A Framework for Digital Data Quality Assessment in Digital Biomarker Research

Hui Zhang, Regan Giesting, Leah Miller, Guangchen Ruan, Neel Patel, Ju Ji, Tianran Zhang and Yi Lin Yang
(zhang_hui@lilly.com)

DH R&D, Digital Health, Eli Lilly & Co.

ALLDATA 2023, April, Venice, Italy

OUTLINE

- Data Challenges in Connected Clinical Trials
- Wearable Device Data for Digital Measure Development
- Digital Data Quality
  - Non-wear Detection
  - Data Model for Quality
  - Data Quality Outcome Reporting Tool
- Deriving Digital Measures from High Quality Sensor Data
- Conclusion & Future Work

CCT & KEY CHALLENGES

WHAT IS CCT ?	Connected Clinical Trials use digital technologies to enable real time connection with consented patients, to capture clinically meaningful signals via connected devices/apps and wearable sensors to enable optimized clinical trial design, execution, and data. CCT will develop science and technology that serve as the foundation to precision and personalized medicine and enable direct connections with patients.

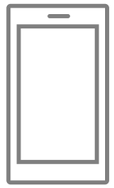
- Unique challenges in connected clinical trials implementing wearable devices
  - **Handling big data**
    - High frequency wearable sensor data, patient reported outcomes, etc.
  - **Delivering good data**
    - Data collected in free-living environment
    - Scanning invalid data points, removing noise, identifying useful portions of data for dBM development
    - Accurate ground truth data is also one essential element of good data
  - **Reusing, reproducing and publishing digital assets**
    - A digital data **cloud**
    - Accessing data from anywhere
    - Slicing, aggregating, and sharing digital data

TYPICAL DIGITAL DATA SETS

Common data collected in connected clinical trials:

Raw sensor data and derived features (Time series data)

- 3-axis accelerometer / gyroscope
- Peripheral thermometer
- EDA (Electrodermal Activity)
- PPG (Photoplethysmogram)
- ECG (Electrocardiogram)



Participant reported data (ePRO)

- Labels
- Events

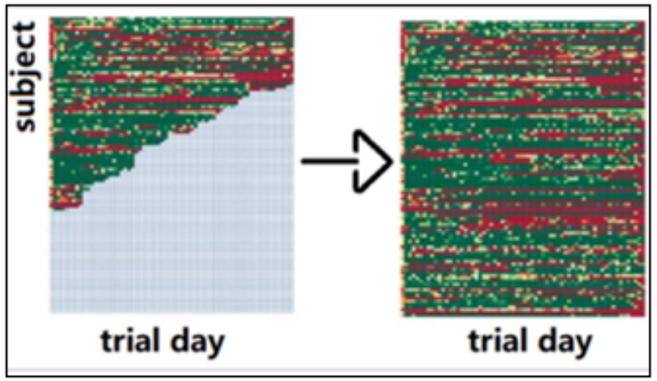
Derived Compliance and Quality Data

DATA QUALITY ASSESSMENT OVERVIEW

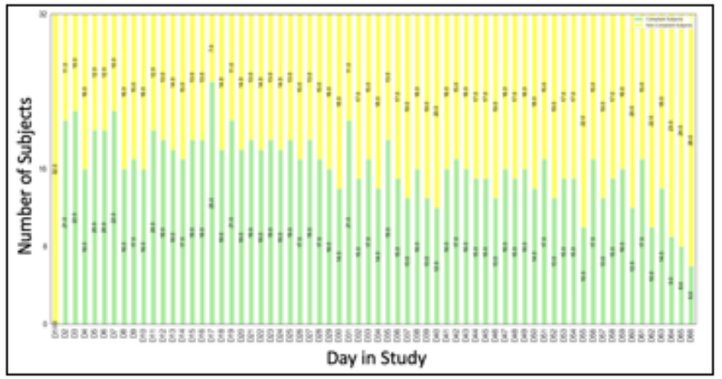
- **Data transfer agreement (DTA)**
  - *valid range*
  - *invalid/error code*
  - *sampling rate*
- **Device wearing compliance criteria**
  e.g., *20 hours wearing / day*

Feature	Freq	Min	Max	Invalid
HR	0.25	30	200	0
BA	0.25	0	100	255
BAV	0.125	0	2500	65535
EV	None	0	65535	None
PO	1	0	12	5
RE	0.25	4	42	0
RSSI	1	-128	126	127
RSVD_01	1	0	65534	65535

- **Monitoring for conformance to expectations and correcting non-conformance**
  - *signal conformance to DTA*
  - *wearing compliance*



- **Quality assessment and reporting**
  - *Cleansing and standardization*
  - *Digital measures generation*
  - *Identifying analyzable data points*
  - *Quality assessment and reporting*



SIGNAL DATA QUALITY METRICS

Sampling Frequency: Preconfigured average number of samples obtained in one second, or the resolution of the data in dBM.

Valid Range:

- **Numerical Values**: minimum and maximum values that can be measured (ex: sensor signals)
- **Enumerated Variables** : a list of predefined categorical values (ex: sleep classifications)

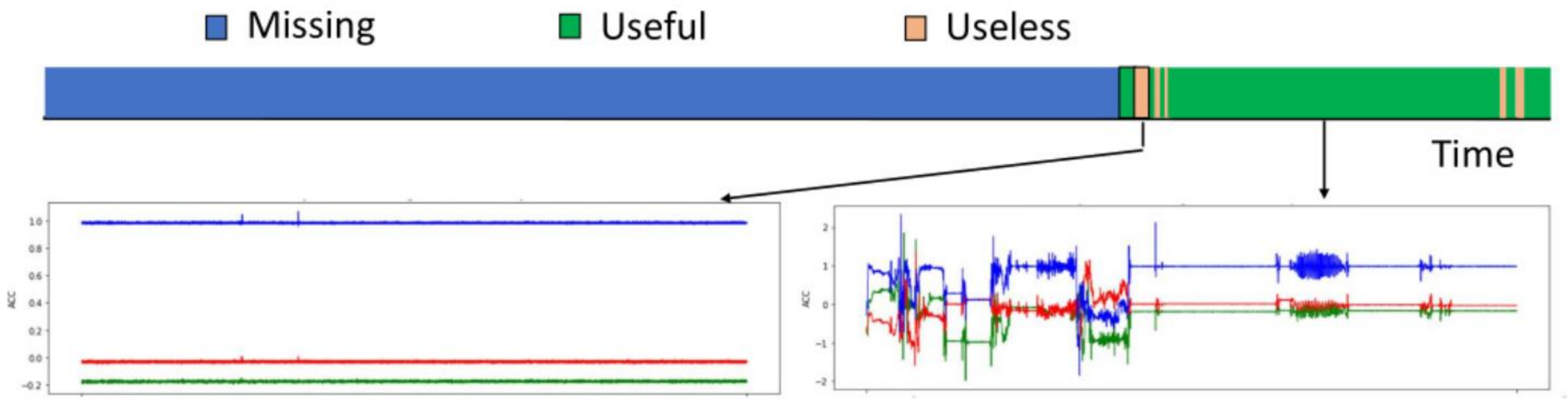
Invalid Value / Error Code: Specific invalid values to indicate different statuses of malfunctioning in devices, in addition to the provided valid ranges.

	Channel	Description	Units	Min Value	Max Value	Invalid Value	Sampling Frequency (Hz)
Numerical	$accel_x$	Accelerometer X Vector	gravity/1024	-32768	32767	None	50
	$accel_y$	Accelerometer Y Vector	gravity/1024	-32768	32767	None	50
	$accel_z$	Accelerometer Z Vector	gravity/1024	-32768	32767	None	50
	ec	ECG signal	μV	-10000	10000	32767	125
	st	Step count	Steps	0	65535	None	1
	hr	Heart rate	beats/min	30	200	0	0.25
	re	Respiration rate	beats/min	4	42	0	0.25
Enumerated	po	Posture • Laying Down = 0 • Standing = 2 • Walking = 3 • Running = 4 • Unknown = 5 • Leaning = 11	Enum	0	11	5	1

SIGNAL DATA QUALITY ASSESSMENT

- Valid signals can mix with invalid signals in the data collection and how they differ when plotted.

- **Correctly** vs **incorrectly worn** → **useful** data vs **useless** data



- Filtering out invalid values with a valid value range results in valid data coverage, or the coverage of valid data points.

NON-WEAR DETECTION

- Detect moments when a device is not correctly worn.
- Utilize *Biobank* to detect non-wear:
  - Estimate accelerometer non-wear based on the standard deviation and the value range of the raw data from *each* accelerometer axis.
  - 30-second epoch classification in larger windows
  - Stationary periods used to define whether a window is stationary or not

DATA QUALITY MODEL

Epoch Level

Generated form Biobank's 30 second epoch classification

Subj	Timestamp	Non-wear
1002	2021-09-15 19:15:00	false
...	...	...
1005	2021-10-18 09:45:30	true

Hourly Level

Keep only correctly worn epochs, resulting in hourly data coverage derived from number of compliant minutes

Subj	Date	Hr	Cvge. $(min.)$
1002	2021-09-15	19	45
...	...	...	...
1005	2021-10-18	09	60

Daily / Intraday Window Level

Summary coverage derived from applying time filters to hourly quality table

Subj	Date	Cvge. $(min.)$	Window
1002	2021-09-15	1440	pa_daily
...	...	...	...
1005	2021-10-18	720	sleep_night

DATA QUALITY MODEL

Extended Quality with External Mappings

Additional data quality mapping comes as additional data becomes available, including visit or site information for patients.

Added mappings allow for thresholds/filters to be applied for data aggregation.

Ex: Minimum of 20 hours of data for a day to be qualified as a valid day

Site	Subj.	Date	Trial Day Index	Visit	Cvge. $(min.)$	Window
101	1002	2021-09-15	1	0 $(PreTreatment)$	1440	pa_daily
...	...	...	...	...	...	...
103	1005	2021-10-18	32	4	720	sleep_night

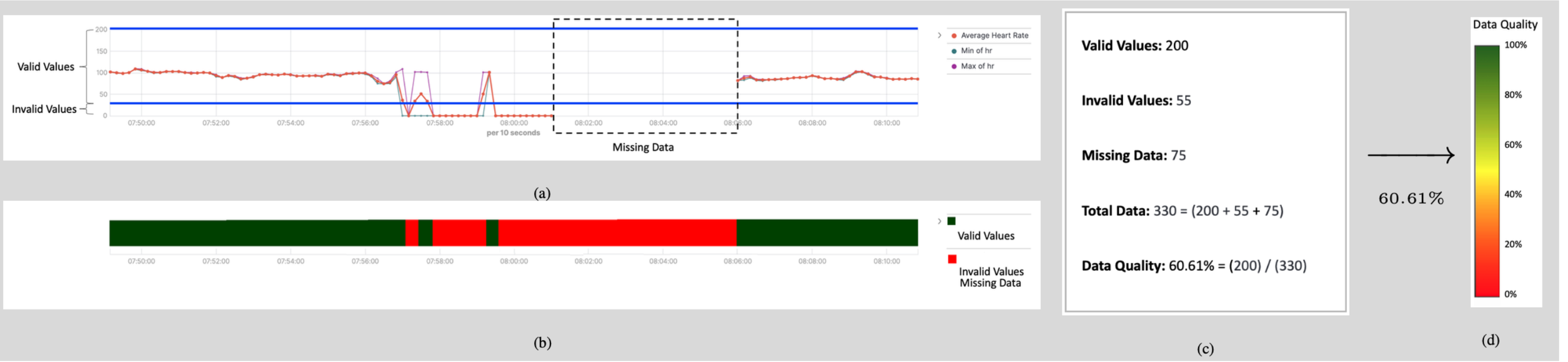
DATA VISUALIZATION – MISSING DATA & OUTLIERS

Identifying outliers in Heart Rate data:

Heart Rate valid values range between 30 to 200

Visualize the data to easily identify:

a) Missing data
b) Sensor signal quality
c) Numeric representation of the data
d) Interpreting data quality on a color scale



DATA VISUALIZATION – DATA QUALITY MAP

Sensor signals can be visualized in order to examine data quality patterns & analysis:

Visualize the data to easily identify:

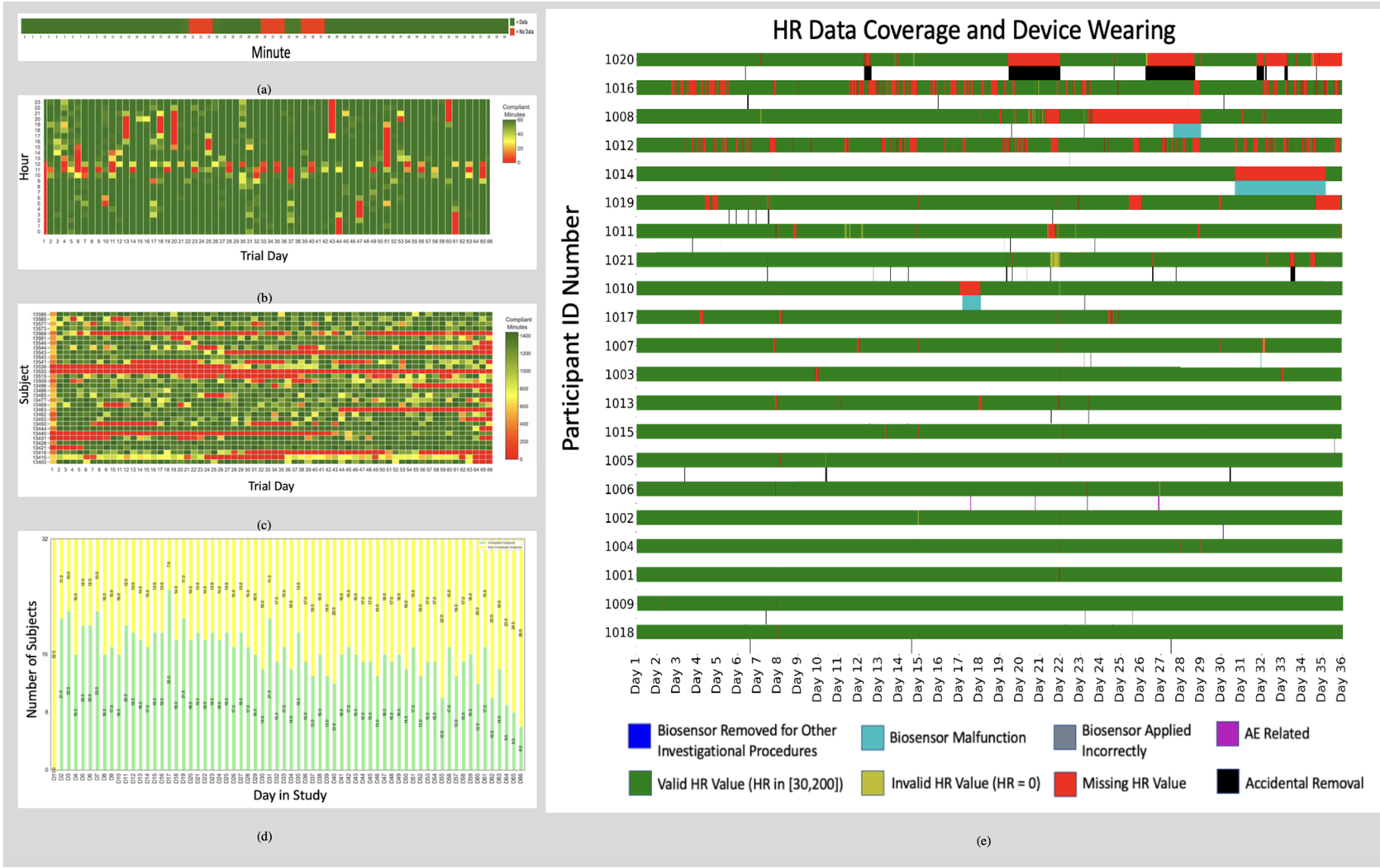
a) Minute by Minute Quality
   Ex: Individual patient, one day

b) Hour by Hour Quality
   Ex: Individual patient, all days

c) Day by Day Quality
   Ex: Population level

d) Compliant Days Throughout Trial
   Ex: Individual patient, all days

e) Identify & Align Data Issues
   Ex: Derive potential wearing patterns and/or device issues



DATA VISUALIZATION – COMPLIANCE REPORTING

Study Level Compliance

CPMP ISA Compliance Report

ISA Information

ISA: BP02

Total Patients: 131

Date: 09/22/2022

Compliance Table

Compliance is calculated for a patient as % of days with >= 20 hours of sensor data. Each patient is expected to have 66 days. Below, a patient is categorized as compliant if they have at least 50% of those 66 days are meet this criteria.

Compliance	# Patients	% of Patients
Patients with >= 50% of days compliant	75	57.25 %
Patients with < 50% of days compliant	56	42.75 %

Site Based Compliance

Compliance is % of days the patient has completed thus far with >= 20 hours of sensor data.

Site	# Patients	Average Compliance
148	3	90.91 %
138	3	85.86 %
102	3	85.35 %
123	2	84.85 %
128	2	68.18 %
132	2	65.15 %
134	1	65.15 %
110	7	60.82 %
106	8	60.23 %
122	21	59.52 %
103	7	59.09 %
114	12	58.33 %
108	5	57.88 %
107	19	55.58 %
119	2	49.24 %
117	2	47.73 %
147	2	45.45 %
149	1	40.91 %
111	1	39.39 %
118	6	31.82 %
142	19	20.65 %
109	3	9.6 %

(a)

Site Level Compliance

CPMP Site Compliance Report

Site Information

Site: 106

Completed: 36 Patients

In Progress : 4 Patients

Date: 09/20/2022

Compliance Table for Completed Patients

Compliance is calculated for a patient as % of days with >= 20 hours of sensor data. Each patient is expected to have 66 days. Below, a patient is categorized as compliant if they have at least 50% of those 66 days are meet this criteria.

Compliance	# Patients	% of Patients
Patients with >= 50% of days compliant	23	63.89 %
Patients with < 50% of days compliant	13	36.11 %

In Progress Patients

Compliance is % of days the patient has completed thus far with >= 20 hours of sensor data.

Subject	ISA	Current Visit	Compliance	Issue Identified
13220	NP02	V6	43.75 %	
13767	NP02	V6	52.94 %	
13817	NP02	V5	87.5 %	
13868	NP02	V4	60 %	

(b)

Patient Level Compliance

CPMP Patient Compliance Report

Patient Information

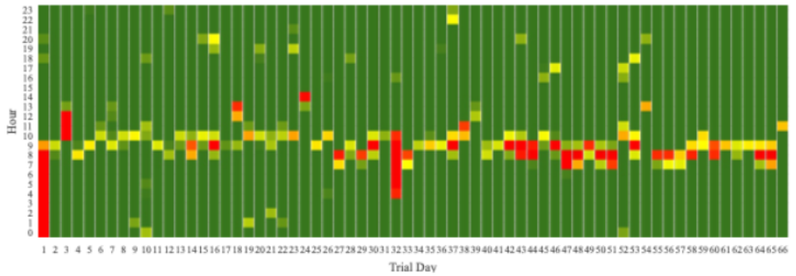
Subject: 12227
ISA: NP03
Site: 122
Date: 09/20/2022

Compliance Table

A compliant day is classified as a day having >= 20 hours of sensor data.

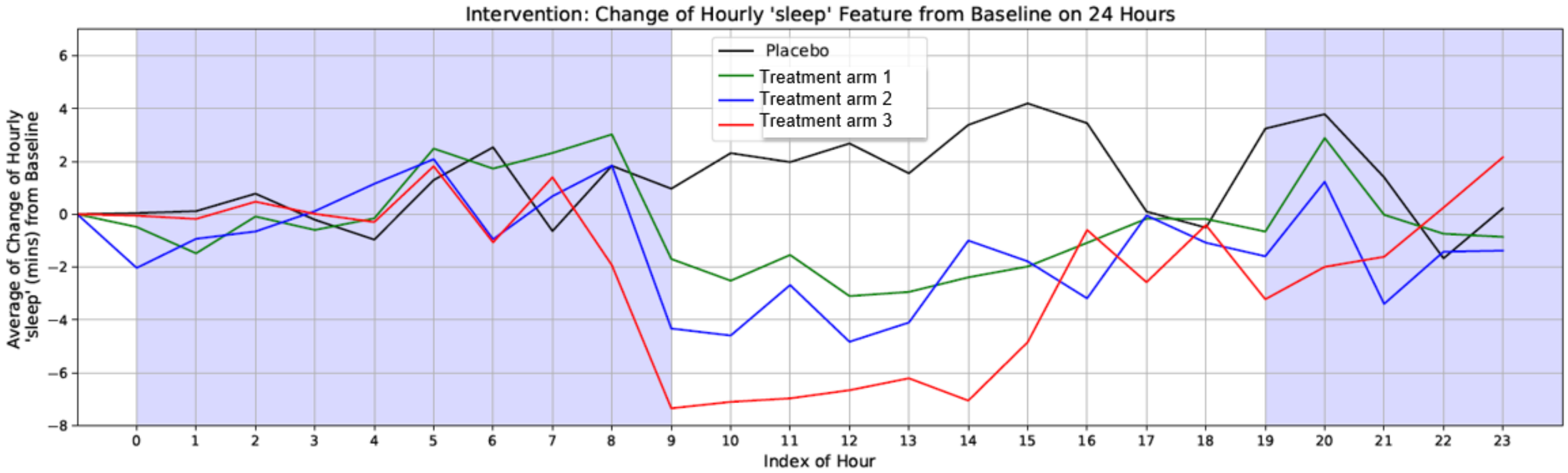
Time Period	Date Range	Number of Compliant Days	Compliance Percentage
PRE-TREATMENT	07/11/2021-07/25/2021	10	66.67%
VISIT 4	07/26/2021-08/09/2021	15	100%
VISIT 5	08/10/2021-08/22/2021	12	92.31%
VISIT 6	08/23/2021-09/06/2021	15	100%
VISIT 7	09/07/2021-09/19/2021	13	100%

Hourly Compliance



(c)

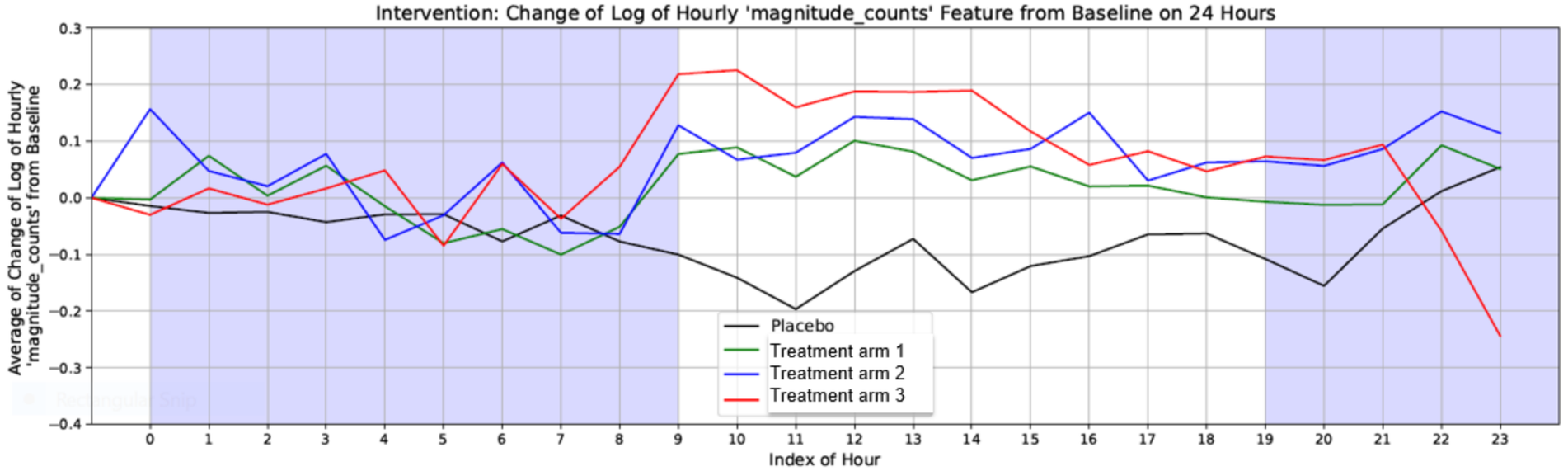
SLEEP DIGITAL MEASURE FROM HIGH QUALITY SENSOR DATA



Derived hourly sleep digital measure demonstrates drug efficacy in treatment phase

- Three treatment arms all show **decreased** daytime sleep change from baseline (more active) compared to placebo cohort

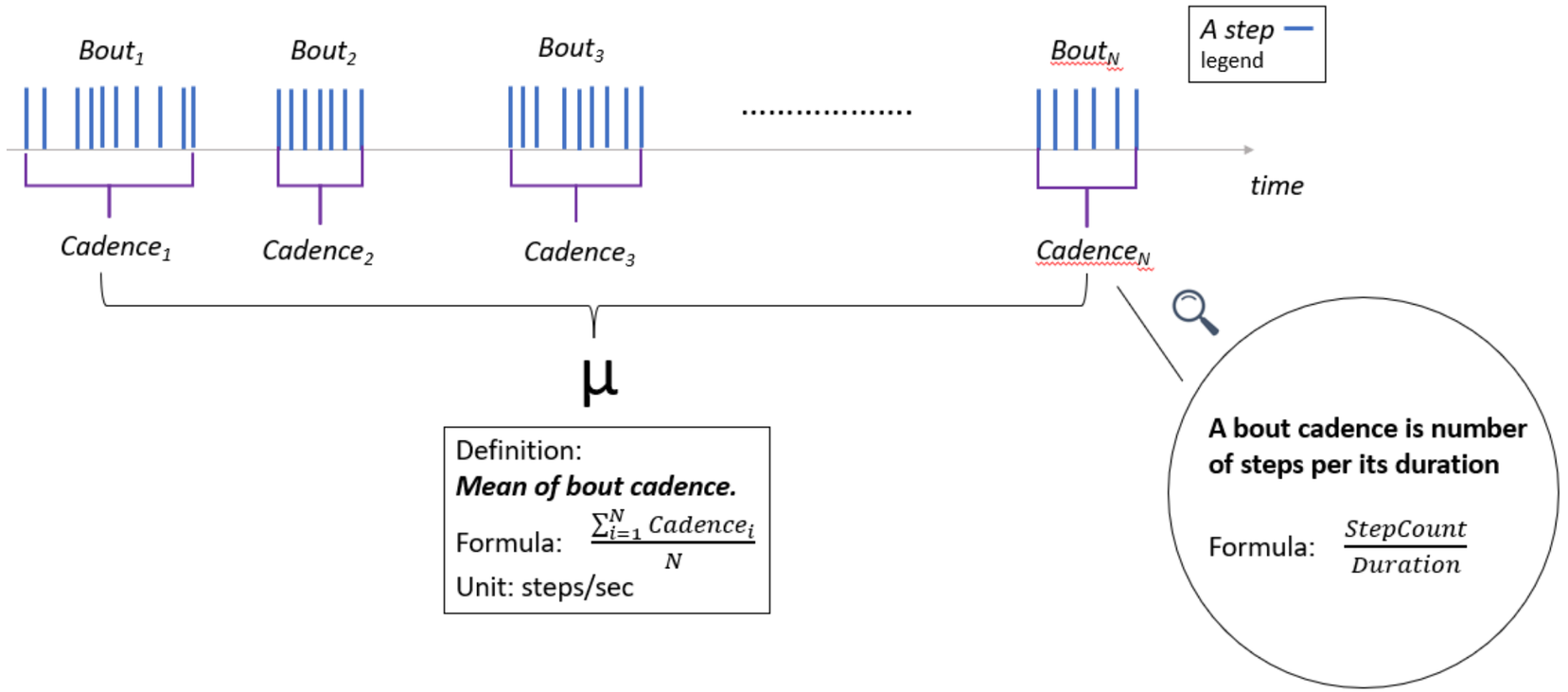
PHYSICAL ACTIVITY DIGITAL MEASURE FROM HIGH QUALITY SENSOR DATA



Derived hourly physical activity (magnitude counts) digital measure demonstrates drug efficacy in treatment phase

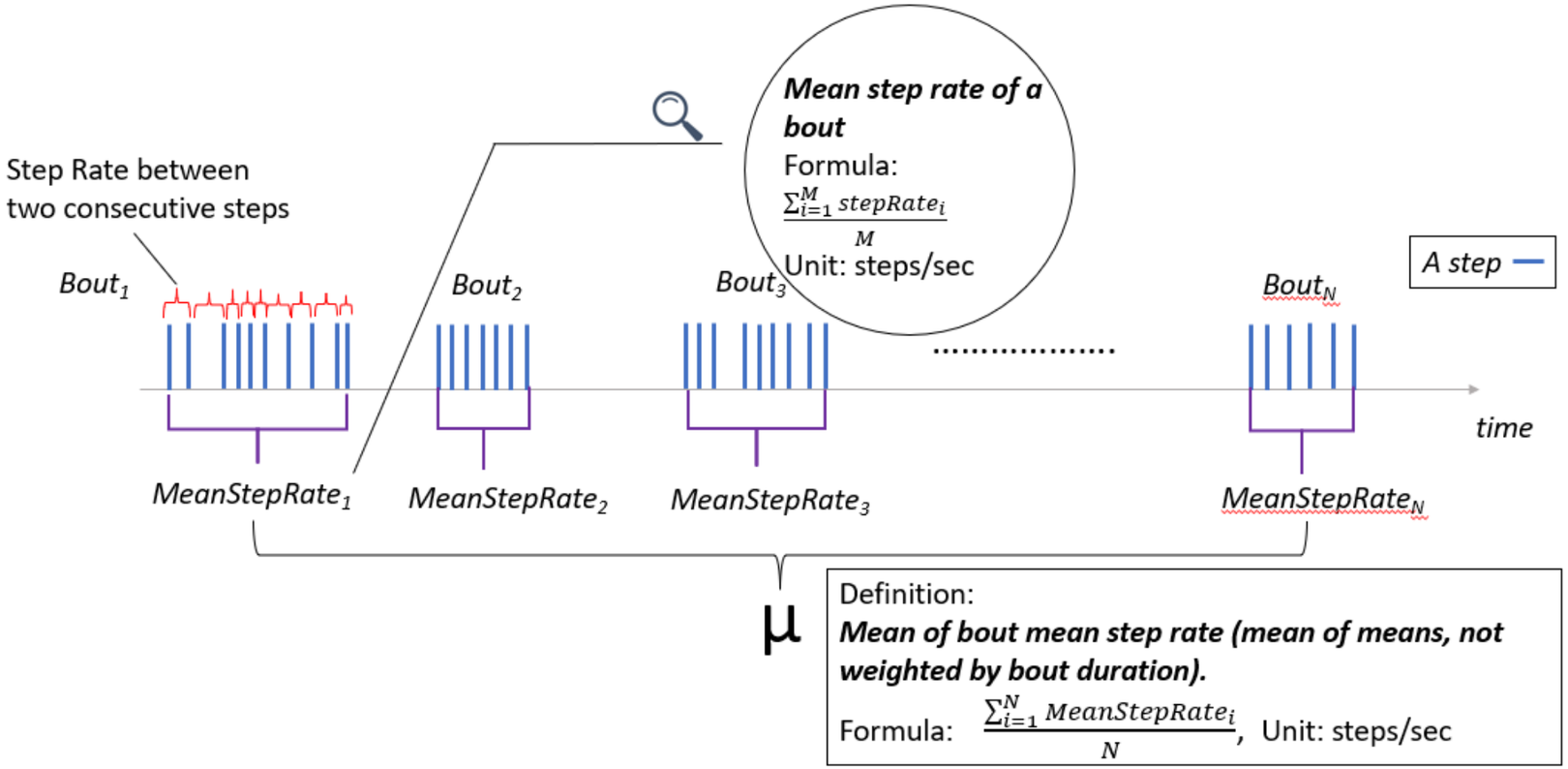
- Three treatment arms all show **increased** daytime physical activity change from baseline (more active) compared to placebo cohort

CADENCE FEATURE FROM DERIVED STEPS



- Individual steps can be derived from raw accelerometer data
- Cadence feature can in turn be derived by averaging across all bouts' cadence

GAIT RATE FEATURE FROM DERIVED STEPS



- Individual steps can be derived from raw accelerometer data
- For each bout we can derive mean step rate
- Gait rate feature is then derived from averaging all bouts' mean step rate

CONCLUSION & FUTURE WORK

- Continue to define and implement the fundamentals of data quality into the digital data quality framework and platform
- Generate automated compliance reports, customizable visualizations, and real-time quality metrics
- Future directions include the use of visual mining and data mining technologies to help identify data quality in a novel way to facilitate data quality assessment

Thank You!