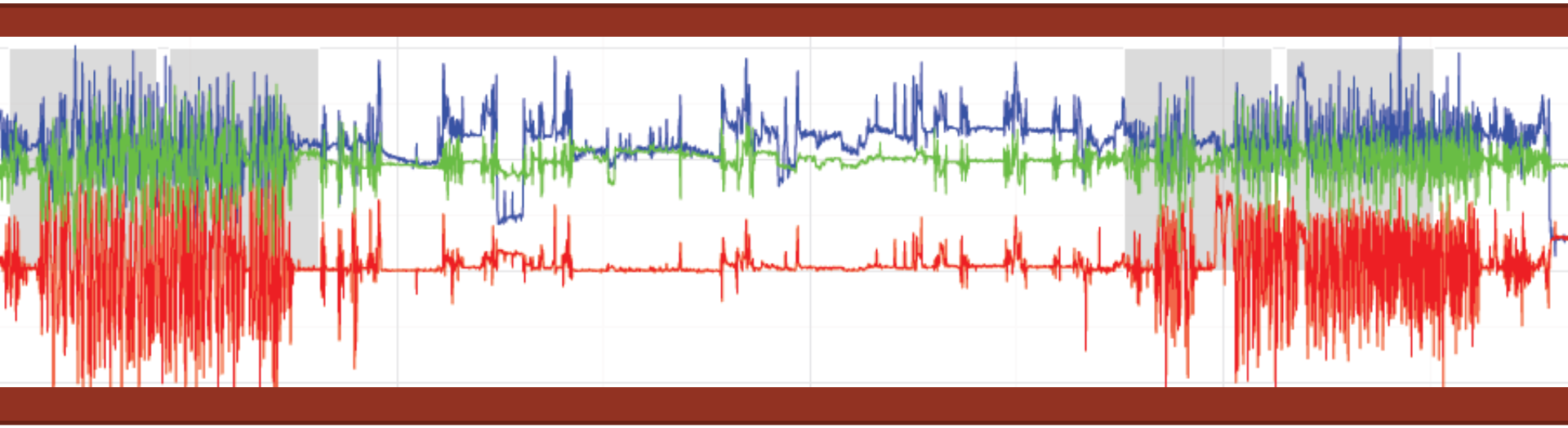
Information extraction and transparency in big data processing

Developing new endpoints for mobility is an important strategic aim for many groups both in industry and academia and the focus of a growing field. Bringing those novel endpoints to health authority acceptance for clinical decision making will require a concerted effort from this research community. This in turn will require openness and transparency; sharing data, methods and findings. Here we discuss challenges within the field to such an open approach and give examples of how they might be overcome.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016



Information extraction and transparency in big data processing

20150305

Ieuan Clay on behalf of the “ActiTeam”

NOVARTIS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

What do we mean by transparency?

- Transparency aims to make the process of research clearly visible to all readers.
  - including data collection, coding, and analysis
- Important for:
  - Peer-review publication process
  - Collaboration
  - Regulatory submission process

...any situation where reproducibility is important...

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Why transparency?

Perspective from Clinic/Industry

- Regulatory trail for the FDA
  - Reproduce any aspect of evidence supporting a submission
- Contributing to a growing field
  - Crowd sourcing
- Collaboration with TRIUM
- Cross divisional, evolving team

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Challenges to transparency?

http://ejournals.bc.edu/ojs/index.php/jtla/pages/view/transparency

- Protection of human subjects (e.g., extent to which data can be made available without infringing on the rights of subjects) – *can we even anonymize accelerometry?*
- Providing access to analytic tools (e.g., Matlab, SPSS, etc.) that are copyrighted – *should we even use them?*
- Providing access to tests and testing tools that are copyrighted or under development – *how far can we standardise QC in a diverse and evolving field?*
- Providing access to large data sets (e.g., TIMSS or NAEP) used as part of a research study – *should we develop open datasets, like kaggle (http://mayer.pro/t-SNE-Samsung)?*
- Developing community spirit and direction – *where is the consensus/synergy in such a diverse and developing field?*
  - ...the potential uses of accelerometry are so huge, expert guidance and direction is crucial (clear pull/push)
  - ...this opens the field to many contributors

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Transparent tools

- (Open) Programmatic frameworks
- Documentation frameworks and standards
- Version control and other good practice
- Data sharing platforms and formats
- Ontologies
- Good project management, e.g. Role transparency within diverse teams
- ...

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Contents

- Intro
  - Project
  - Technology & TRIUM collaboration
  - Team
- Focus on Analytical Challenges
  - Creating an environment for open collaboration
  - Information extraction from clinical tests
- Outlook

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Physical activity

WHO Fact sheet N°385 February 2014

- Physical activity is defined as any bodily movement produced by skeletal muscles that **requires energy expenditure**.
  - Including activities undertaken while working, playing, carrying out household chores, travelling, and engaging in recreational pursuits.
  - At all ages, the benefits of being physically active **outweigh potential harm**, for example through accidents.

Some physical activity is better than doing none.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Physical inactivity

- The fourth leading risk factor for global mortality and causes 6% of all deaths
- Appr. 3.2 million people die each year due to inactivity
- Physical inactivity is on the rise in many countries.
- 20% to 30% increased risk of death compared to people who engage in at least 30 minutes of moderate intensity physical activity on most days of the week.
- Physical inactivity is the main cause for approximately:
  - 21–25% of breast and colon cancers
  - 27% of diabetes
  - 30% of ischaemic heart disease

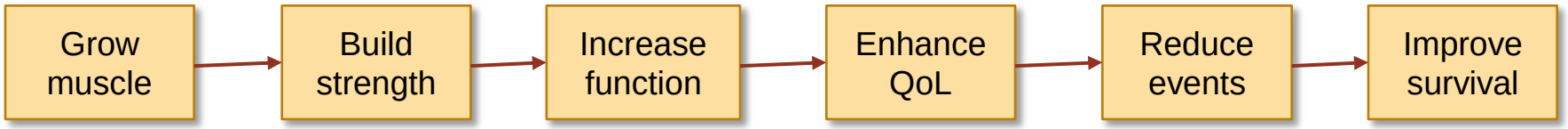
Vos et al. Lancet 2012; 380: 2163–96

2010

Disorder	Mean rank (95% UI)	% change (95% UI)
1 Low back pain	1·1 (1 to 2)	43 (34 to 53)
2 Major depressive disorder	1·9 (1 to 3)	37 (25 to 50)
3 Iron-deficiency anaemia	3·3 (2 to 6)	−1 (−3 to 2)
4 Neck pain	4·3 (3 to 7)	41 (28 to 55)
5 COPD	5·8 (3 to 10)	46 (32 to 62)
6 Other musculoskeletal disorders	5·9 (4 to 8)	45 (38 to 51)
7 Anxiety disorders	6·4 (4 to 9)	37 (25 to 50)
8 Migraine	8·9 (6 to 15)	40 (31 to 51)
9 Diabetes	9·1 (6 to 13)	68 (56 to 81)
10 Falls	10·1 (7 to 14)	46 (30 to 64)
11 Osteoarthritis	12·3 (9 to 17)	64 (50 to 79)
12 Drug use disorders	12·5 (9 to 16)	40 (27 to 54)
13 Hearing loss	13·5 (7 to 20)	29 (22 to 36)
14 Asthma	15·3 (10 to 20)	28 (21 to 34)
15 Alcohol use disorders	15·8 (12 to 21)	32 (16 to 50)
16 Schizophrenia	16·0 (9 to 22)	48 (37 to 60)
17 Road injury	16·1 (12 to 20)	30 (13 to 49)
18 Bipolar disorder	16·6 (9 to 23)	41 (31 to 51)
19 Dysthymia	18·6 (13 to 26)	41 (34 to 48)
20 Epilepsy	21·8 (18 to 27)	36 (27 to 47)
21 Ischaemic heart disease	21·9 (17 to 29)	48 (40 to 57)
22 Eczema	22·3 (16 to 35)	29 (19 to 39)
23 Diarrhoea	23·1 (19 to 28)	5 (−1 to 11)
24 Alzheimer's disease	25·9 (21 to 33)	80 (71 to 88)
25 BPH	26·3 (20 to 35)	84 (48 to 120)
26 Tuberculosis		
27 Neonatal encephalopathy*		

Communicable, maternal, neonatal, and nutritional disorders
Non-communicable diseases
Injuries

Challenges for study design

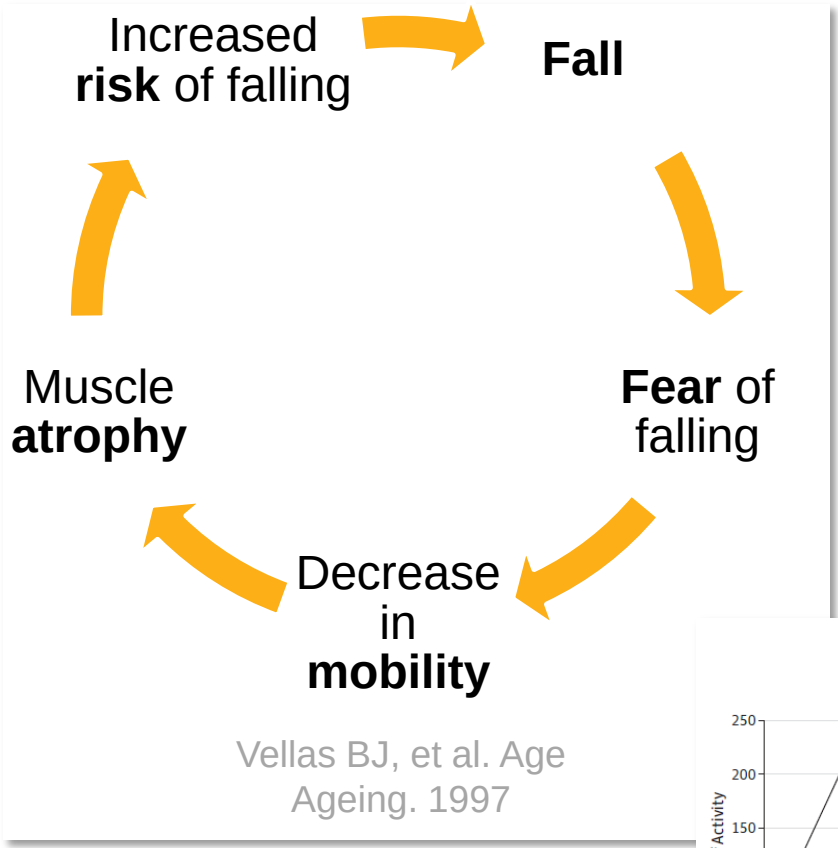


Establishing the chain of evidence

NOVARTIS

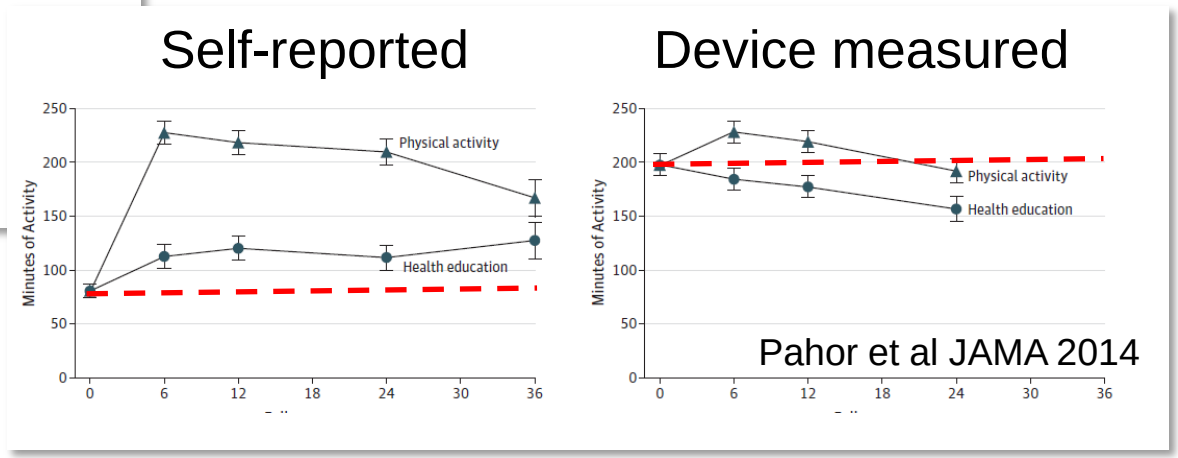
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Falls and deterioration of physical activity



Vellas BJ, et al. Age Ageing. 1997

Many pertinent research questions unanswered due to inaccessible endpoints...



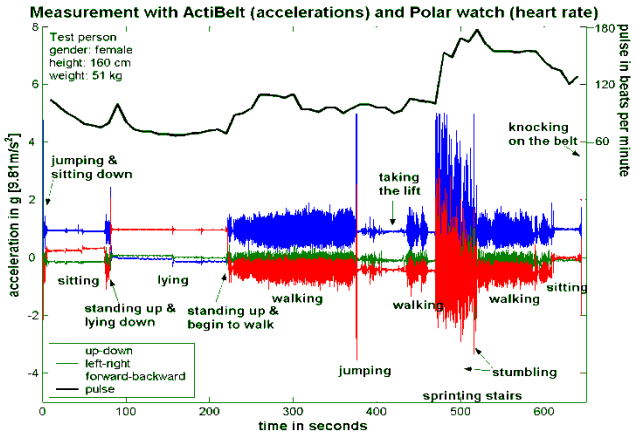
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

The challenge: Added value of accelerometry

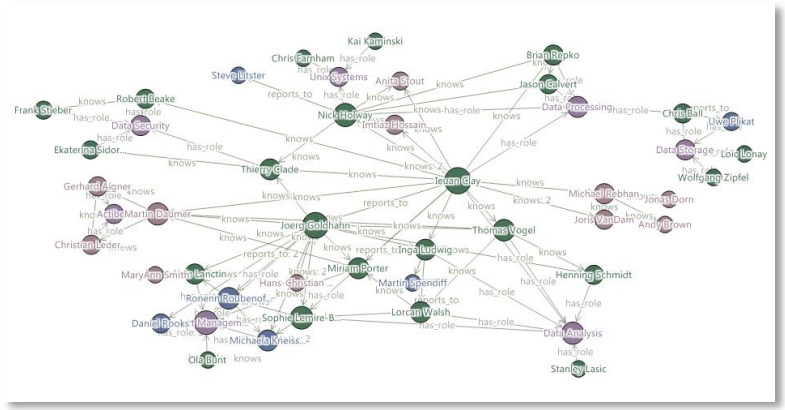
Objectively captures real-world movement patterns throughout the day

Accelerometry



- Define a new field of **research for drug development**
- Establish new, **robust clinical endpoints** together with the scientific community
- Set new **registerable standards for phase III trials**

ActiBelt



ActiTeam

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

External collaboration

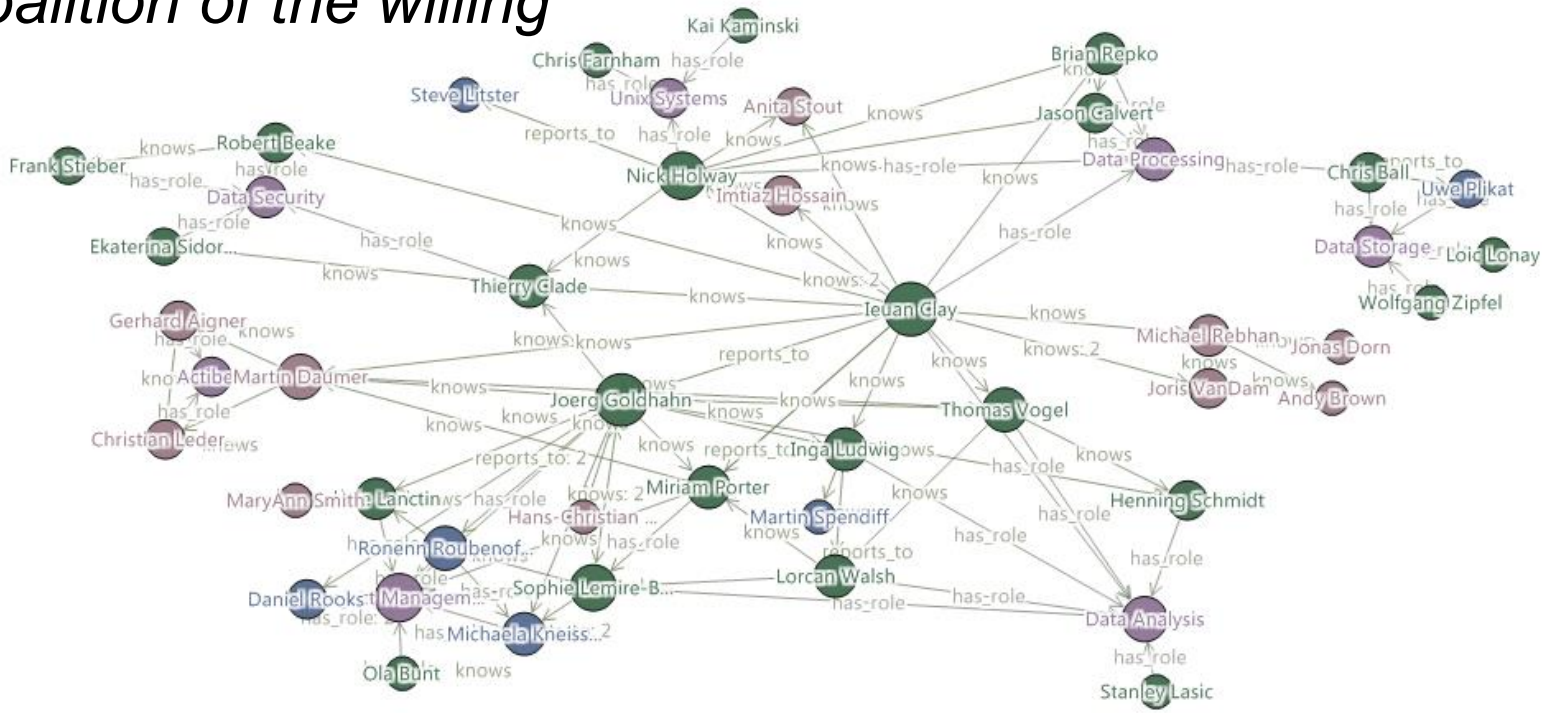
ActiBelt and TRIUM

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Internal working group

ActiTeam

- *"Coalition of the willing"*



- Jörg Goldhahn, Daniel Rooks, Sophie Lemire-Brachat, Valerie Lanctin, Miriam Porter, Thomas Vogel, Henning Schmidt, Inga Ludwig, Lorcan Walsh, Thierry Clade, Nick Holway, Ieuan Clay
- Collaborators at TRIUM / The Human Motion Institue
- Collaborators across NIBR

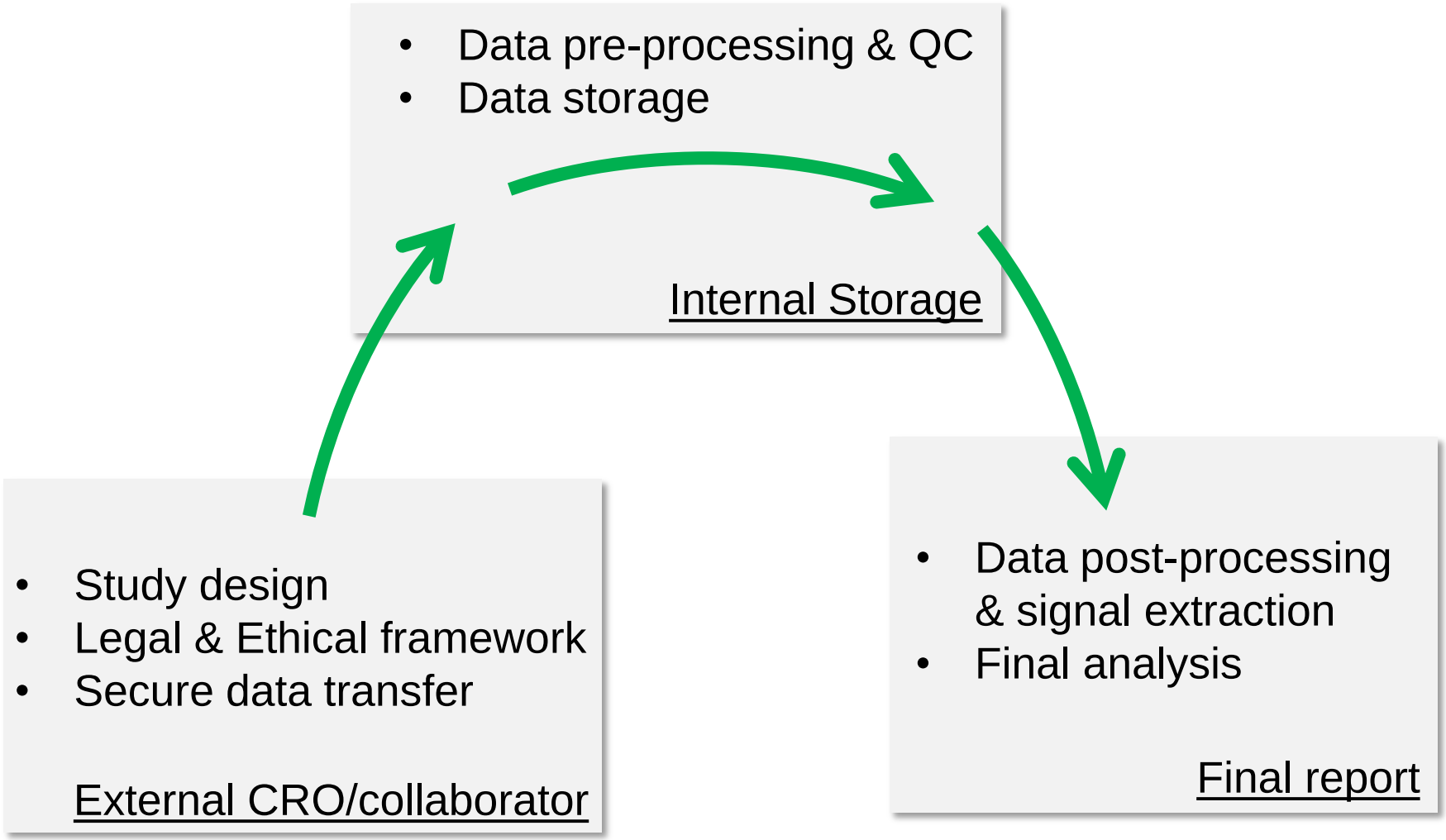
Challenges

- FDA/regulatory approach
- Translational strategic approach
- Collaborative challenges
- Competence challenges
  - Technical:
    - Data transfer & volumes
    - Data storage/access/security
  - Analytic challenges:
    - Data QC
    - real world streaming data
    - Variability & complexity
    - etc

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Data Flow: overview



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

Data Flow I: External to Novartis

Documentation and coordination

Lorcan Walsh &
Miriam Porter

1 /Month visit to clinic & swap of device

Trium manage the sensor and forward data

trium

NOVARTIS

Clinical Protocol

1. Patient self-selects to wear device continuously between clinic visits
2. In clinic, standard clinical tests performed
    1. 6MWT
    2. SPPB
    3. Sway
    4. etc

Device deployment

Device preparation

Trium Recycle Actibelt

- Data Transfer Specifications
- Secure Hard Drive / File transfer

Device collection

Data extraction and analysis

- Accelerometry data
- Gait/Mobility parameters

NOVARTIS

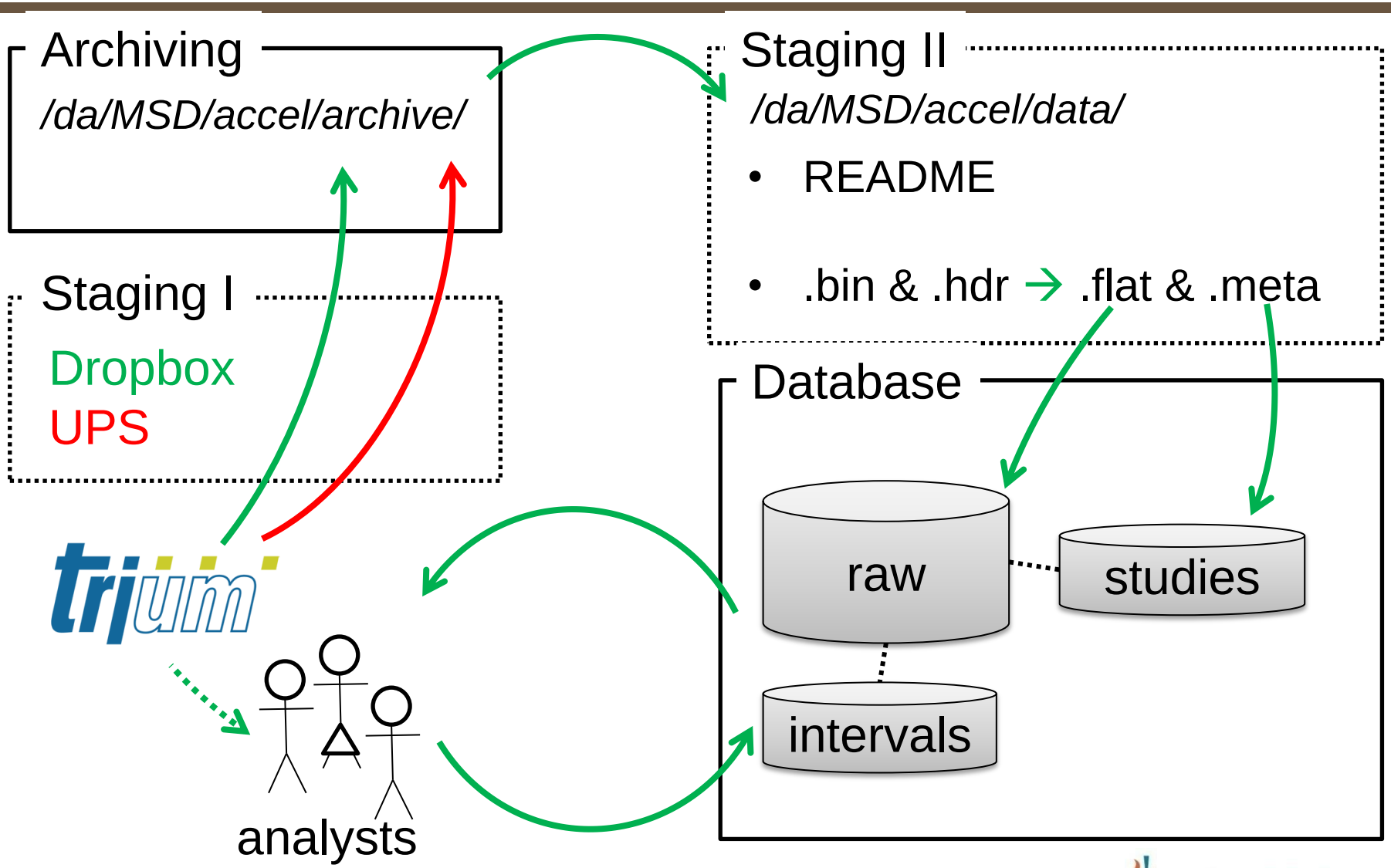
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Data Flow II: Processing & storage

Programmatic framework

transient

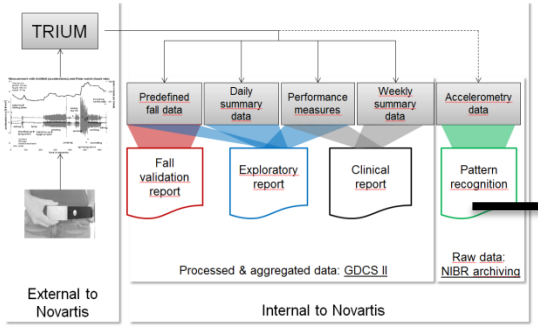
permanent



NOVARTIS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Data Flow III: Archiving of exploratory data



NIBR Basel archiving

File system

Data on unix root:

- **/da/MSD/accel/**
  - transfer_archive
    - 20141013
    - 20141120
  - data
    - trium_pilot
      - README.txt
      - actibelt_videos
      - exercise
      - fall data
      - long term recordings
      - short term recordings

Subversion

- docs
- analysis
- examples
- engineer

Vertica DB

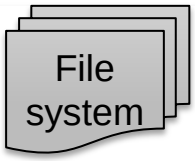
- msd_accel

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Data overview: Database structure

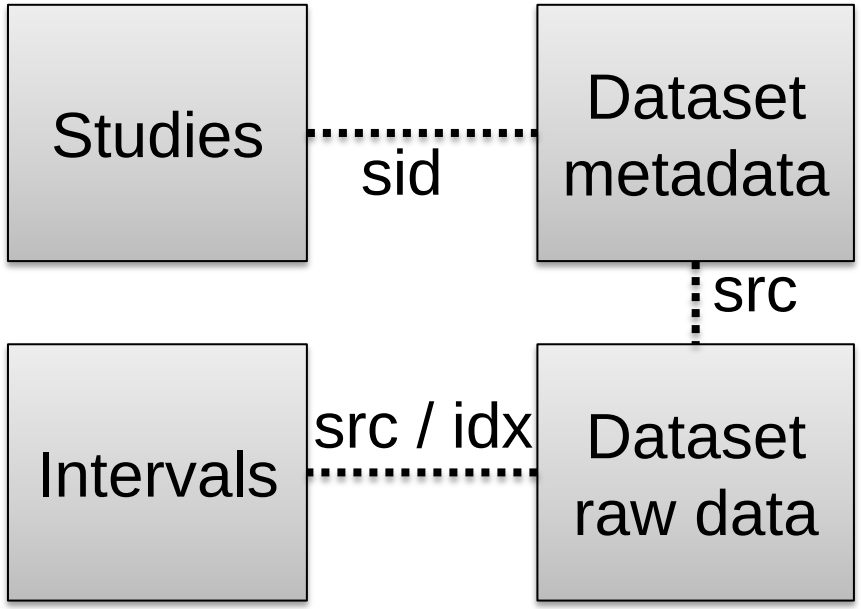
Data storage

file	BIN	FLAT
54256DDB.BIN	117M	1.4G
MDfall20110901.bin	1.1M	15M
1408001464.bin	150M	1.5G
1408396914.bin	512M	6.1G
1408396625.bin	62M	748M
1408555060.bin	449M	5.3G
53318EC1.BIN	1.4G	22G
54368376.BIN	39M	594M
1408001464a.bin	150M	1.5G



64 hours of data
~2.9Gb binary data
~40Gb 'flat' data
=>15x (de-)compression

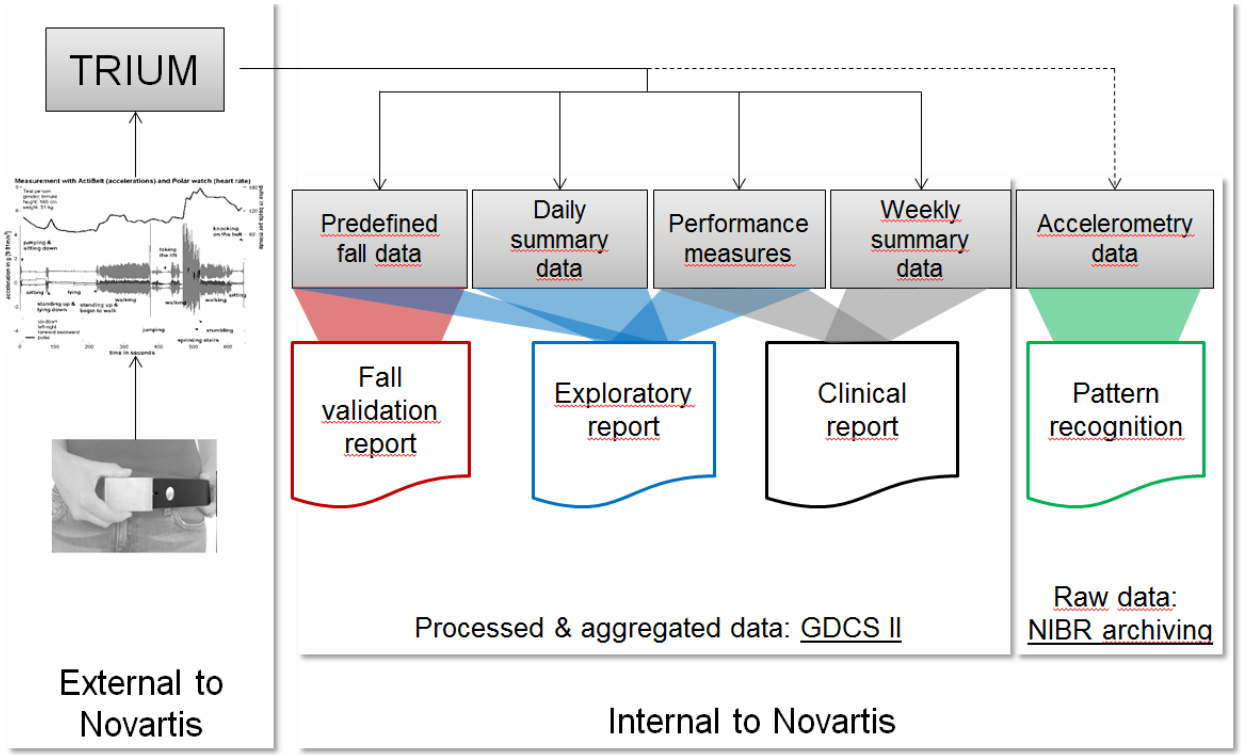
Vertica DB • msd_accel



5.5Gb on disk
⇒ 2x compression from binary
⇒ 0.09Gb per hour data
⇒ **2Gb per day per patient**

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Data Flow IV: Division within Novartis



- Aims ("Added Value")

Through deeper understanding of the physical activity of patients, we aim to deliver:

- Better patient selection
- Simpler trials and reduced patient burden
- Better outcome measures

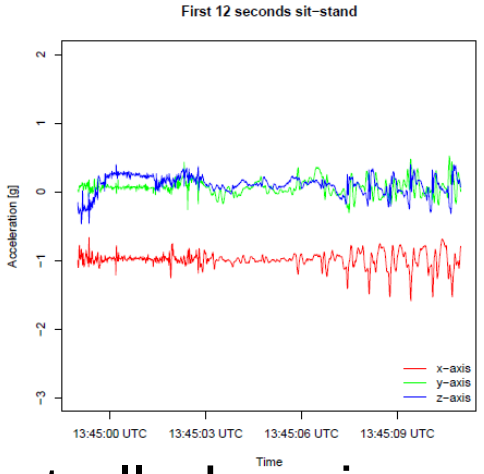
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Analytic challenges

- Aims ("Added Value")

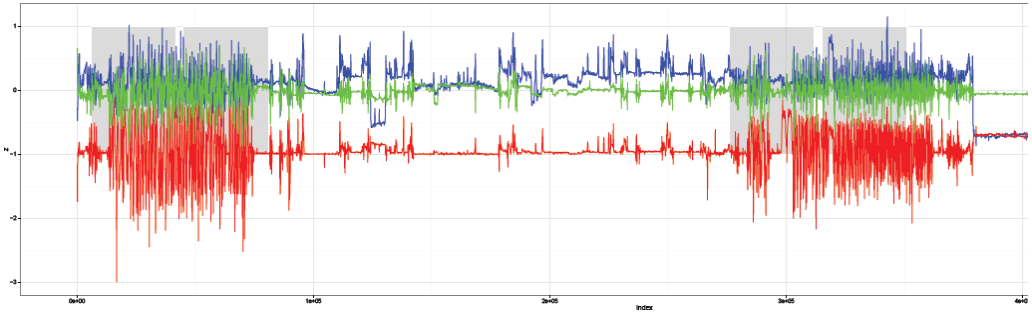
Through deeper understanding of the physical activity of patients, we aim to deliver:

- Better patient selection
- Simpler trials and reduced patient burden
- Better outcome measures



Controlled environment (sit-stand) clinical test

- Challenges:
  - Data QC (What is normal?)
  - Real-world, streaming data (Variability & complexity)
  - etc



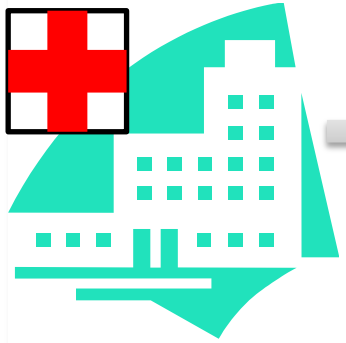
Real-World environment

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

NOVARTIS

6-minute walk test

Current gold-standard

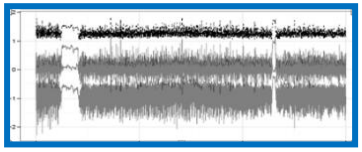


6-minute walk test

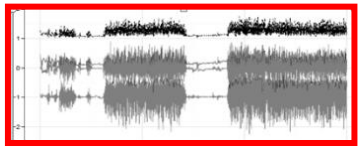
One outcome variable measured by observer

- **Total Distance walked**

TRIUM ActiBelt
actibelt.com



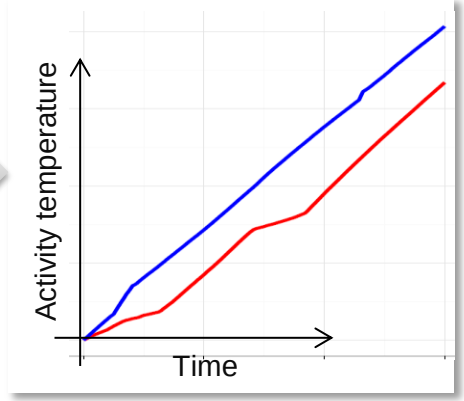
Patient 1



Patient 2

Total Distance walked

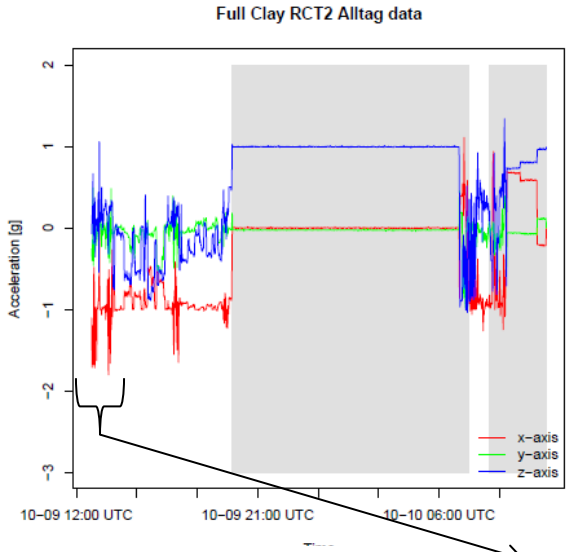
\+ *Fatigue, Gait, Balance, etc*



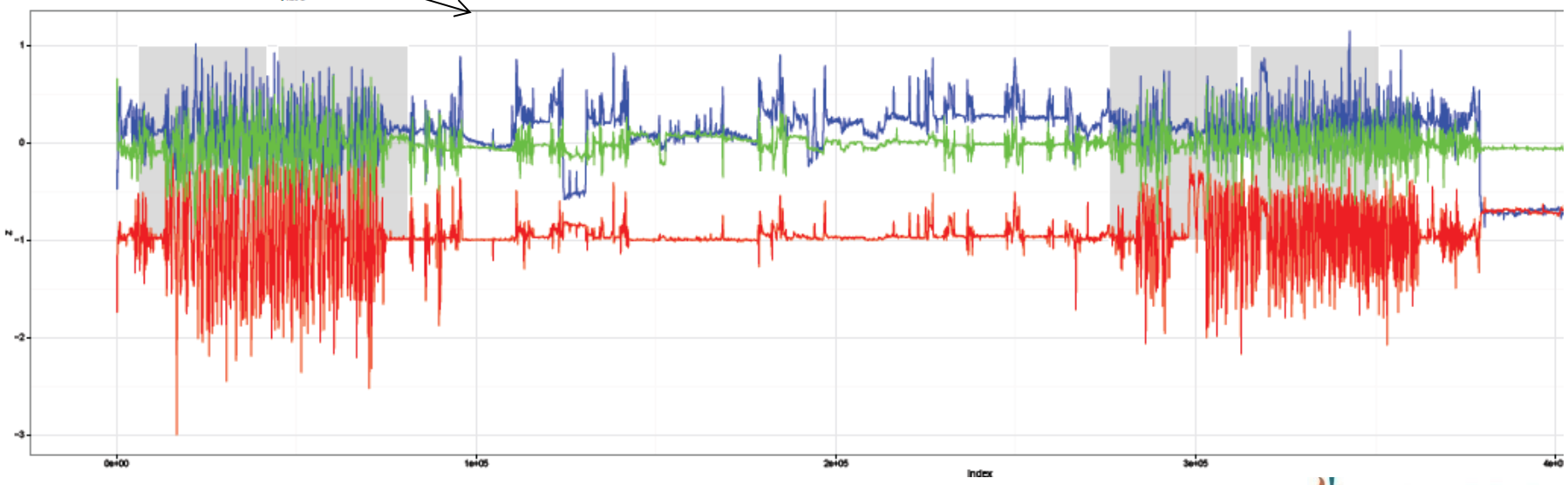
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

6-minute walk test II

Exercise (Clay "6 minute walk")



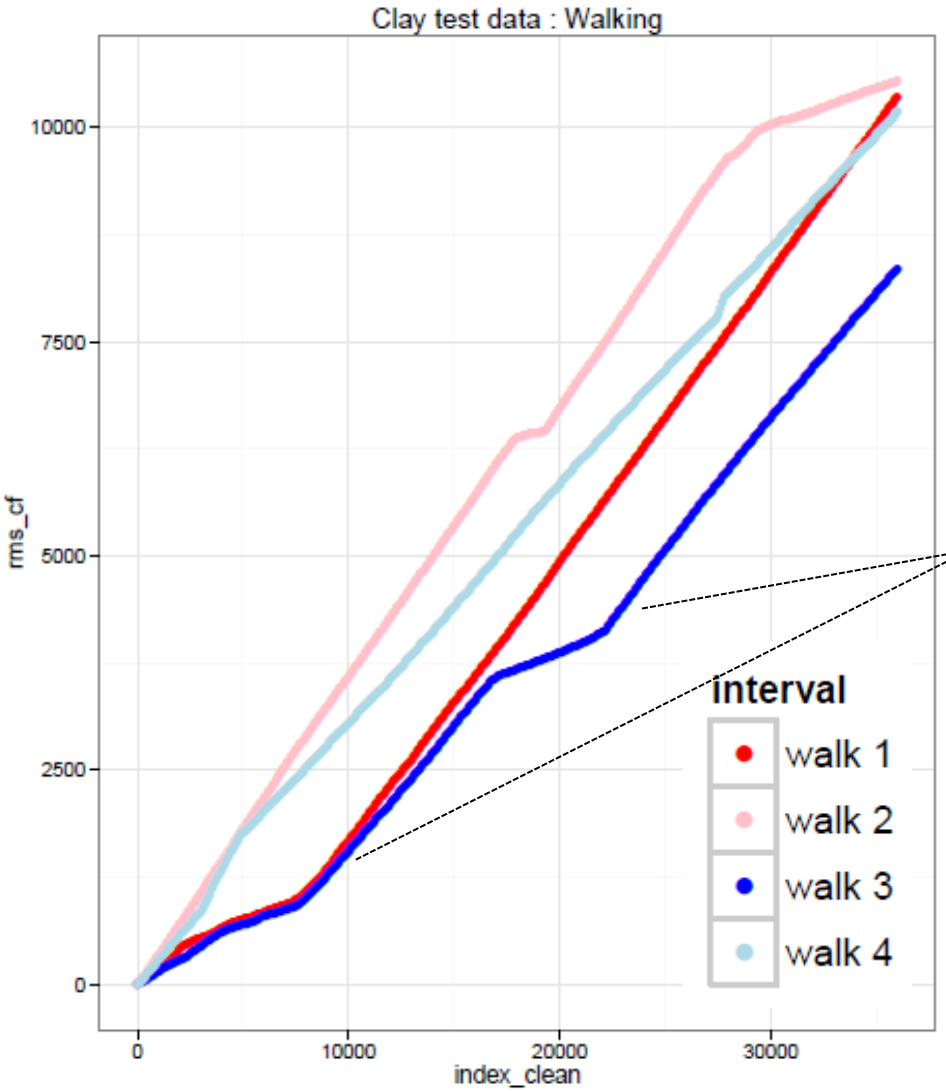
- 4 x 6 minute walking periods extracted from data
  - Grey boxes (below)
- Cumulative activity* over time
  - * = RMSq over all axes

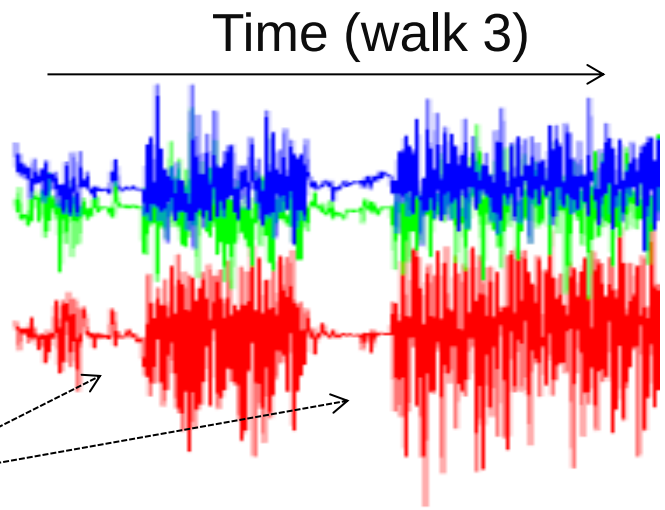


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

6-minute walk test III

Exercise (Clay "6 minute walk")

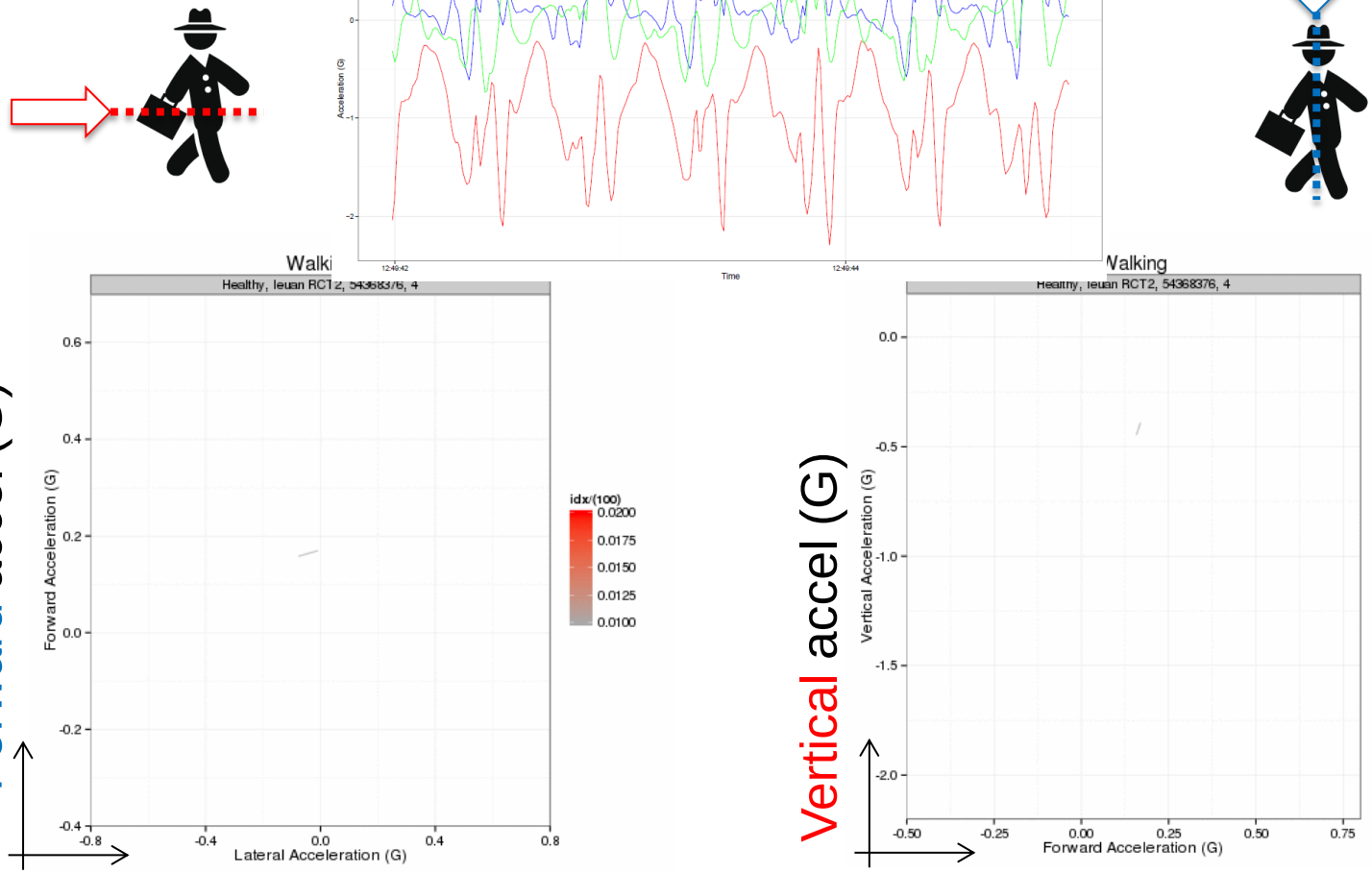




- Cumulative plots could be used to track different velocities and find "pauses" in 6 minute test

NOVARTIS

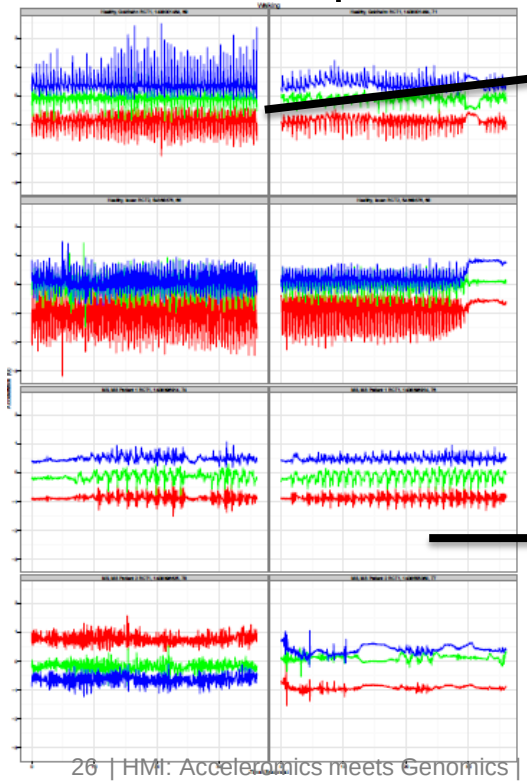
Walking assymetry (gait 'quality')

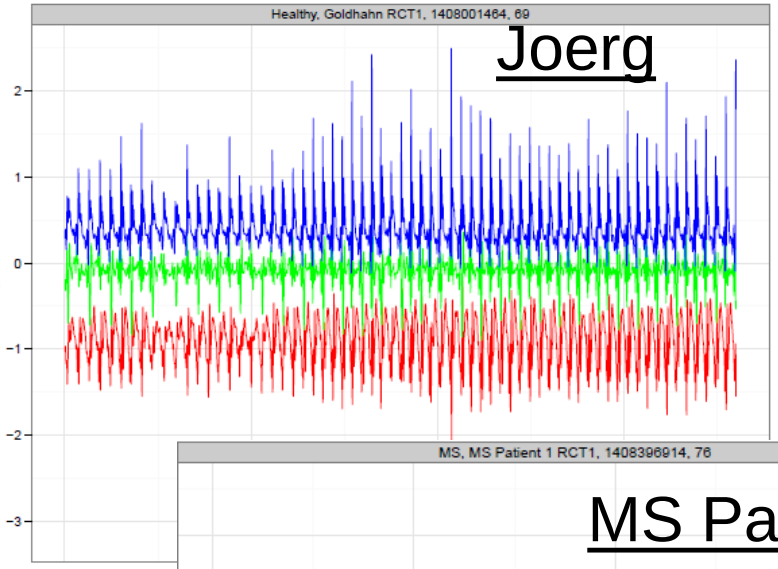


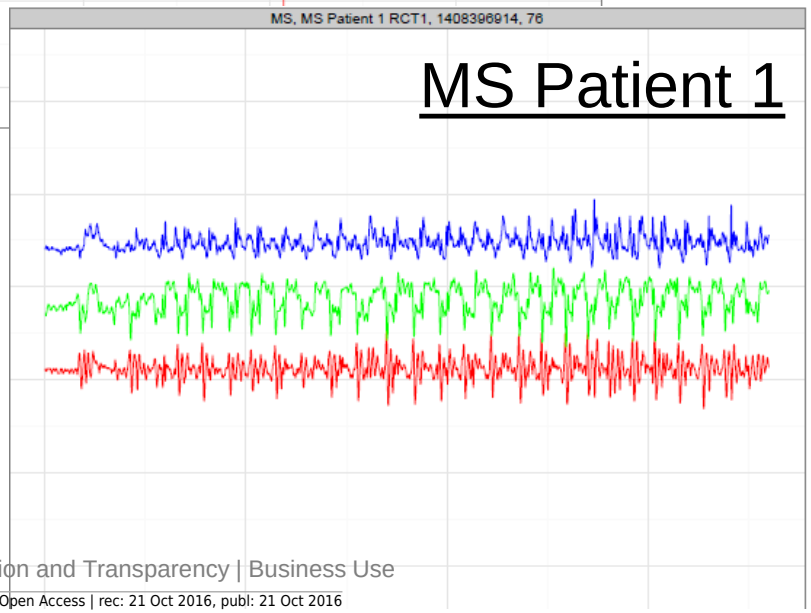
Gait II

Real life or 6MW

- 2 x ~6 minute ‘walking’ periods hand-curated from long-term observation data
  - 2 x healthy (Ieuan & Joerg)
  - 2 x MS patients



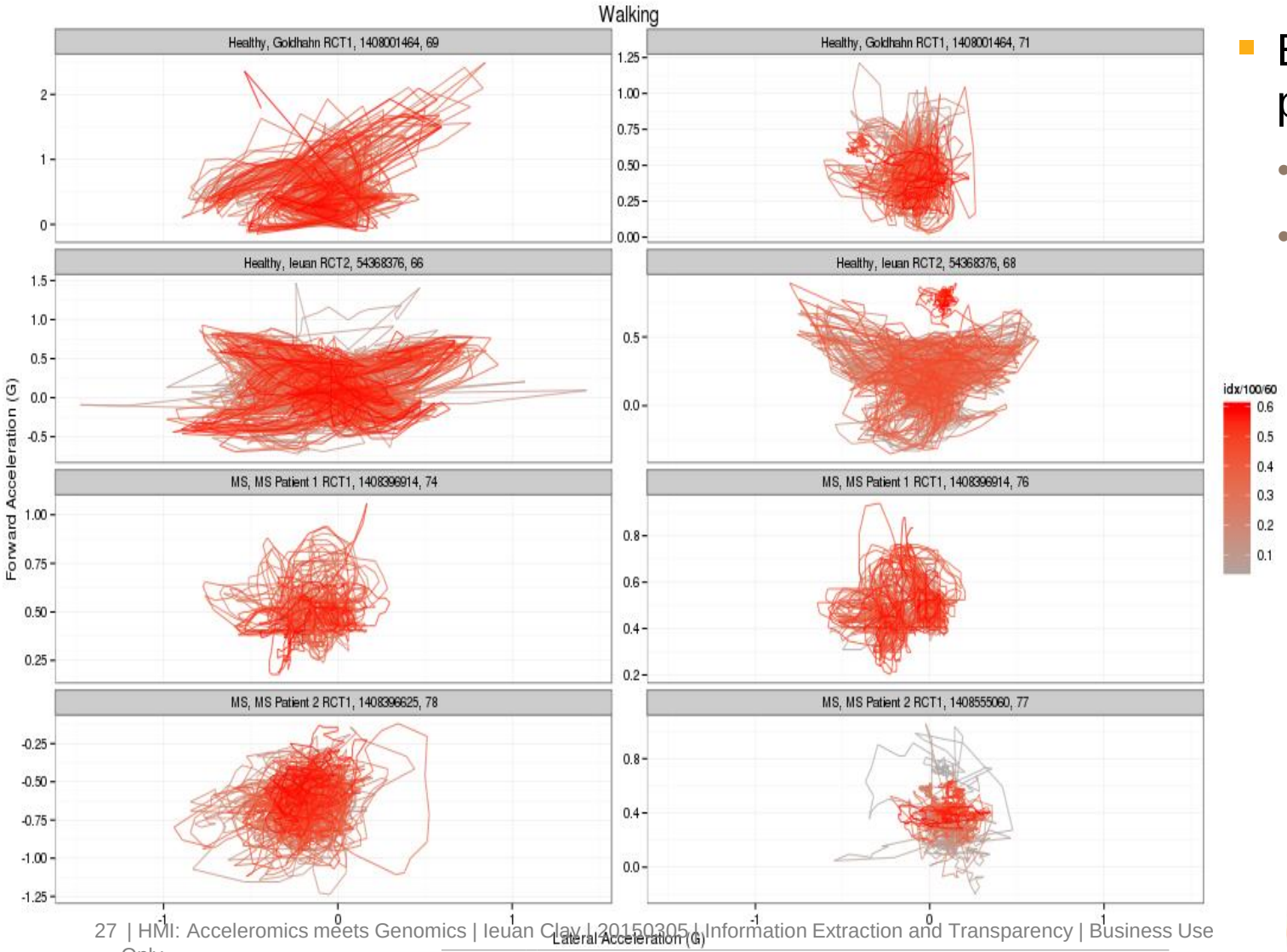




PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Gait II

Variation within and between individuals



- Each row = 1 person
  - Top 2 = healthy
  - Bottom 2 = MS

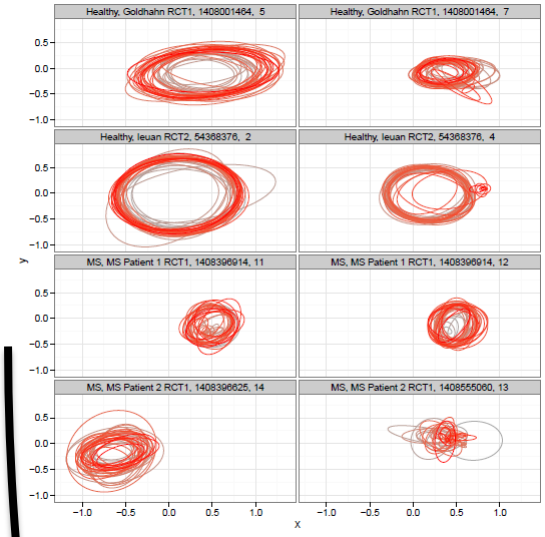
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

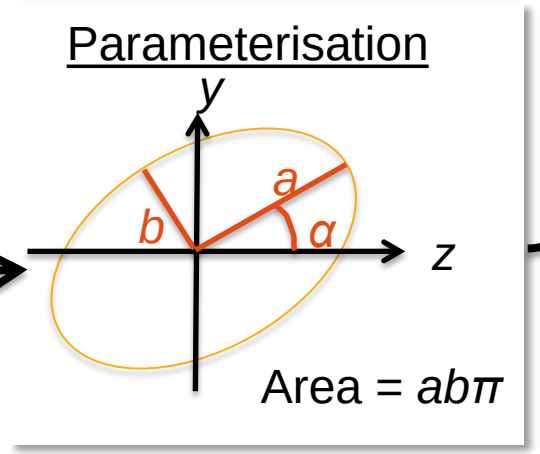
Gait II

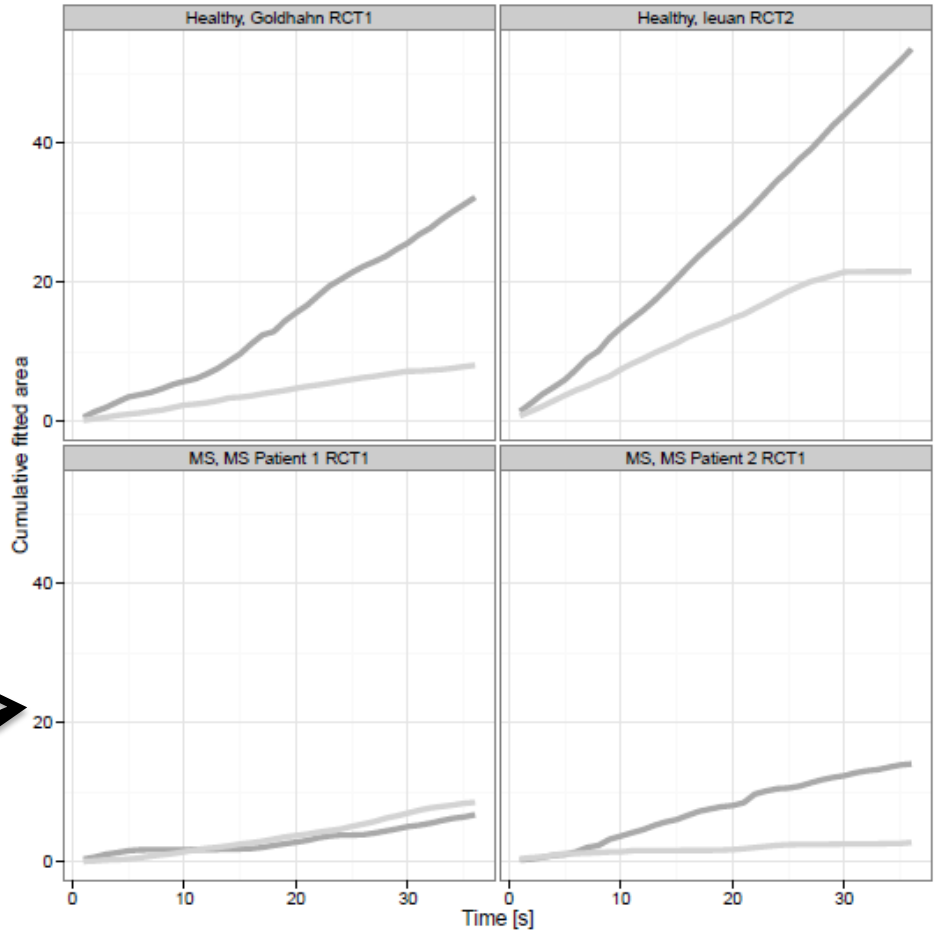
Variation within and between individuals

Each panel = 1 person

- Top 2 = healthy
- Bottom 2 = MS







Inga Ludwig
NOVARTIS

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Outlook

- The possibilities for information extraction are huge
- Documentation and other best practices are crucial to enabling analysis
- The tools exist
- Will bigger data = bigger problems (scaling infrastructure)?

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016

Discussion

Questions, suggestions, feedback!

- Jörg Goldhahn, Daniel Rooks, Sophie Lemire-Brachat, Valerie Lanctin, Miriam Porter, Thomas Vogel, Henning Schmidt, Inga Ludwig, Lorcan Walsh, Thierry Clade, Nick Holway, Ieuan Clay
- Collaborators at TRIUM / The Human Motion Institue
- Collaborators across NIBR & Dev

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2546v1 | CC BY 4.0 Open Access | rec: 21 Oct 2016, publ: 21 Oct 2016