MRC Biostatistics Unit
UNIVERSITY OF CAMBRIDGE
Funded by NIHR | National Institute for Health and Care Research

Eight Methodological Questions for Digital Outcome Measures

Basel Biometric Society

Mia Tackney, MRC-Biostatistics Unit

Joint work Prof. James Carpenter and Prof. Sofía Villar

The Six-Minute Walk Test

Gold-standard outcome for clinical trials in many diseases e.g, Duchenne Muscular Dystrophy and Pulmonary Hypertension.

Some recognised limitations:

- Places high burden on patients;
- May not meaningfully measure patient experience;
- 'Snapshot' measure: affected by fatigue, motivation.

Digital Outcome Measures

Digital Health Technologies enable trial outcomes to be measured remotely in patient's chosen locations.

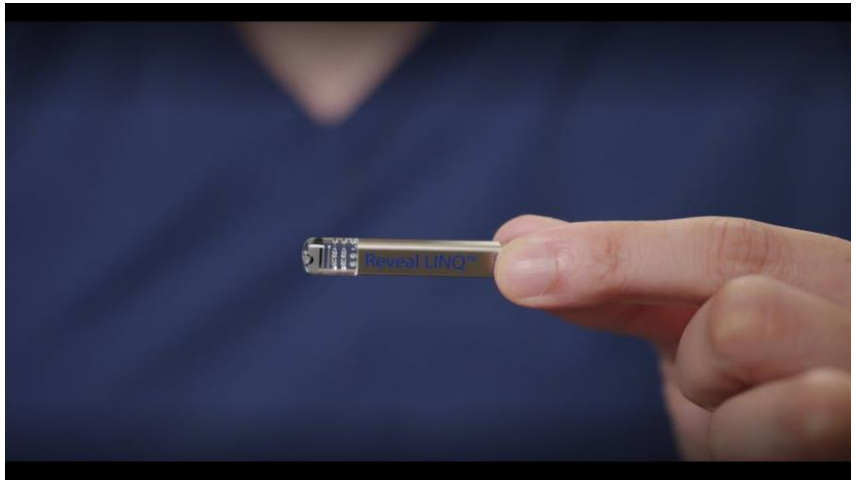
This allows:

- Decentralisation of the trial;
- More frequent and convenient measurements;
- Potential for more sensitive measures.

Passive Monitoring

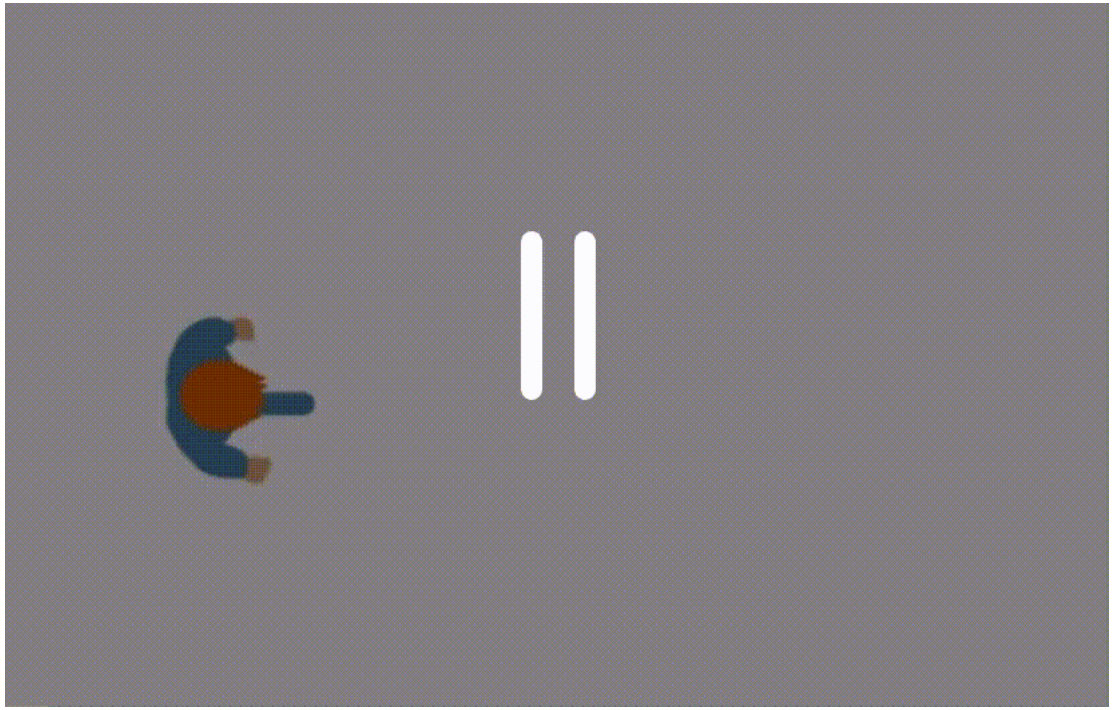
SV95C for Duchenne Muscular Dystrophy. Currently the only EMA-qualified digital outcome measure [Servais et al 2022].

Passive Monitoring



Insertable Cardiac Monitor used in Pulmonary Hypertension [Rothman et al (2025)].

Active Tests



Floodlight: Digital platform for Multiple Sclerosis (Roche).

Many Developments...

V3+ Framework

EMA Qualification of SV95C

28 July 2023
EMADOC-1700519818-1127132
Committee for Medicinal Products for Human Use (CHMP)

Qualification Opinion for Stride velocity 95th centile as primary endpoint in studies in ambulatory Duchenne Muscular Dystrophy studies

Guidance From FDA

Digital Health Technologies for Remote Data Acquisition in Clinical Investigations

Guidance for Industry, Investigators, and Other Stakeholders

Eight methodological questions

Q1: What is the appropriate outcome granularity?

Q2: What is the appropriate measurement period?

Q3: How is the summary measure chosen?

Q4: How do we exploit the granularity of digital outcome measures to learn more about the treatment?

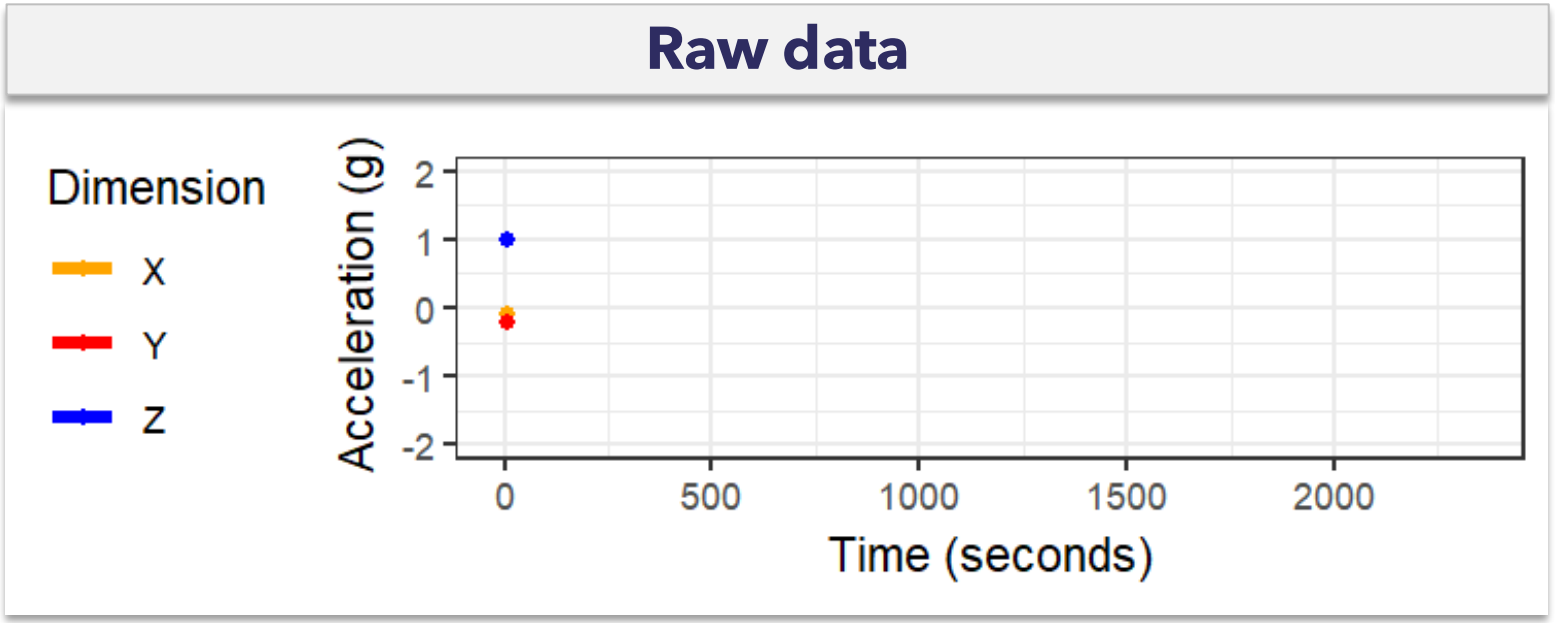
Q5: How do seasonal/behavioral variations affect design/analysis?

Q6: How are missing data defined and handled?

Q7: What are appropriate estimands for digital outcome measures?

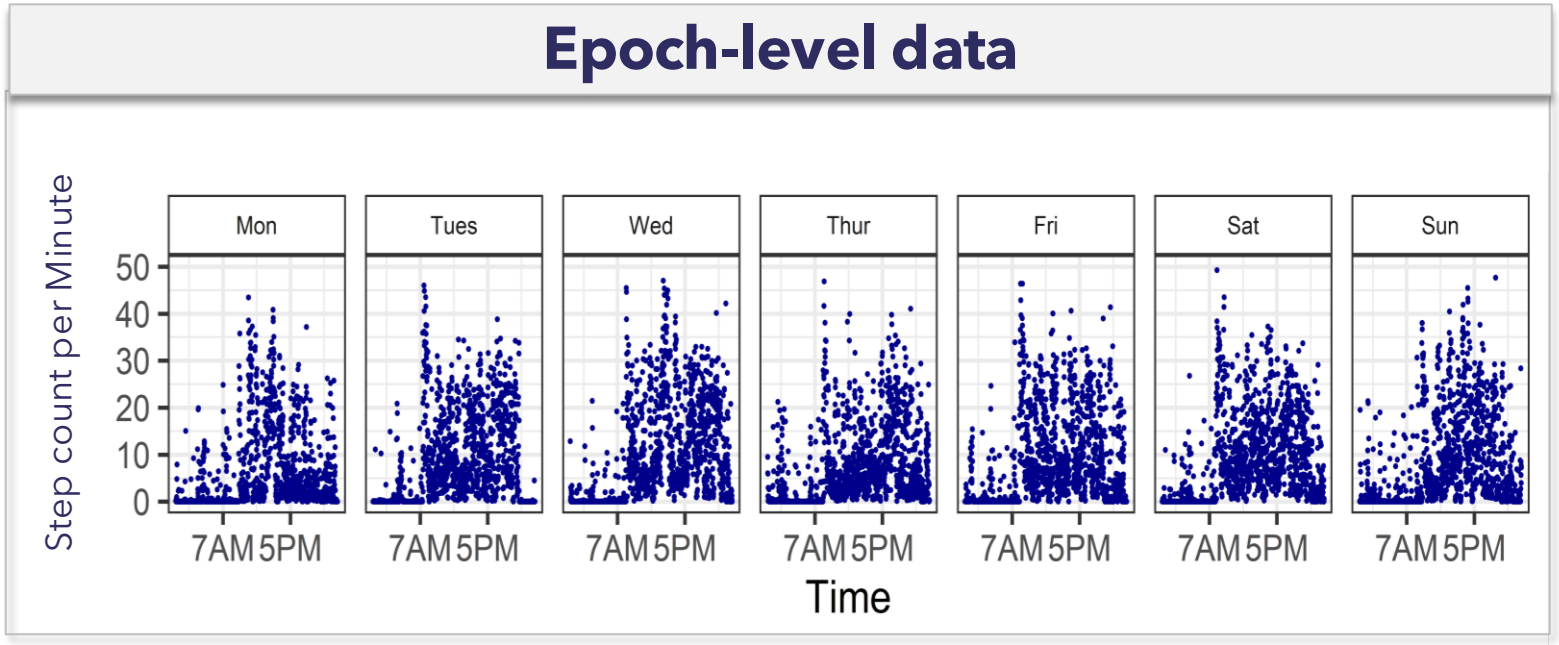
Q8: Is evidence synthesis possible?

Question 1: What is the appropriate granularity of the outcome?



- Raw data: physical signals at high frequency
- Example: Acceleration in three dimensions at 100Hz
- **Technical** validation required
- Not always available to researcher

Question 1: What is the appropriate granularity of the outcome?



- Epoch level data: raw data processed into a physiological metric for a small unit of time
- Example: Step-count at the minute-level
- **Analytical** validation required

Question 1: What is the appropriate granularity of the outcome?

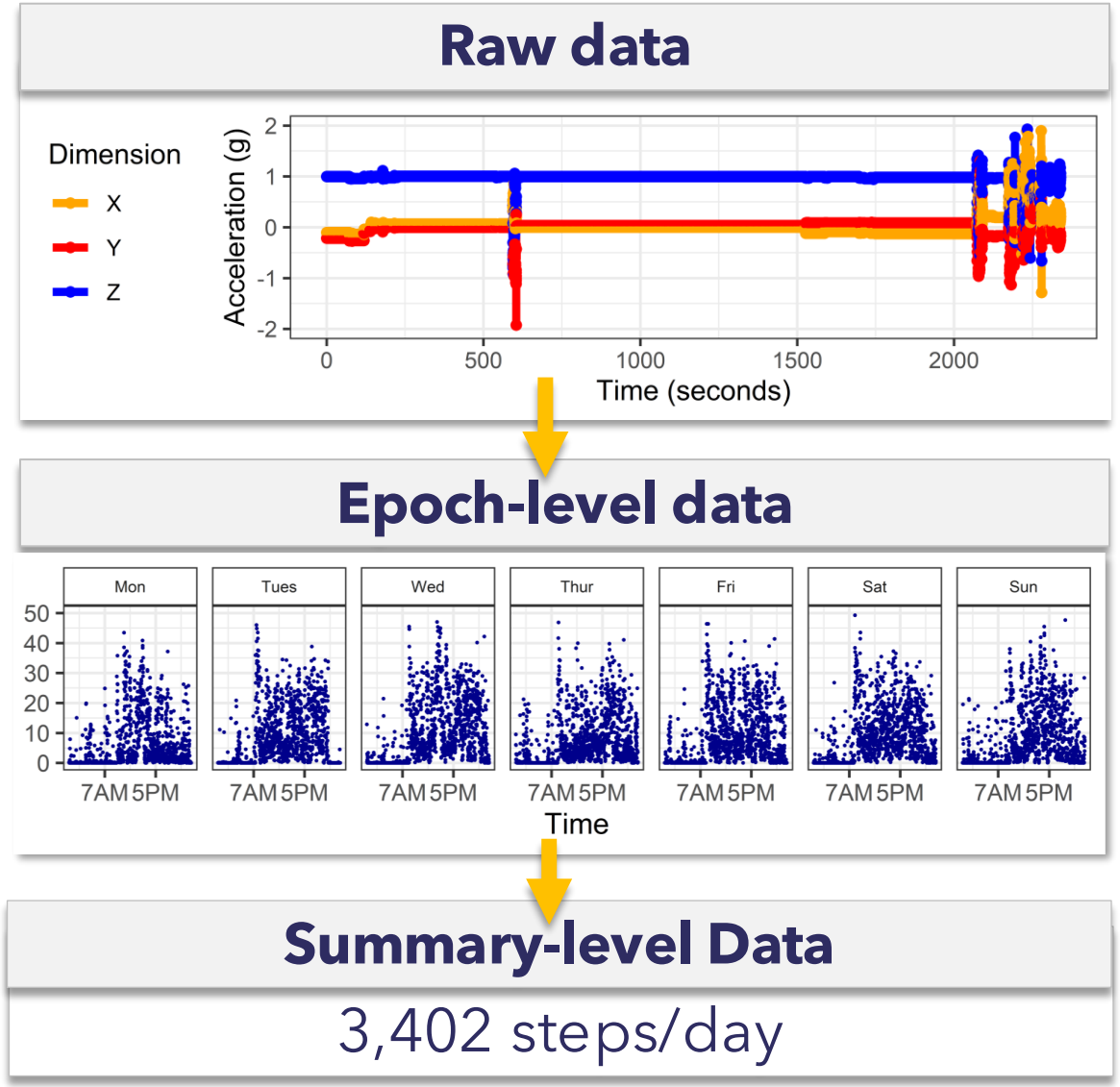
Summary-level Data

Average total step count per day:

3,402

- Scalar summary to represent physiological state
- Example: average total step count
- **Analytical** and **clinical** validation required: reliability, convergent validity, known-groups validity, sensitivity to change

Question 1: What is the appropriate granularity of the outcome?



Many decisions are being made as you aggregate the data, e.g.:

- Filtering outliers
- Handling of missing data…

Additional clinical questions can be answered with granular data.

Question 2: What is the appropriate measurement period?

Longer measurement periods reduce variability in summary metrics ...
… but this must be balanced against **compliance and burden.**

Studies with longer term monitoring show that engagement tends to drop off quickly, particularly for active tests.

[Midaglia et al. 2019, Taylor et al. 2025]

Timing of follow-up periods are crucial, given that digital outcomes may be more sensitive to change.

Question 3: How is the ideal summary measure chosen?

Pre-specified scalar summaries are typically validated.

Recently, **data-driven approaches** have been used to create summaries:

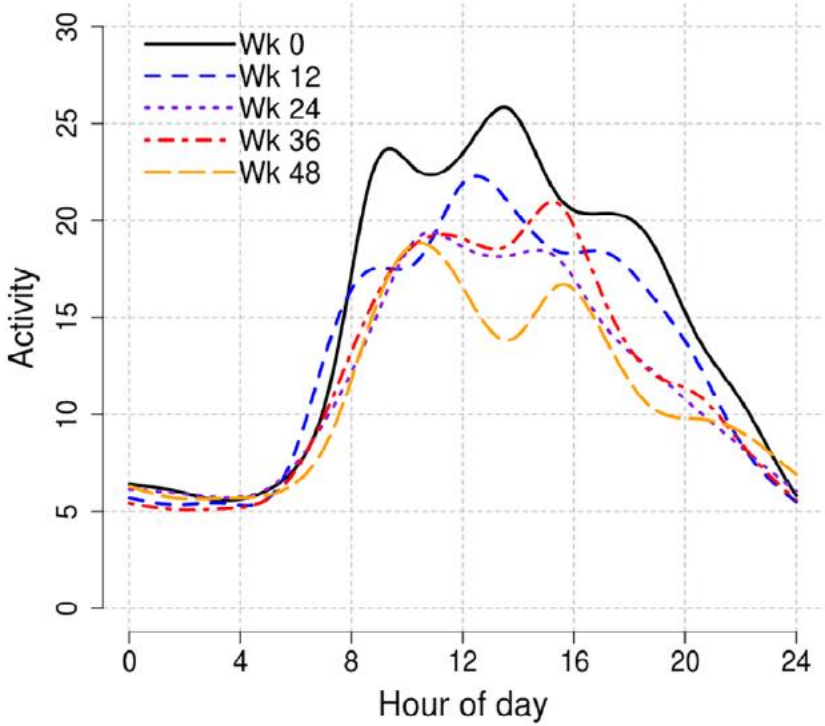
- Machine-learning based severity scores from speech data have been developed for MND [Webber et al 2026].
- Machine Learning to combine data from different digital active tests in Parkinson's Disease [Zhai et al 2025].
- Scores from a Multi-level Functional Principal Components Analysis in ECGs and Gait data [Tackney et al 2026].

Question 4: How do we exploit the granularity of digital outcome measures to learn more about clinical change?

We may gain insight into **diurnal patterns.**

For example, Generalised Additive Models, have shown that daily patterns in physical activity can dip in the middle of the day with progression of ALS.

Potential to inform times of day when outcome measures may be more sensitive.

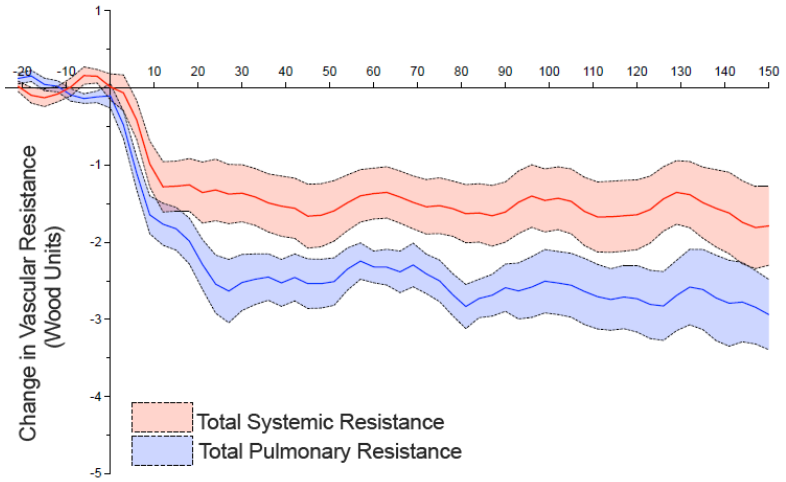


Estimated smooth curves of physical activity for 5 selected week-long periods in an observational study on ALS [Lisi and Abellan (2023)]

Question 4: How do we exploit the granularity of digital outcome measures to learn more about clinical change?

We may gain insight into the **mechanism of action** of a treatment.

Data from implantable devices demonstrated that changes in physiology occurred **earlier** than expected following treatment initiation/withdrawal.

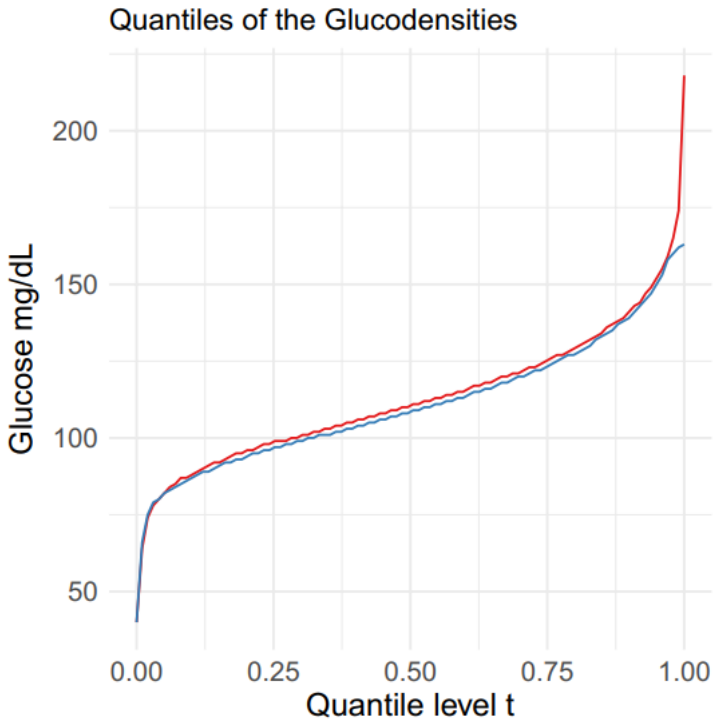


Insertable cardiac monitor outcomes following initiation and withdrawal of Imatinib in Pulmonary Hypertension patients [Rothman et al 2025 PIPAH]

Question 4: How do we exploit the granularity of digital outcome measures to learn more about clinical change?

We may gain insight into the **distributional profiles** of digital biomarkers.

Glucodensities have been shown to discriminate between individuals with diabetes and healthy controls [Matabuena et al (2025)].



Glucodensity representations for individuals with diabetes (red) and healthy controls (blue) [Matabuena et al (2025)]

Question 5: What are appropriate estimands for digital outcome measures?

Population

Variable

Intercurrent Events

Treatment

Population level summary

Izem et. al (2024)

Question 5: What are appropriate estimands for digital outcome measures?

Population

Variable (digital): high-frequency measurements may be more sensitive to clinical change

Intercurrent Events

Treatment

Population level summary

Izem et. al (2024)

Question 5: What are appropriate estimands for digital outcome measures?

Population

Variable (digital): high-frequency measurements may be more sensitive to clinical change

Intercurrent Events

Treatment

Population level summary: validated summary needed; considerations around seasonal/behavioral variation, missing data.

Izem et. al (2024)

Question 5: What are appropriate estimands for digital outcome measures?

Population: may broaden or shift target population; selection bias

Variable (digital): high-frequency measurements may be more sensitive to clinical change

Intercurrent Events

Treatment

Population level summary: validated summary needed; considerations around seasonal/behavioral variation, missing data.

Izem et. al (2024)

Question 5: What are appropriate estimands for digital outcome measures?

Population: may broaden or shift target population; selection bias

Variable (digital): high-frequency measurements may be more sensitive to clinical change

Intercurrent Events: new ICs such as software updates, device malfunction

Treatment

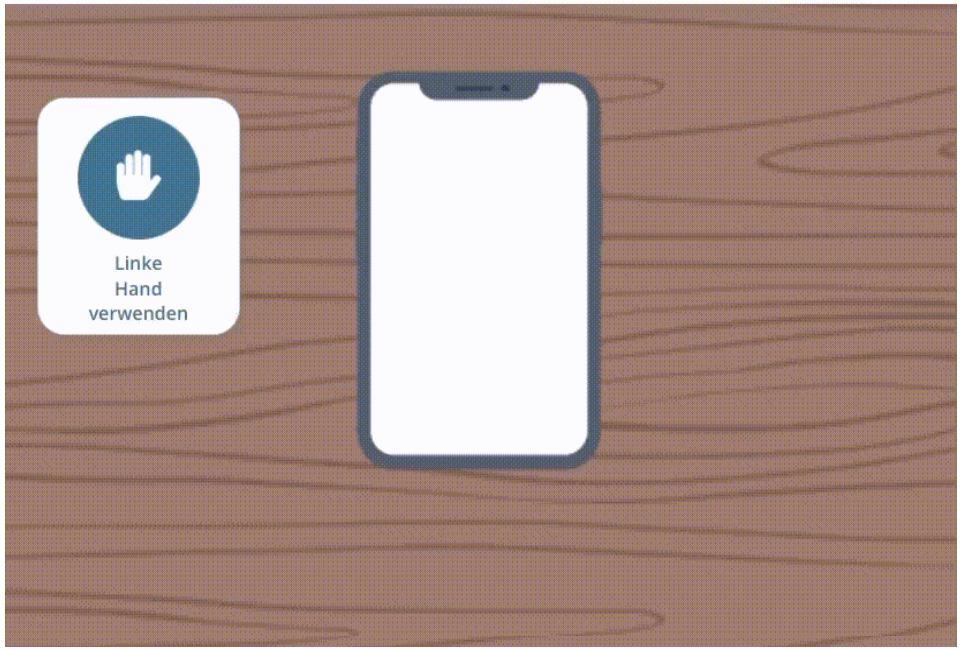
Population level summary: validated summary needed; considerations around seasonal/behavioral variation, missing data.

Izem et. al (2024)

Question 6: How do seasonal/behavioural variations affect design/analysis?

Bounded growth mixed models demonstrated [Woelfle et al 2021]:

- Strong **long-term** practice effects for the Symbol Digit Modalities test;
- Small to modest practice effects for finger pitching test;

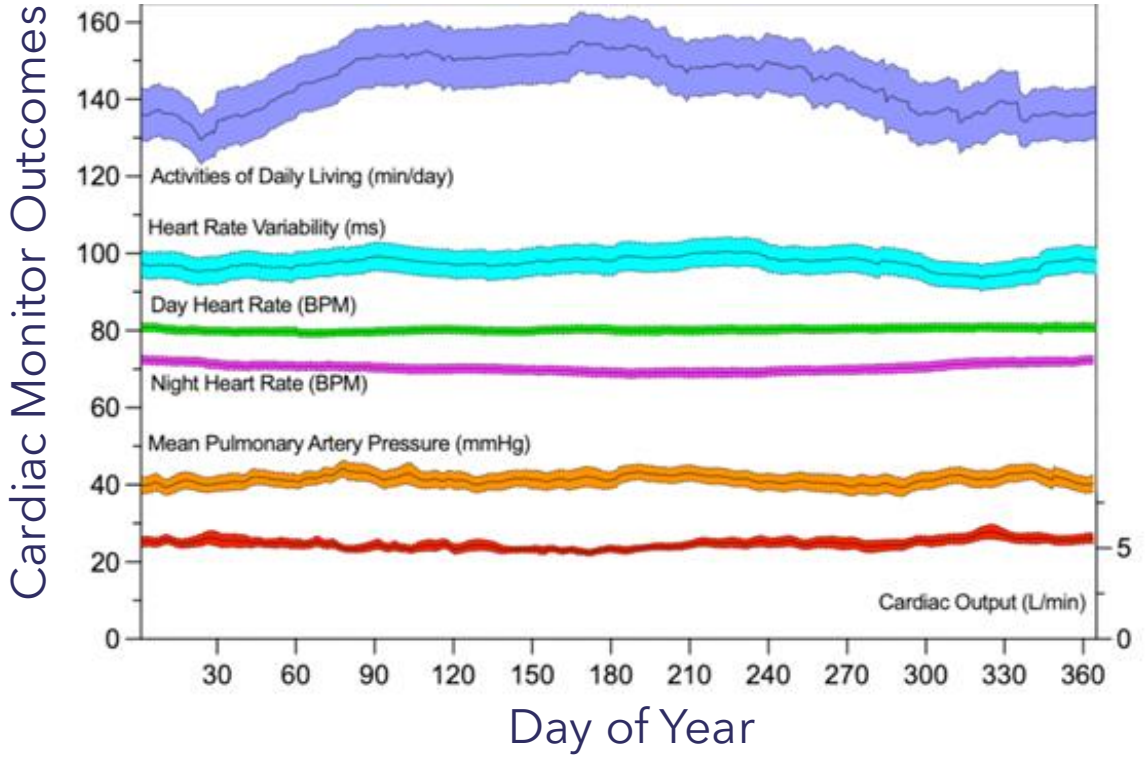


Question 6: How do seasonal/behavioural variations affect design/analysis?

Passive measures of physical activity will vary by **season**.

In observational or single-arm studies: potential for bias.

Need to consider carefully the timing of recruitment; adjusting for season in the analysis may be beneficial.



Insertable cardiac monitor outcomes in patients with Pulmonary Hypertension by day of the year (mean +/- SEM) [Rothman et al (2025)]

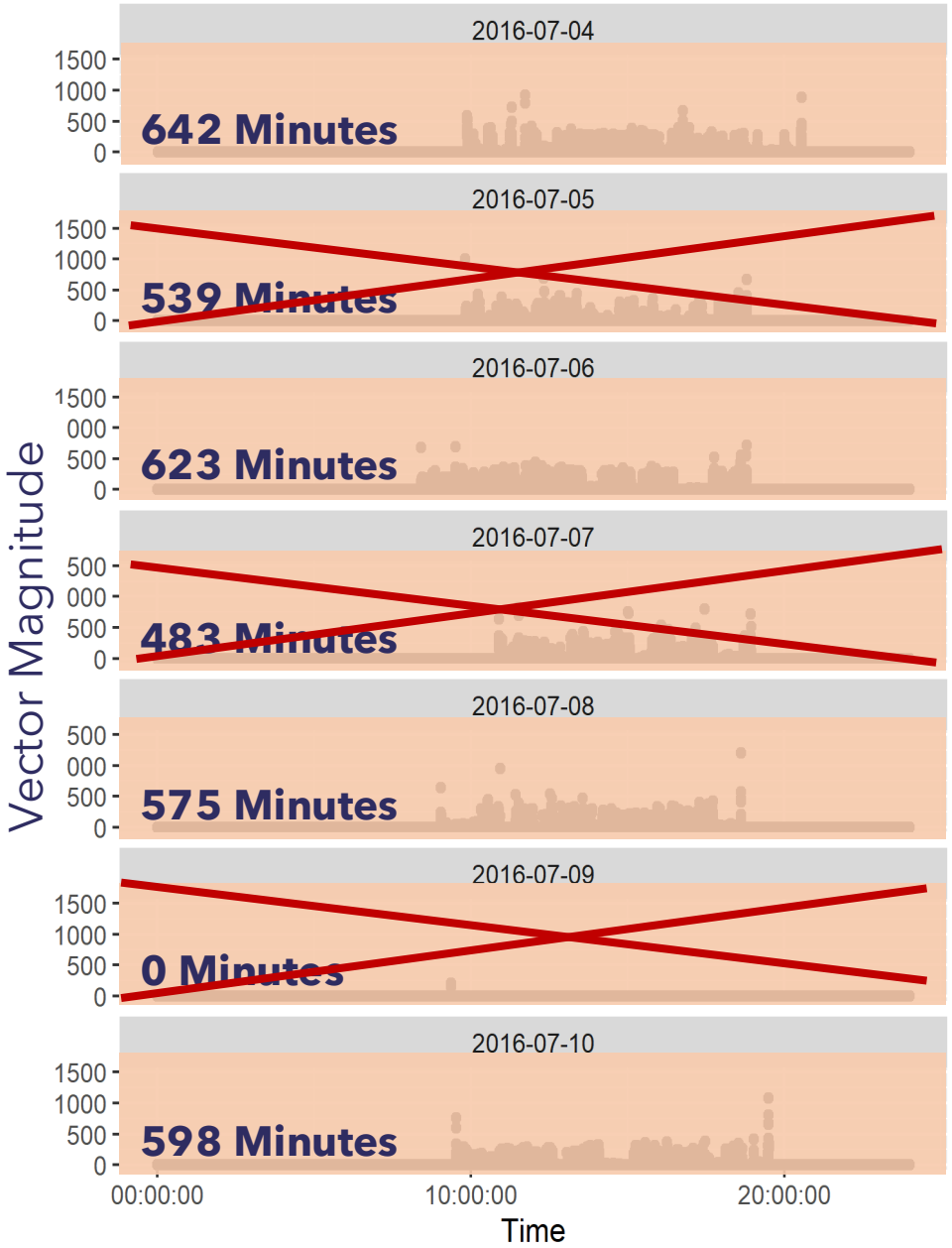
Question 7: How are missing data defined and handled?

Defining missing data is may be complex for digital outcomes.

Weartime thresholds often chosen ad-hoc and may not be suitable for the chosen population [Airlie et al 2022].

Imputation framework on the epoch-level data has been proposed [Tackney et al 2021, 2023].

Sensitivity analyses to the Missing at Random (MAR) assumptions are encouraged; multi-modal settings may provide insights.



Question 8: Is evidence synthesis possible?

A review identified **266 digital outcomes** from 21 diabetes trials, mostly as exploratory outcomes [Graña Possamai et. al, 2020].

Even for a specific summary measure, there many points of discretion which pose challenges for evidence synthesis going forward.

Need for standardisation: should we draw from existing frameworks, such as core outcome sets [Kirkham and Williamson 2022]?

Considerations for a Primary Analysis

Simulation study motivated by INOPulse Phase II trial [King et al, 2022], where the primary outcome was Time spent in Moderate-to-Vigorous Physical Activity.

We illustrate the impact on bias and standard error of the treatment effect estimator when:

- The measurement period is changed
- There are behavioural and seasonal effects
- There is missing outcomes

Eight methodological questions

Q1: What is the appropriate outcome granularity?

Q2: What is the appropriate measurement period?

Q3: How is the summary measure chosen?

Q4: How do we exploit the granularity of digital outcome measures to learn more about the treatment?

Q5: How do seasonal/behavioral variations affect design/analysis?

Q6: How are missing data defined and handled?

Q7: What are appropriate estimands for digital outcome measures?

Q8: Is evidence synthesis possible?

REVIEW Open Access

Unleashing the full potential of digital outcome measures in clinical trials: eight questions that need attention

Mia S. Tackney[1*], James R. Carpenter[2,3] and Sofía S. Villar[1]

References (selected)

Bakker et al, 2024. V3+: An extension to the v3 framework to ensure user-centricity and scalability of sensor-based digital health technologies.

European Medicines Agency. Draft Qualification Opinion for Stride velocity 95th centile as primary endpoint in studies in ambulatory Duchenne Muscular Dystrophy studies. 2023.

Graña Possamai C, Ravaud P, Ghosn L, Tran VT. Use of wearable biometric monitoring devices to measure outcomes in randomized clinical trials: a methodological systematic review. BMC Med. 2020;18(1):1–11. doi: 10.1186/s12916-020-01773-w

Izem, R., Zuber, E., Daizadeh, N. et al. Decentralized Clinical Trials: Scientific Considerations Through the Lens of the Estimand Framework. Ther Innov Regul Sci 58, 495–504 (2024). doi: 10.1007/s43441-024-00615-8

King, et al, 2022. A Phase-2 Exploratory Randomized Controlled Trial of INOpulse in Patients with Fibrotic Interstitial Lung Disease Requiring Oxygen. doi: 10.1513/AnnalsATS.202107-864OC
[Phase II study: NCT01457781, Phase III study: NCT03267108]

Lisi E, Abellan JJ. Statistical analysis of actigraphy data with generalised additive models. Pharm Stat. 2023. doi: 10.1002/pst.2350

Servais L, Yen K, Guridi M, Lukawy J, Vissière D, Strijbos P. Stride Velocity 95th Centile: insights into gaining regulatory qualification of the first wearable-derived digital endpoint for use in Duchenne muscular dystrophy trials. J Neuromuscul Dis. 2022;9(2):335–46. doi: 10.3233/JND-210743

U.S. Food and Drug Administration. *Digital Health Technologies for Remote Data Acquisition in Clinical Investigations*. December 2023. Guidance Document, Docket No. FDA-2021-D-1128. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations