"Why most responder analyses are misleading"

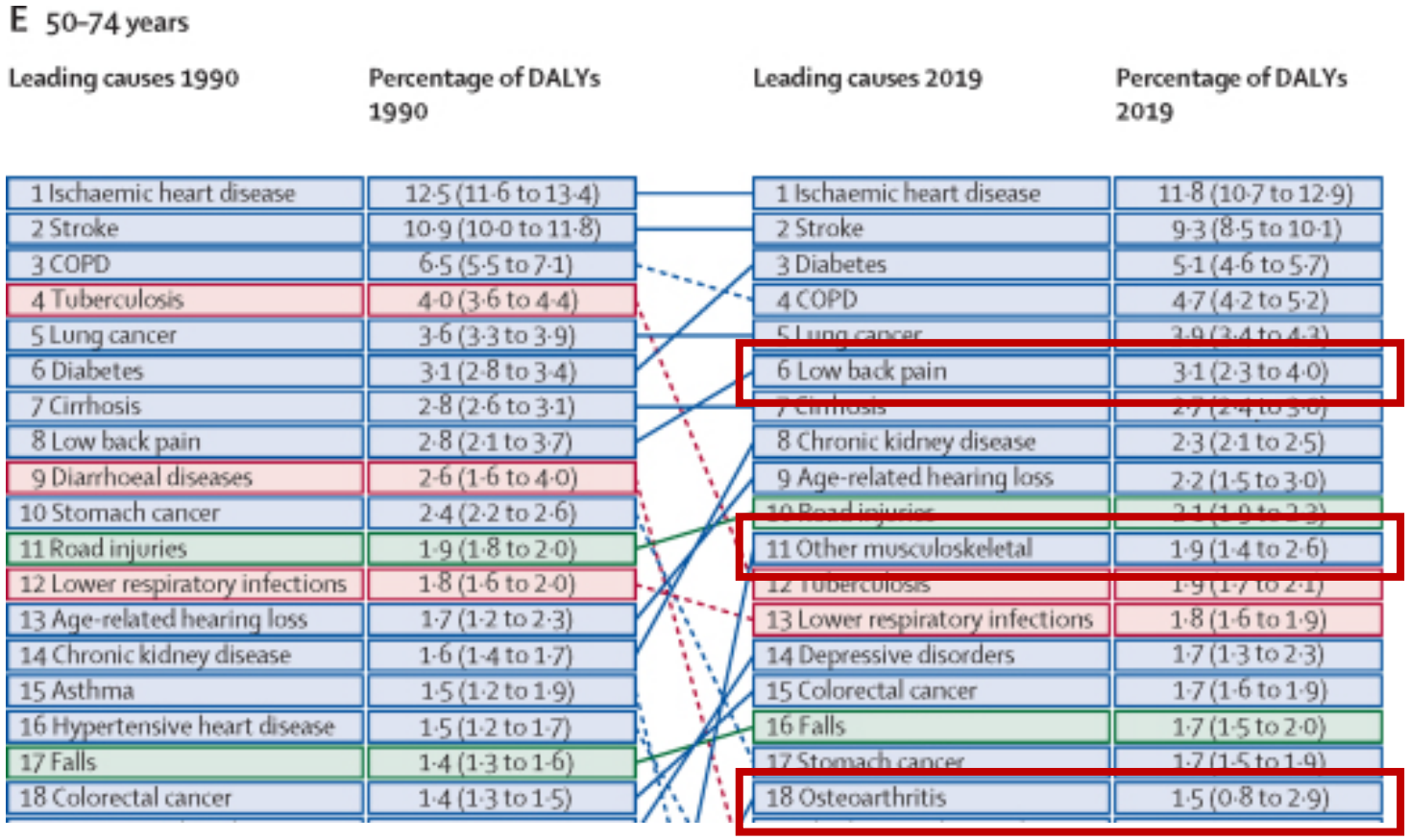
Aleksandra Turkiewicz, LUPOP methods seminars

Work done with Jos Runhaar, Marius Henrikssen and Martin Englund

https://portal.research.lu.se/en/persons/aleksandra-turkiewicz

Clinical background

- Chronic musculoskeletal (MSK) pain is a serious public health problem



PMID: 33069326

There are no disease modifying treatments

- Despite thousands of trials
- Exercise and NSAIDs are core treatment options
- The field is really trying to find something that would work

Pain is the cardinal symptom

- The most common ways of measuring pain in MSK diseases:
  - Visual Analog Scale (VAS, 0 to 100 mm) – how much pain do you have?
  - Numerical Rating Scale (NRS, 0 to 10) – how much pain do you have?
  - Knee (or hip) Osteoarthritis Outcome Score (KOOS) – questions about pain, function, sport participation, activities of daily living, quality of life (total score from 0 to 100)
  - WOMAC (similar to KOOS)
- How ‘stable’ are these measures over time?

Intraclass correlation for biomarkers and pain

Supplementary Table. Intraclass correlation coefficients (ICC) and 95% confidence intervals (CI) of biochemical marker measurements of tibiofemoral osteoarthritis participants from APPROACH[6] across month six, 12, and 24 follow-up visits. Urinary (u) markers have been corrected for creatinine levels.

Biomarker	ICC (95% CI)
ARGS-aggrecan	0.93 (0.91, 0.94)
C10C	0.90 (0.88, 0.92)
C1M	0.94 (0.92, 0.95)
C2M	0.92 (0.90, 0.93)
C3M	0.89 (0.87, 0.91)
Coll2-1	0.77 (0.72, 0.81)
u-creatinine	0.77 (0.72, 0.81)
uCTX-II	0.44 (0.31, 0.54)
VICM	0.82 (0.78, 0.85)

Supplementary table. Intraclass correlation coefficients (95% confidence intervals) for two measurements of pain, WOMAC physical function and KOOS quality of life (QOL), one year apart, in persons with osteoarthritis.

Patient reported outcome	ICC (95% CI), all
NRS knee pain	0.64 (0.60, 0.68)
WOMAC physical function	0.69 (0.66, 0.73)
KOOS QOL	0.72 (0.69, 0.75)

Patient reported outcome	ICC (95% CI), eligible for trial
NRS knee pain	0.54 (0.47, 0.60)
WOMAC physical function	0.63 (0.57, 0.68)
KOOS QOL	0.69 (0.64, 0.73)

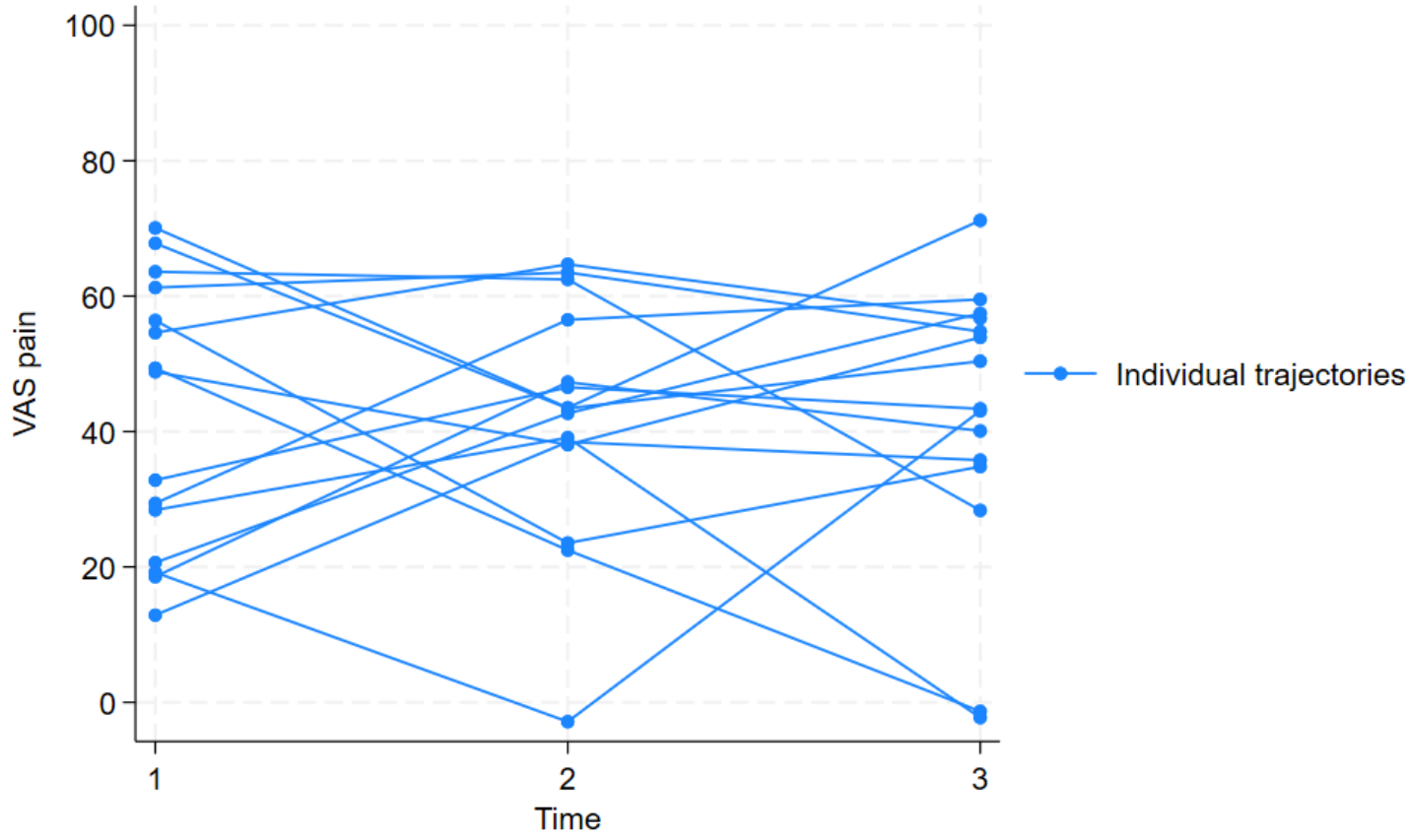
DOI: 10.1016/j.joca.2024.11.002

https://doi.org/10.1016/j.joca.2024.11.006

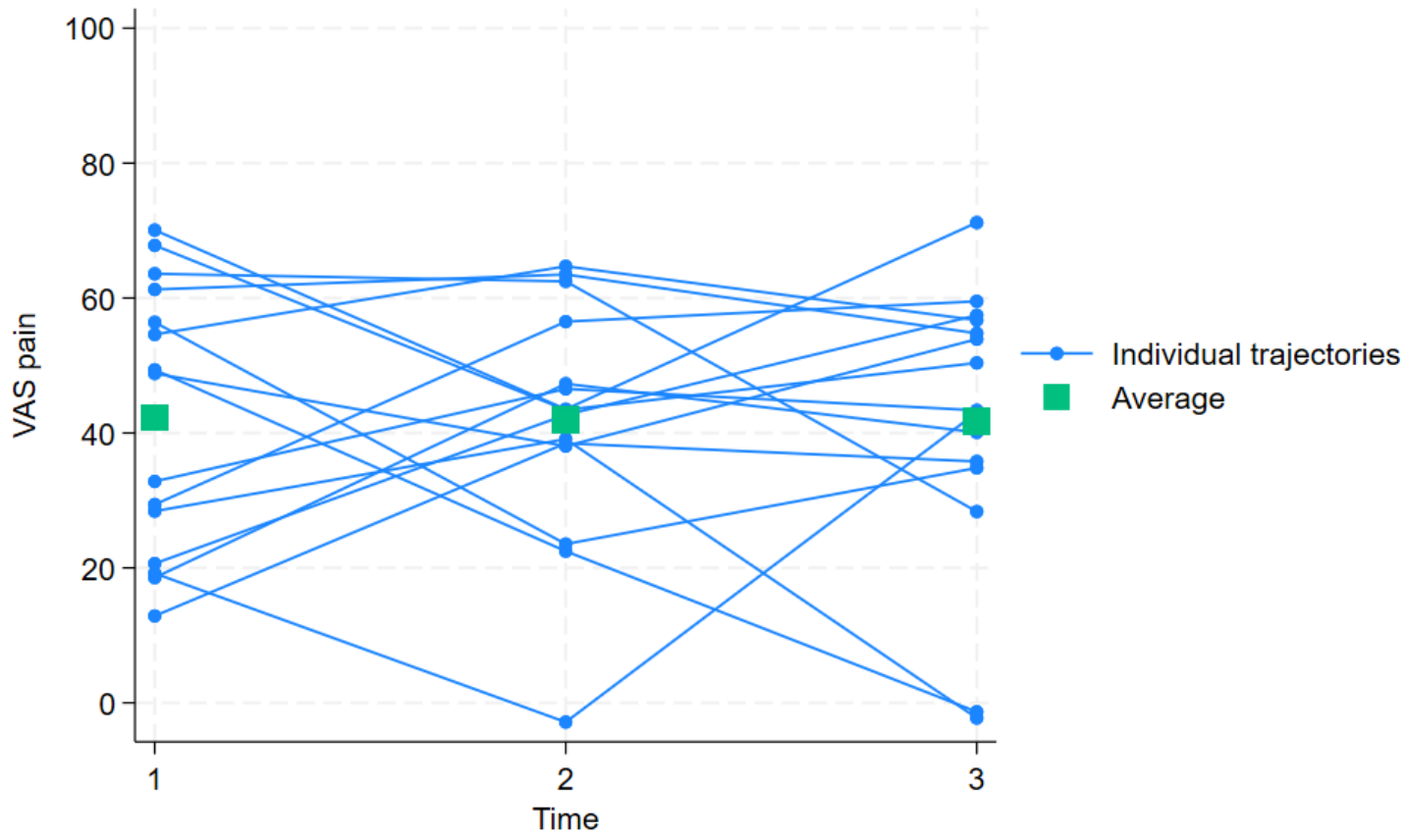
Consequences

- Regression to the mean (RTM)
- Occurs when:
  a) there is variability in measurements and
  b) there is selection at baseline
- The greater variability and selection, the larger RTM

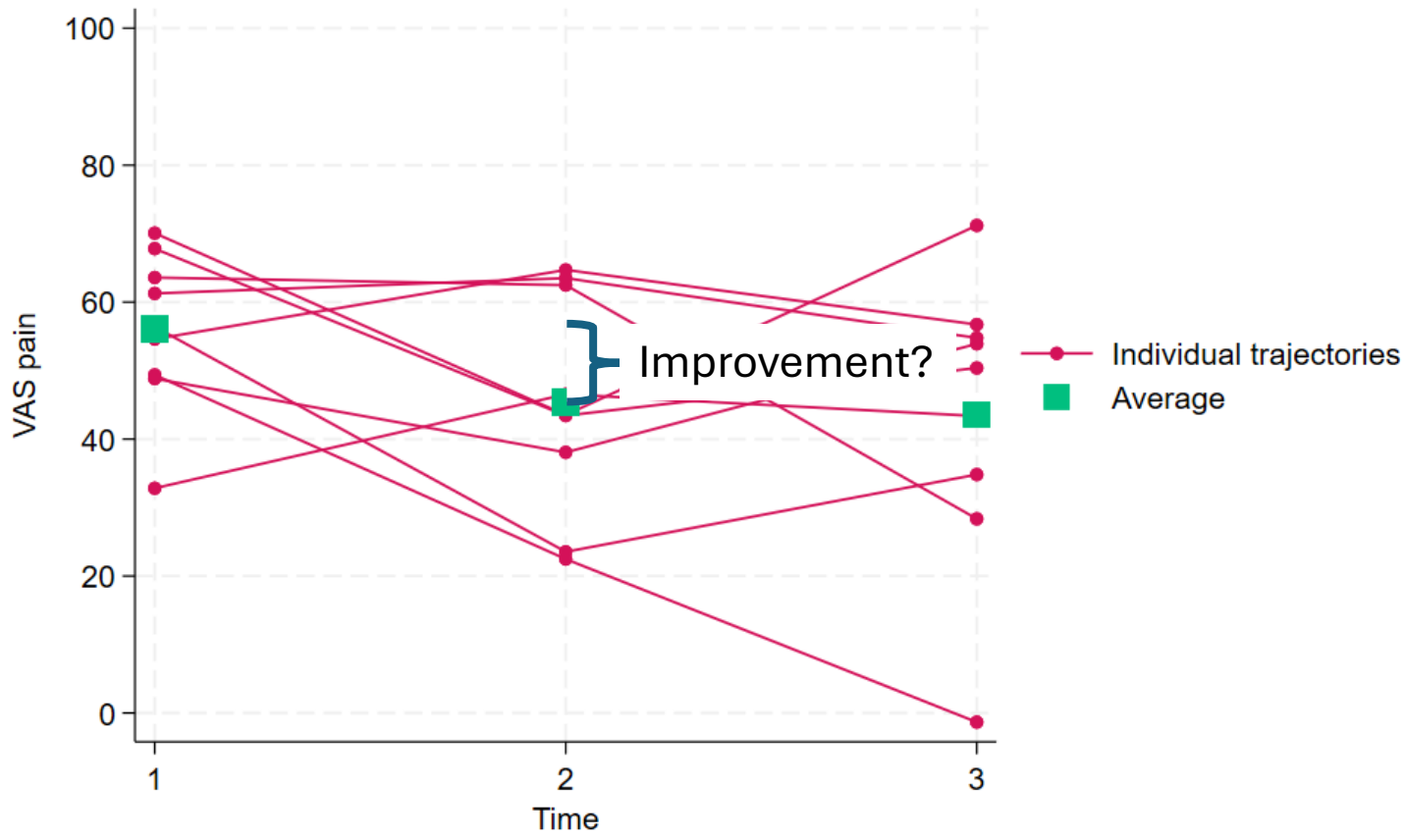
RTM



RTM

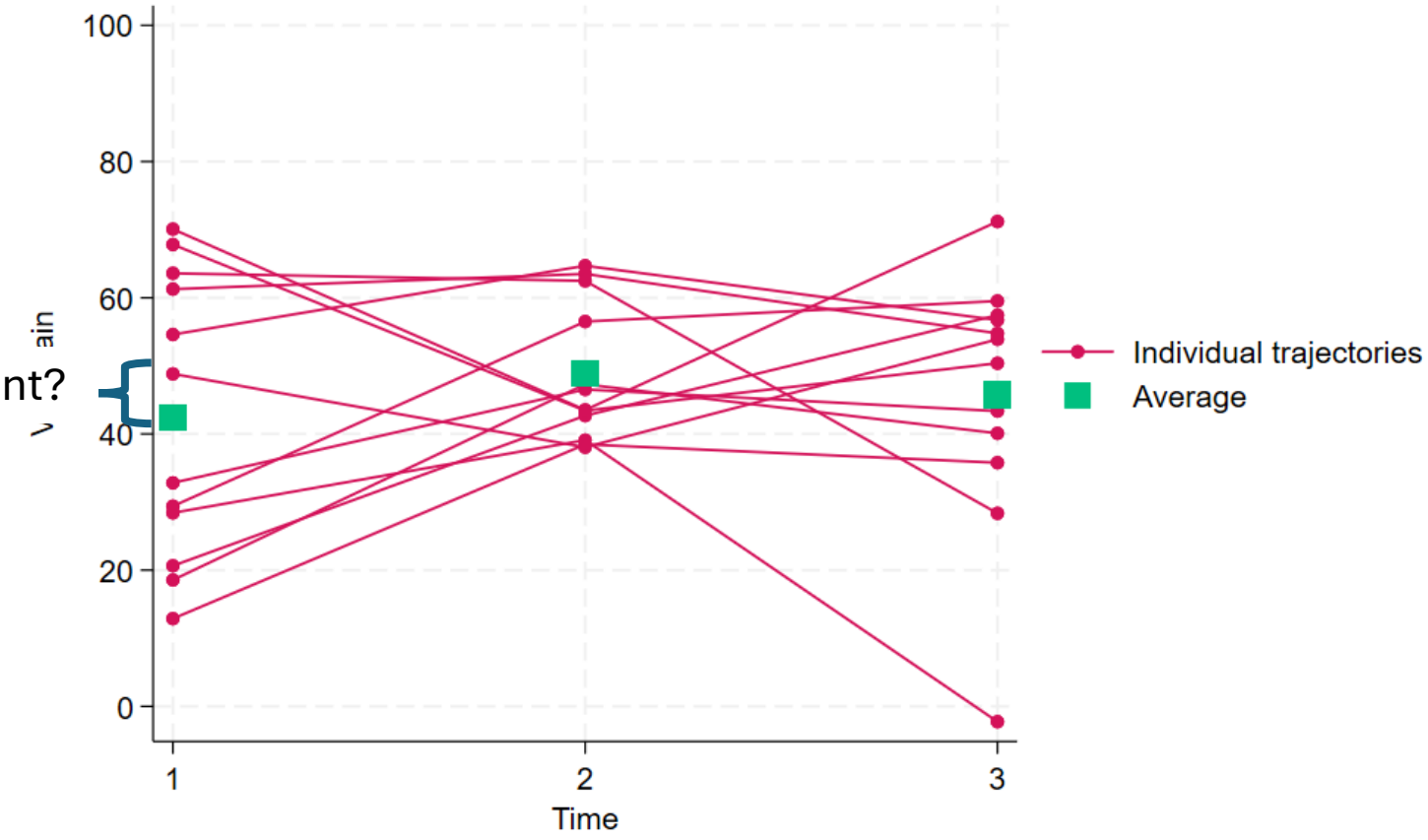

100
80
60
40
20
0
VAS pain
1
2
3
Time
Individual trajectories
Average

RTM


100
80
60
40
20
0
VAS pain
1
2
3
Time
Improvement?
Individual trajectories
Average

RTM backwards in time

Improvement?



Consequences in reporting of RCTs

Results: All groups improved WOMAC Pain, WOMAC Physical Function, PGA-OA, and average pain in the index joint over the 56-week treatment period relative to baseline. (...)
PMID: **35351194**

Results: We randomized 349 participants to SSM (n = 116), SSM+S (n = 116) or SSM+PS (n = 117) and 292 (84%) provided AUSCAN Osteoarthritis Hand Index hand pain scores at the primary end point (8 weeks). All groups improved, with no mean treatment difference between groups: (...)
PMID: **33254239**

Results: (...) All groups had improved WOMAC pain scores, (...)
PMID: **24740822**

Results: (...) There were no differences in MVIC between groups at Post 1 (P=0.32) or Post 2 (P=0.45), while all groups significantly improved from baseline
DOI: 10.1016/j.joca.2019.02.493

The actual results from these trials are

- There is no clinically relevant difference between the groups
- Size of RTM in OA trials with pain as outcome

THE LANCET
Rheumatology

COMMENT · Volume 5, Issue 6, E309-E311, June 2023 Download Full Issue

Pain in clinical trials for knee osteoarthritis: estimation of regression to the mean

Martin Englund · Aleksandra Turkiewicz

Painful MSK diseases

- Patient groups are heterogenous, different phenotypes or endotypes
- It is reasonable to speculate that there may be subgroups of patients that benefit from a treatment
- Subgroup analyses of RCTs are challenging with the main two problems being multiplicity and lack of precision
- Personalized medicine is a hot topic
- “Responder analyses” have become popular

Responder analysis

- Responder analysis aims at identifying individual responders
- Typically, a ”responder” is defined as a person with change from baseline larger than a (pre-defined) threshold

Responder analysis in MSK diseases

- There exist OMERACT-OARSI responder criteria for clinical trials (DOI: 10.1016/j.joca.2004.02.001)
- There exist “guidance for industry” from the FDA that states among others “*There may be situations where it is more reasonable to characterize the meaningfulness of an individual's response to treatment than a group's response, and there may be interest in characterizing an individual patient as a responder to treatment*” (PMID: 17034633)

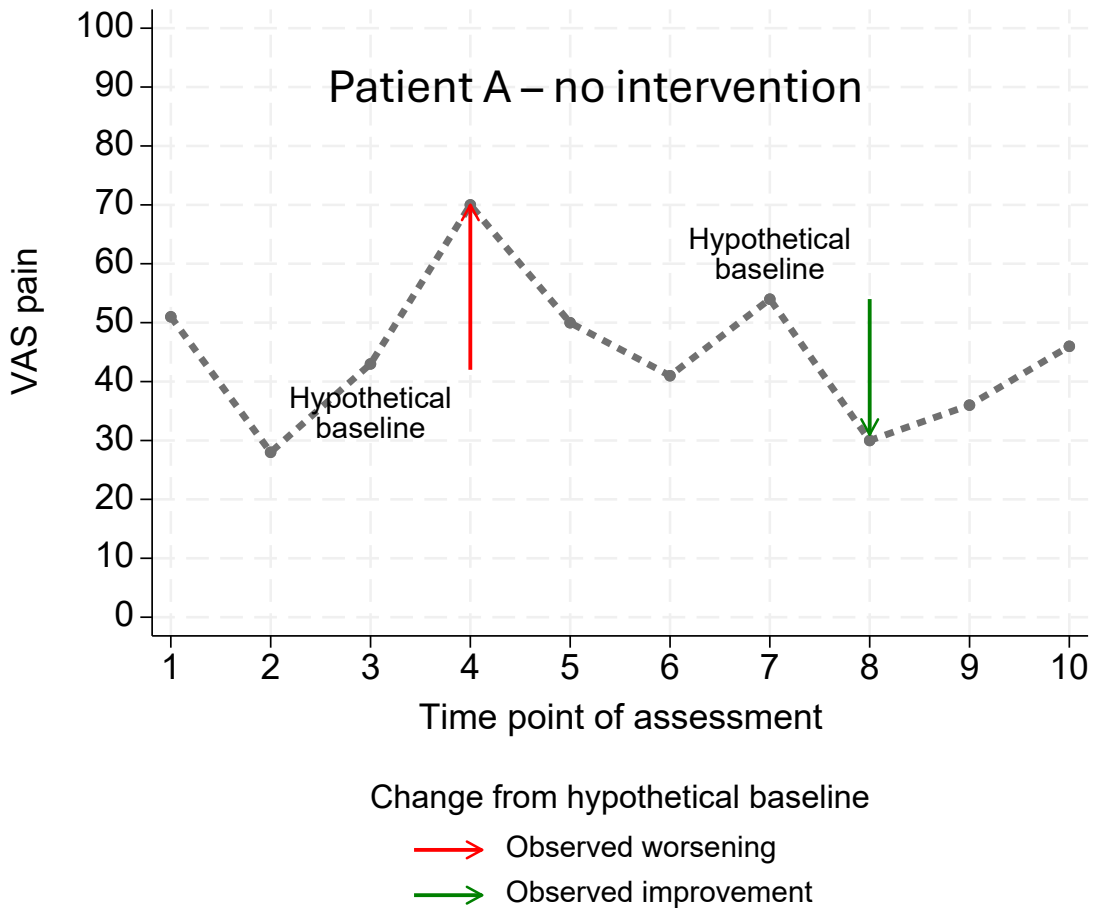
Aim

- Check how does naive definition of ”responder” work when applied to pain data from patients with osteoarthritis
- What happens when we apply the ”responder” analysis to data from RCTs?

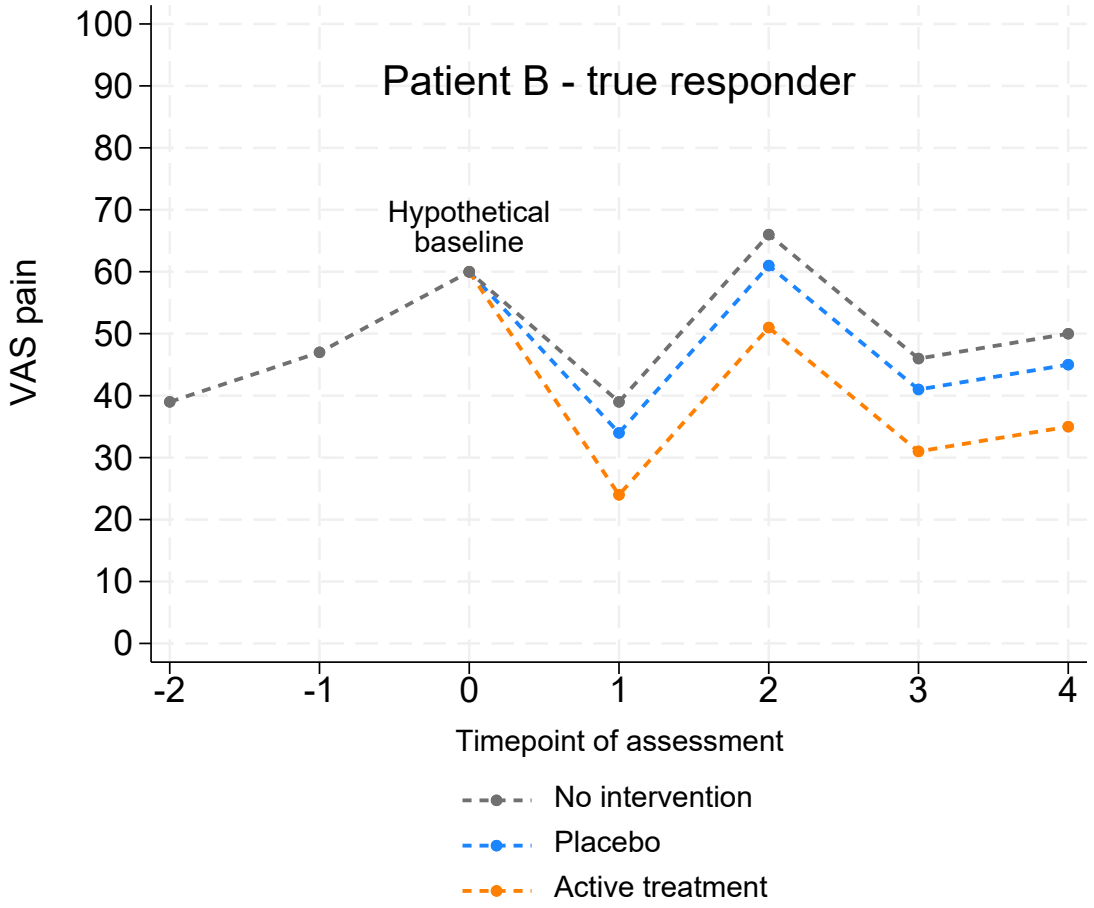
Assumptions

- Pain measured with VAS (0 to 100), follows normal distribution, but in the measurements are truncated between 0 to 100
- Within-persons SD is 12
- Between-person SD is 25
- Average pain in the underlying population is 45
- Clinically relevant change/difference is 10
- Placebo effect is 5
- Within-person constant response (i.e. the same person always has the same size of response)

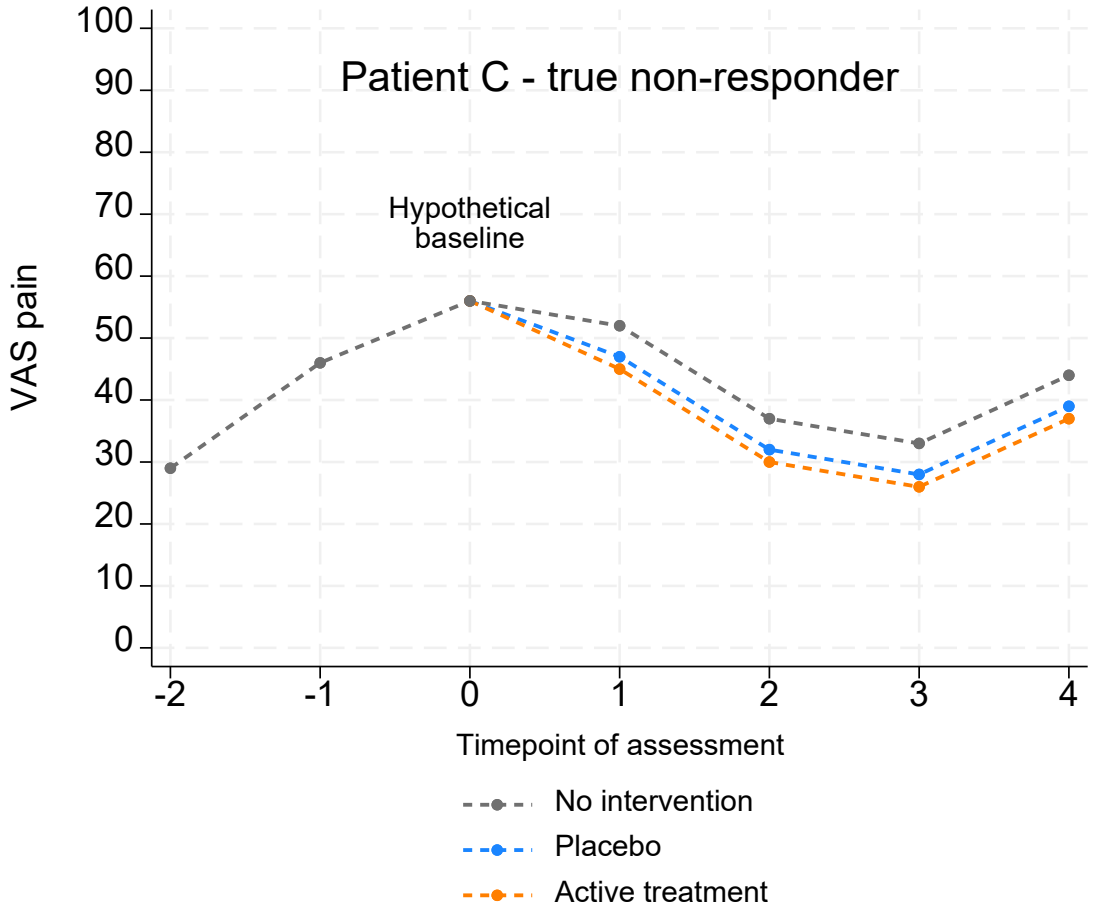
True and ”observed” responder



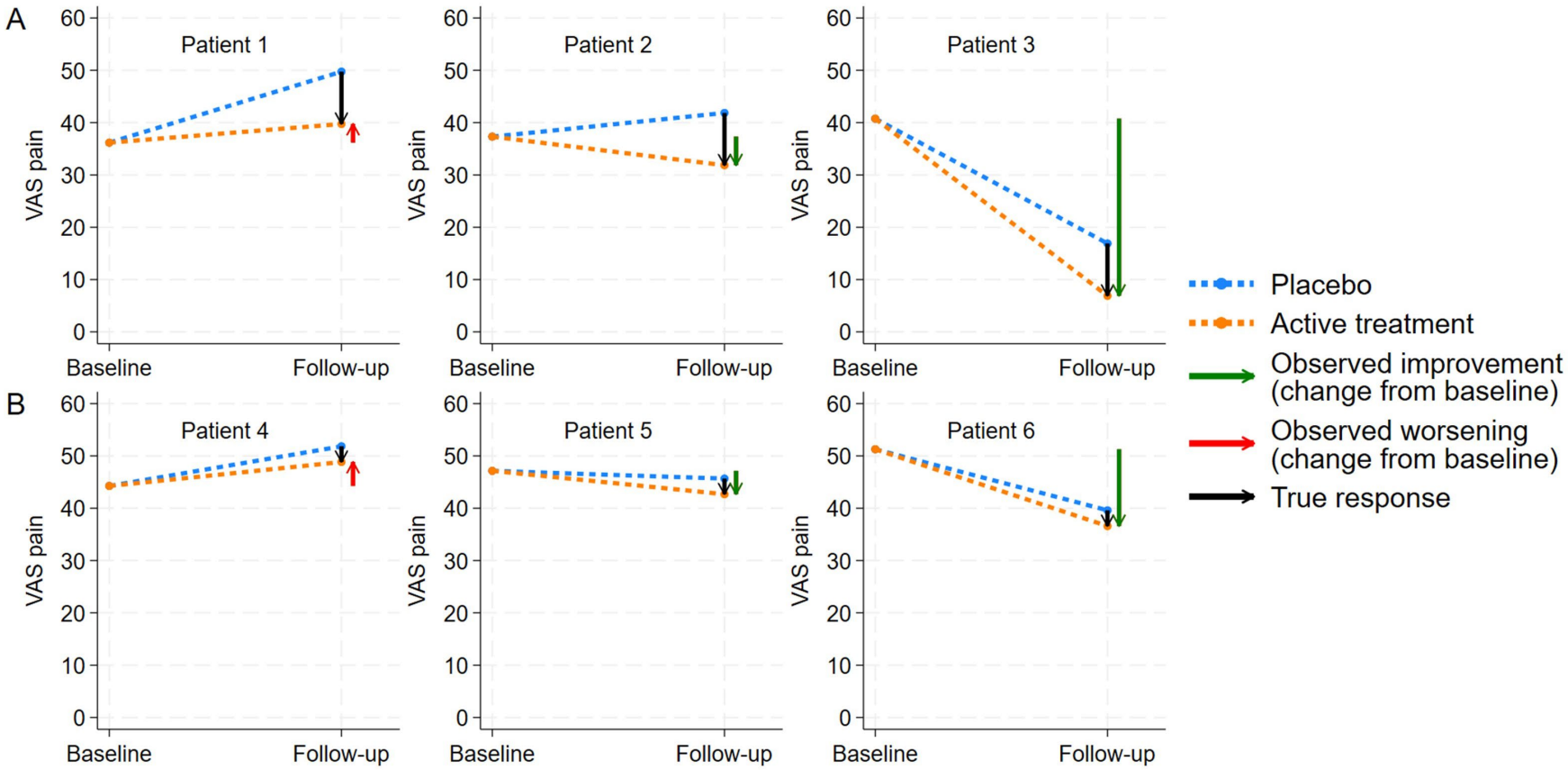
True and ”observed” responder



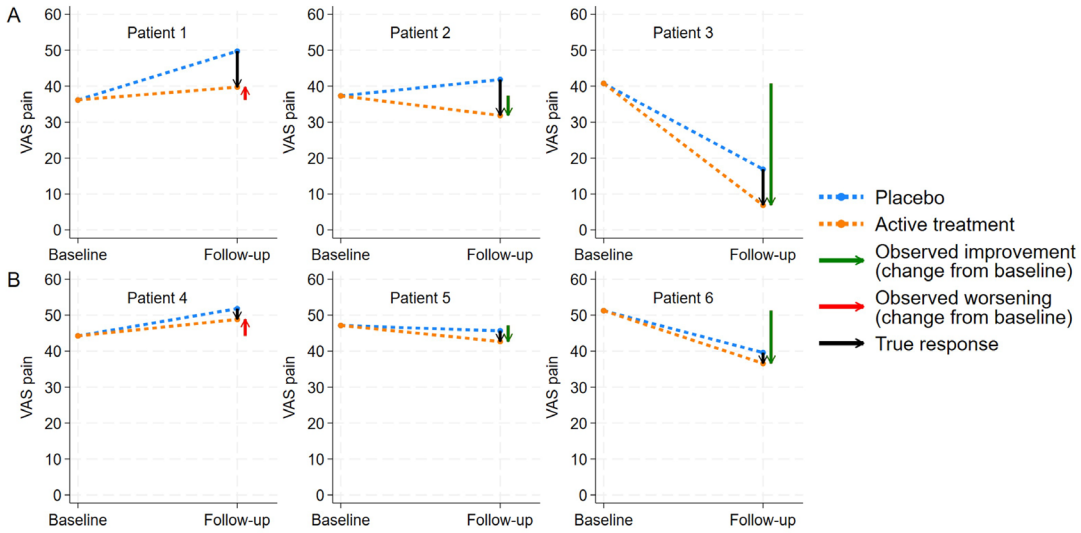
True and ”observed” responder



What happens in a trial



What happens in a trial



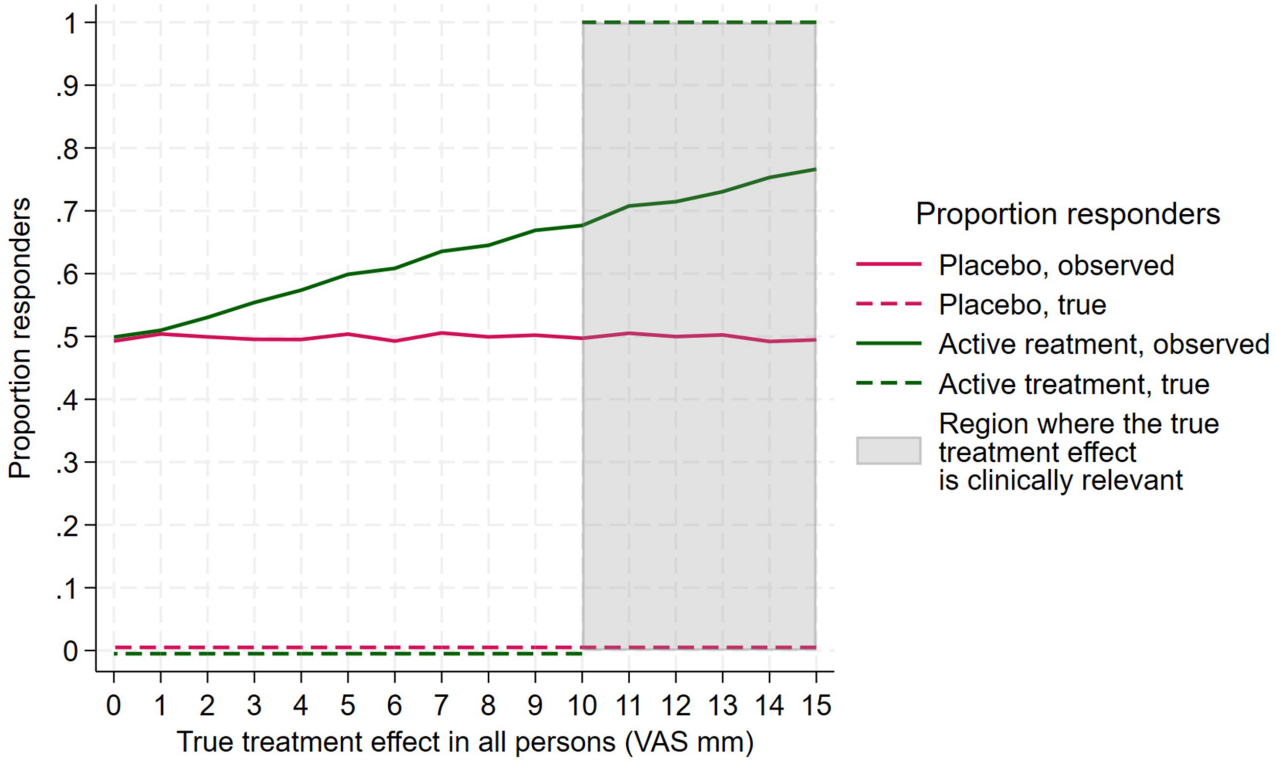
- There will be a lot of observed variability in change from baseline
- Should not be confused with variability in ”response”

Simulate trials

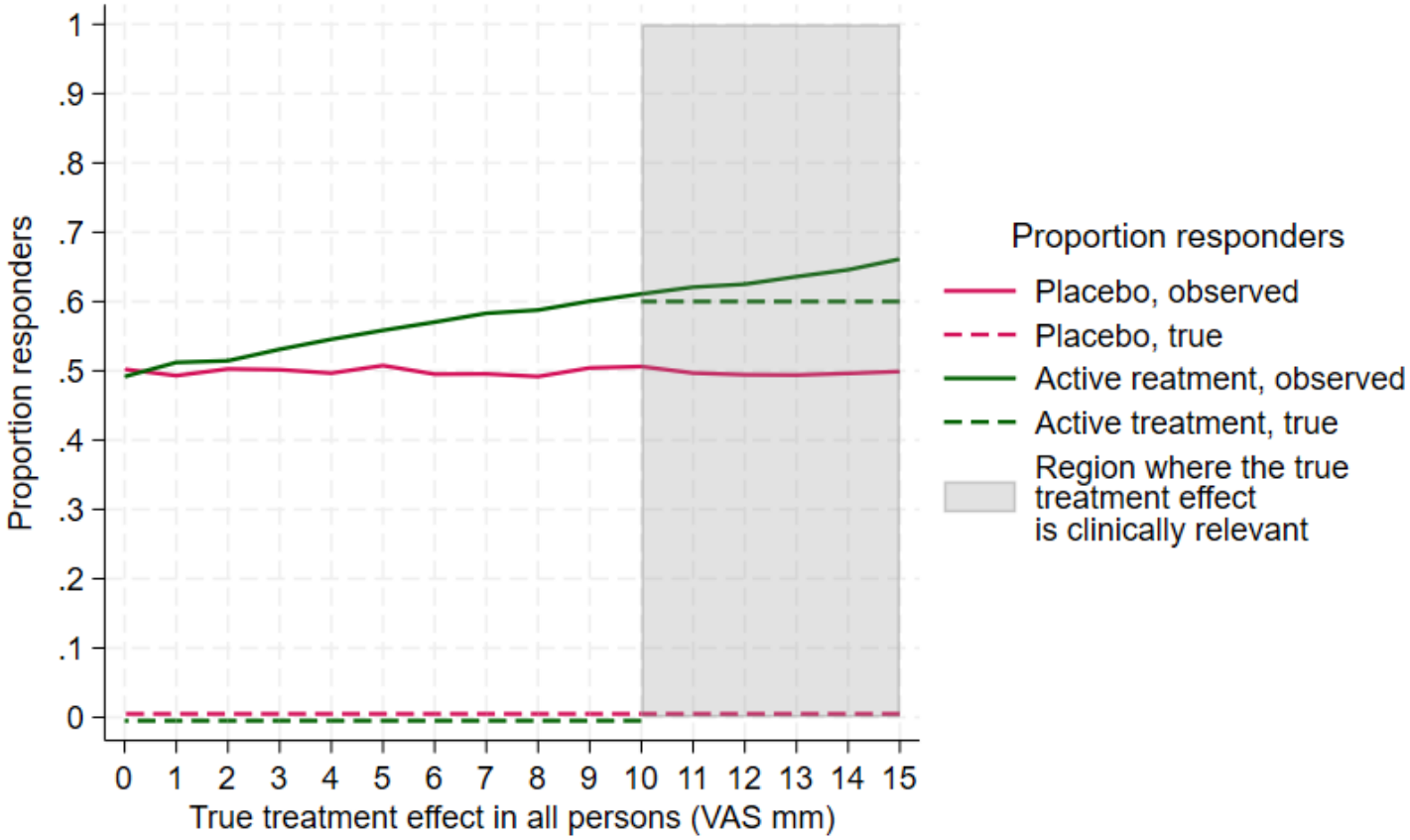
- 10 000 persons in each to minimize random variation
- Response the same in all persons (and constant within each person)
- Response (and thus treatment effect) from 0 (in trial nr 1) to 15 (in trial nr 16), i.e. 16 trials in total
- Apply typical responder analysis

Percentage responders per group in a trial

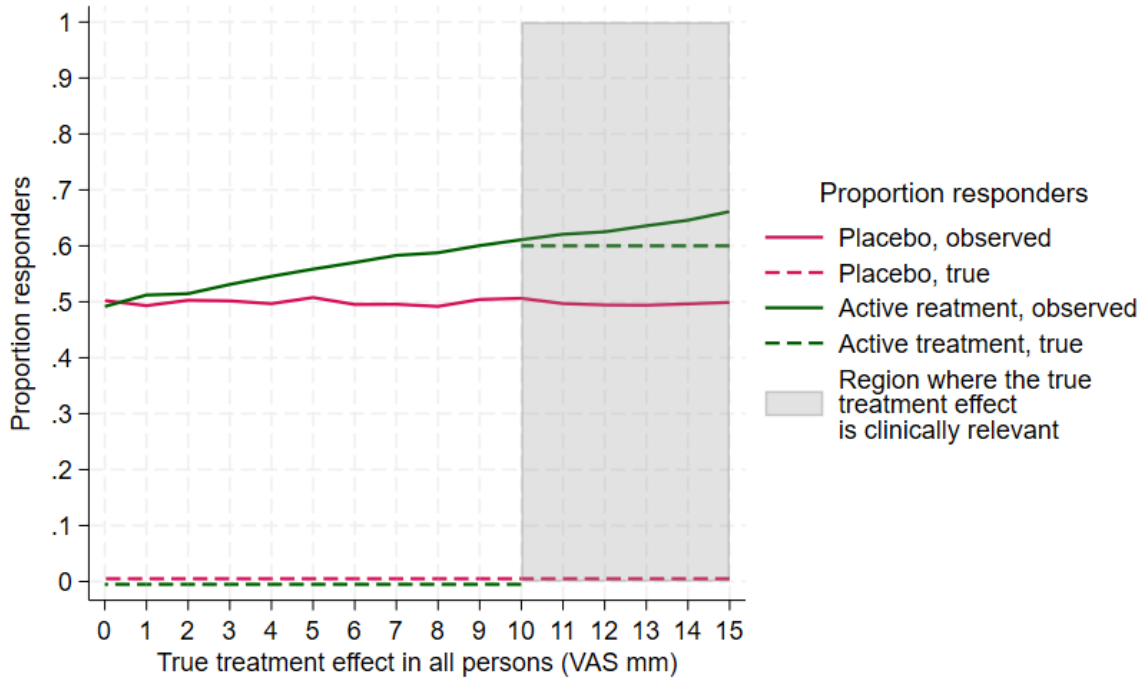
Figure 2. Simulation of a parallel-group randomized controlled trial with pain as outcome. On the x-axis is the size of true response to treatment in all persons (i.e. the treatment effect on top of placebo effect), and the response is the same in all persons. On the y-axis is both the proportion of true and observed responders per group. The increase in percentage true responders from 0% to 100% is when the response reaches level of clinical relevance (10 VAS points).



What if we assume that 60% of persons respond to treatment?



What if we assume that 60% of persons respond to treatment?



- The sensitivity of naive responder definition is ~73%
- The specificity is ~50%

You may think I used ”extreme” assumptions

- Two types of NSAIDs vs placebo (two trials) in osteoarthritis, OMERACT-OARSI criteria (https://doi.org/10.1093/rheumatology/kel296)

” *There were significantly greater proportions of responders in the etoricoxib (66.2%) and celecoxib (63.5%) groups compared with the placebo group (43.0%; P < 0.001).*”

- Exercise for OA, GLA:D® program (no control/placebo group):

”45% responders, 10% deteriorated”

- Other NSAIDs vs placebo (https://doi.org/10.1186/ar2394):

” *There was a very wide individual range of responses to the various treatments*.”

You may think I used ”extreme” assumptions

- Two types of NSAIDs vs placebo (two trials) in osteoarthritis, OMERACT-OARSI criteria (https://doi.org/10.1093/rheumatology/kel296)

”*There were significantly greater proportions of responders in the etoricoxib (66.2%) and celecoxib (63.5%) groups compared with the placebo group (43.0%; P < 0.001).*”

- Exercise for OA, GLA:D® program (no control/placebo group):

”45% responders, 10% deteriorated”

- Other NSAIDs vs placebo (https://doi.org/10.1186/ar2394):

~~*” There was a very wide individual range of responses to the various treatments.”*~~

”*There was a very wide individual range of* ~~*responses*~~ ***changes from baseline*** *after the various treatments*.”

Can we find true responders?

Special Issue Paper | Open Access

Mastering variation: variance components and personalised medicine

Stephen Senn

Can we find responders?

Table 1. Sources of variation in a clinical trial.

	Type of variation	Definition
A	Between treatments	The variation between treatments averaged over all patients
B	Between patients	The variation between patients given the same treatments
C	Patient-by-treatment interaction	The extent to which the effects of treatments vary from patient to patient
D	Within patients	Variation from occasion to occasion when the same patient is given the same treatment

	Type of variation
A	Between treatments
B	Between patients
C	Patient-by-treatment interaction
D	Within patients

Table 2. Identifiable sources of variation according to trial design.

Type of trial	Description	Identifiable components of variation	'Error' term
Parallel group	Patients are randomised to a course of one of the treatments being compared, which they then follow for the period of the trial.	A	B + C + D
Classical cross-over	Patients are randomised to sequences of treatments (to be taken in different periods) with the purpose of studying differences between treatments. Each treatment being compared is studied in one period.	A, B	C + D
Repeated period cross-over	Patients are randomised to sequences in which they are treated by each treatment in more than one period.	A, B, C	D

Repeated period cross-over designs

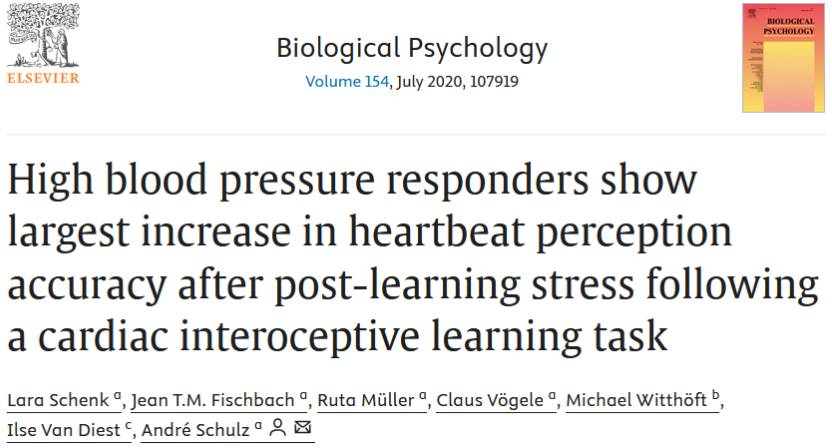
- Are most often not feasible for treatments for MSK pain, especially for interventions such as exercise
- I.e. there is no evidence that “responders” and “non-responders” exist, i.e. no evidence of patient by treatment interactions in MSK pain
- Such indirect evidence could be inferred from larger variance in the treatment group than placebo, but I have not seen this

Example of a repeated period cross-over

- March et al.
  (PMID: 7950736)
- 25 patients with osteoarthritis treated during 6 periods of two weeks each, 3 with paracetamol and 3 with diclofenac, double blinded and randomized
- Inefficient analysis, with separate tests for each patient and erroneous interpretation of lack of significance as equivalence
- Senn has attempted a partial reconstruction of the original data and attempted a random effects analysis, which does suggest some heterogeneity
- So we still do not know…

Only MSK pain?

- Variable outomes are common in biomedicine (blood pressure, biomarkers, etc.) – beware of similar problems



There are good examples as well

Original Investigation

FREE

April 11, 2023

Heterogeneity in Blood Pressure Response to 4 Antihypertensive Drugs

A Randomized Clinical Trial

Johan Sundström, MD, PhD[1,2]; Lars Lind, MD, PhD[1]; Shamim Nowrouzi, MD[1]; et al

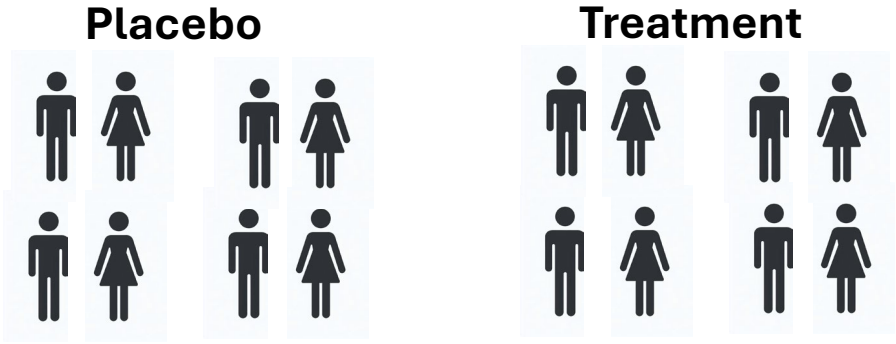
» Author Affiliations | Article Information

JAMA. 2023;329(14):1160-1169. doi:10.1001/jama.2023.3322

Design, Setting, and Participants A randomized, double-blind, repeated crossover trial in men and women with grade 1 hypertension at low risk for cardiovascular events at an outpatient research clinic in Sweden. Mixed-effects models were used to assess the extent to which individuals responded better to one treatment than another and to estimate the additional blood pressure lowering achievable by personalized treatment.

How can we use parallel group trials to find true responders?

- Subgroup analyses – if a subgroup is the unit of interest (as opposed to individual patient), then a ”standard” parallel group trial provides replication required to estimate interaction between subgroup and treatment



- Consequences for the study design (or post-hoc pooling of studies) with respect to e.g. sample size and pre-specification of subgroups

Take home messages (1)

- Self-reported pain (function, quality of life etc.) is highly variable in persons with MSK conditions
- Change from baseline is not ”response” to treatment
- Individual responders can not be identified from a parallel group trial
- There is almost no evidence that there exist differential response to treatments in MSK pain (i.e. the search for ”responders” could be futile!)

Take home messages (2)

- When you see the term ”responder” in a paper – be suspicious
- If you see the term “personalized medicine” - be suspicious and check that it is not individual “responder” analysis that is meant

Clinical Epidemiology Unit

FACULTY OF MEDICINE | LUND UNIVERSITY

https://clinicalepidemiology.se/