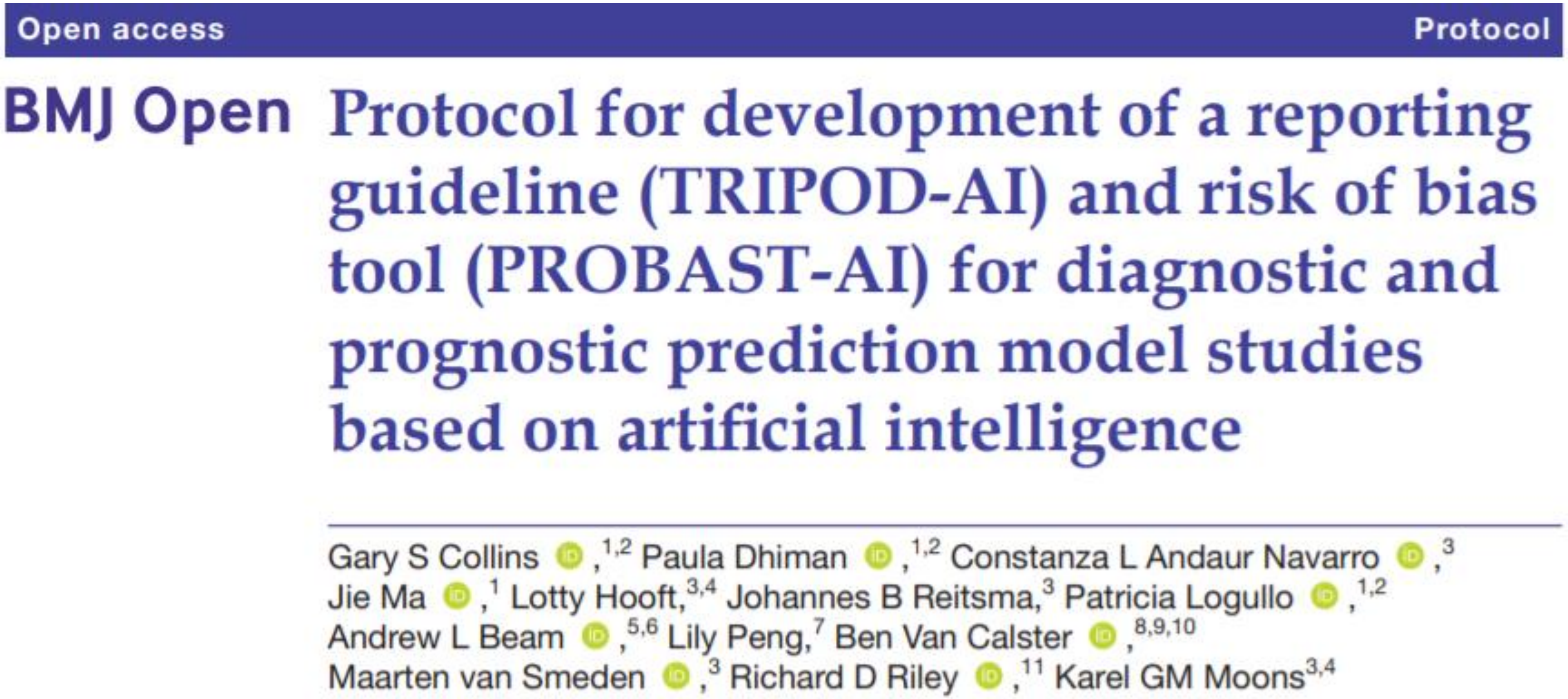
TRIPOD

- The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis)

- 22-items checklist

- https://www.tripod-statement.org

TRIPOD -AI



Open access Protocol

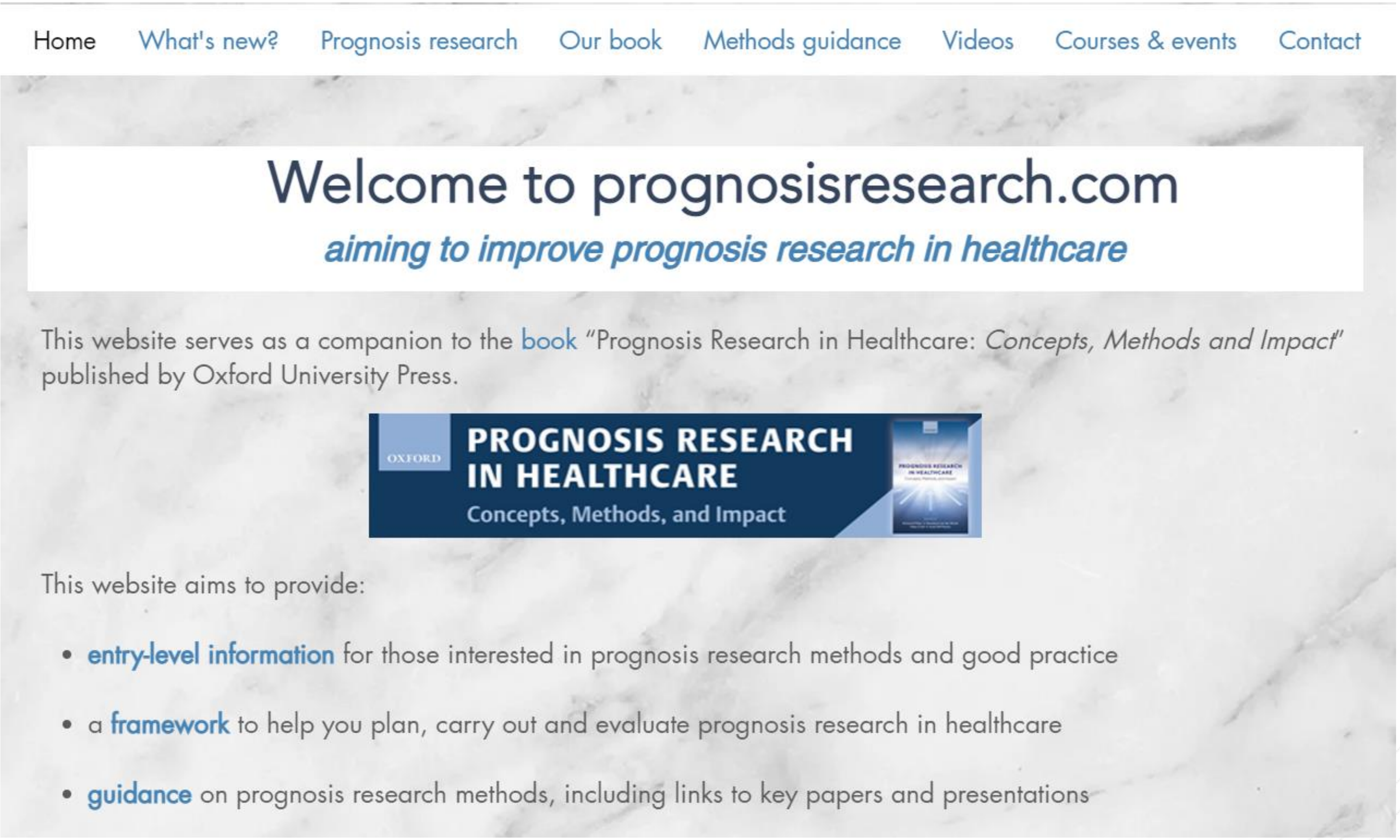
BMJ Open

Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence

Gary S Collins,[1,2] Paula Dhiman,[1,2] Constanza L Andaur Navarro,[3] Jie Ma,[1] Lotty Hooft,[3,4] Johannes B Reitsma,[3] Patricia Logullo,[1,2] Andrew L Beam,[5,6] Lily Peng,[7] Ben Van Calster,[8,9,10] Maarten van Smeden,[3] Richard D Riley,[11] Karel GM Moons[3,4]

https://bmjopen.bmj.com/content/11/7/e048008

https://www.prognosisresearch.com/



Welcome to prognosisresearch.com

aiming to improve prognosis research in healthcare

This website serves as a companion to the book "Prognosis Research in Healthcare: *Concepts, Methods and Impact*" published by Oxford University Press.

This website aims to provide:

- entry-level information for those interested in prognosis research methods and good practice
- a framework to help you plan, carry out and evaluate prognosis research in healthcare
- guidance on prognosis research methods, including links to key papers and presentations

All models are wrong but some are useful

George E.P. Box

Kind reminder!

Work on the activities till 25/02/2022

qa.auth.gr