A Framework for Digital Data Quality Assessment in Digital Biomarker Research

Hui Zhang[1], Regan Giesting[1], Leah Miller[1], Guangchen Ruan[1], Neel Patel[1], Ju Ji[2],
Tianran Zhang[1,3], Yi Lin Yang[1,4]
[1]Digital Health Office, Eli Lilly & Company, Indianapolis, Indiana, USA
[2]Advanced Analytics and Data Science, Eli Lilly & Company, Indianapolis, Indiana, USA
[3]Department of Computer Science, Brown University, Rhode Island, USA
[4]Carmel High School, Carmel, Indiana, USA
email: {zhang_hui, rgiesting, miller_leah, ruan_guangchen, patel_neel_k, ji_ju yang_yi_lin}@lilly.com,
tzhang96@cs.brown.edu
all authors contributed equally

***Abstract*—Digital Health Technology (DHT) utilizes a combination of computing platforms, connectivity, software, and sensors for healthcare-related uses. Today, these technologies collect complex digital data from participants in clinical investigations, including wearable sensor signals and electronic Patient-Reported Outcomes (ePRO)s. These collected data are used to develop digital biomarkers (dBMs), which can act as health outcomes indicators for diagnosing and monitoring disease state and life quality. One essential step towards realizing the full potential of these complex digital data is to define the fundamental principles and methods to demonstrate sufficient data quality and fidelity needed for the research. This paper aims to develop a digital data quality assessment framework across the complete data life cycle in dBM research, including data quality metrics and methods to derive, visualize, and report digital data quality. Aggregating and reporting digital data quality is often challenging and error-prone. We developed a data quality assessment and reporting tool that defines data compliance criteria and views automatically generated quality reports at different levels in a consumable fashion. Combining all these methods helps to establish our digital data quality assessment framework to facilitate dBM research.**

***Keywords—digital health technology*; *connected clinical trial*; *sensor data*; *data quality assessment*; *data visualization*; *digital biomarker*.**

I. Introduction

Digital biomarkers (dBMs) are patient-generated physiological and behavioral measures collected through connected digital devices. The collected data are then used to explain, influence, or predict an individual's health-related outcomes (see [1]). While the development of dMBs invests heavily in advanced analytics, effective results depend on trusted and understood data collected from digital devices. An established data quality assessment framework is thus needed to define the expectation of data, monitor the data for conformance to expectations throughout the trials, and report various measures to assess the data quality (see, e.g., [2]). Establishing a meaningful data quality function will help reduce the risk throughout the dMB research activities and ultimately ensures the success criteria are met.

Today, we use DHT (see, e.g., [3]) to collect some of the most complex digital data from patients for dBM research. There has been an overall need for better data understanding and easier access to quality and trusted digital data to support operational and analytical activities in the research. Establishing a data quality assessment framework and building tools to facilitate the assessment is an emerging industry capability, and some unique challenges for this class of data quality strategy include:

- **Complexity of digital data** — We collect some of the most complex digital data in the dBM context, including sensor signals from wearables, patient-reported outcomes from hand-held devices, and labels and annotations processed and used as ground truth information for algorithm development and model building. Handling these data could be a big data problem. For example, with a sampling frequency of 50Hz, over 4 *million* 3-axial data points are collected from an accelerometer sensor for a single day to understand a patient's daily activities. Similar sensor data streams include, e.g., continuously collected photoplethysmography (PPG) and electrocardiogram (ECG) signals from trial participants.
- **Full-spectrum quality expectations** — Defining quality expectations for digital data and monitoring their conformance to expectations are full-spectrum in the data life cycle. For example, given that data can be collected in a free living environment, scanning the invalid values and noises in wearable sensor signals is often the first profiling step. Identifying the wearable sensor signal's useable (wear-compliant) portions is also a leading data quality function. The ultimate answer to the digital data quality question is the extent our digital data satisfies the specific dBM analysis requirement.
- **Aggregation and reporting** — Generating various measures to assess digital data quality is not trivial. For example, aggregating compliance information from signal level to the number of analyzable digital measures at the visit and study levels can often be tedious and error-prone. Equally challenging is to report data quality in an efficient and effective means across the data life-cycle.

Our task in this paper is to present a data quality assessment

Copyright (c) IARIA, 2023.   ISBN: 978-1-68558-041-4

framework and demonstrate a reporting platform to facilitate dBM research. The paper starts with an overview of the typical categories of digital data that our research is concerned with, then focuses on the metrics we use to profile digital data quality and later for aggregation at different levels. We also demonstrate how we put together all functions into a data quality reporting platform to support the work of all data quality assessment functions. Then, we elaborate on how the data quality reporting platform associates data stewards and quality analysts with particular study data, allowing them to run processes via interactive workflows and pull out consumable data quality reports in a central location.

This paper is organized as follows. Section II presents the related work. We present our digital data quality assessment framework and platform in Section III and Section IV, respectively. In Section V, we showcase digital endpoints and, finally, we conclude the paper in Section VI.

II. Related Work

Developing dBM requires conducting studies in a lab or free-living settings to collect raw sensor data, often with appropriate labels and annotations (*e.g.*, reported patient outcomes). Collection and analysis of wearable sensor data, together with other digital data sets, has thus become an emerging capability needed in dBM development. Industry players have begun exploring cost-effective and purpose-built solutions in the past few years. For example, the Medidata sensor cloud [4] is used to manage wearable sensor and digital health technology data for clinical trials. The Koneksa platform [5] provides support to improve compliance monitoring and patient engagement, and other representative efforts to store and deliver raw or processed data from devices in trials, including Evidation [6] and DHDP [7]. Furthermore, good data is more important than big data in dBM development. Given that data are collected in a free-living environment, noise in wearable sensor signals is inherent. To make sensor data useful, we need to monitor the quality and eventually standardize and process them to support dBM discovery, as digital data quality is of fundamental importance to developing algorithms for new dBMs (see, e.g., [8] [9] [10]). In this paper, we are mainly concerned with digital data sets that fall into four general categories:

1) *Raw Sensor Signals*. A device typically collects data from multiple sensor signals at varied pre-configured sampling frequencies to minimize study participants' burden under free living conditions. In most cases, the sensor signals are collected in a nonstop $24 * 7$ fashion throughout the entire study, which generally runs between weeks to months. Therefore, assessing potential issues, such as sensor malfunctioning, or wear non-compliance due to participants' behaviors, is critical to ensure data quality can satisfy the downstream analytics needs. Meanwhile, the quality and coverage of sensor data directly correlate to the dBM derivation, which will be discussed in the later sections of this paper.
2) *Scored Data, or Digital Biomarkers*. In addition to raw sensor signals, device companies usually have their proprietary algorithms to analyze sensor data and derive dBMs from it. For example, heart rate and blood volume pulse can be derived from the raw photoplethysmography (PPG) sensor signal. Derived dBMs are at a much lower resolution than the sensor signal, often at the minute or half-minute level.
3) *Labels/Annotations*. As algorithms and machine learning models used in developing dBMs become more complex, requirements for large annotated data sets grow. Annotating data for machine learning applications is especially challenging in the biomedical domain as it requires the domain expertise of highly trained specialists to perform the annotations. Annotations can come as interval-based events, with precise timestamps to label the onset and offsets of disease events.
4) *Clinical Records*. Apart from raw sensor data and derived dBMs, one yet important piece of data is clinical records that provision key mappings, *e.g.*, device ID to participant ID, participant ID to the treatment cohort, visit dates to treatment phases, *etc*.

Unique challenges arise from these digital data and have made a case for us to develop a data quality assessment framework to define the expectation of these digital data (e.g., completeness, uniqueness, validity, integrity), to monitor the data for conformance to expectations throughout the dBM trials, and, finally, a user interface to display the findings to support operational and analytical activities.

III. Digital Data Quality Assessment Framework

The key functions in our data quality assessment framework should now be clear in Figure 1. The logical series of modeling steps, the problems they induce, and the ultimate resolution of the problems are in the rest of this section as follows.

A. Signal Data Quality Metrics

In the pre-study phase, we establish the Data Transfer Agreement (DTA), to clearly define data quality metrics regarding signal data, including raw sensor signals and dBMs. Below we list the typical quality metrics, and Table I gives an example of the data quality metrics table we find in a DTA document, where $acce_x$, $accel_y$, $accel_z$ and ec are raw sensor signals, st, po (categorical) are derived dBMs (or, scored data) from accelerometry data, and hr and re are the scored ones from ec.

- **Sampling Frequency** — For raw sensor signals, it is the preconfigured average number of samples obtained in one second. For derived dBMs, it is the resolution of resultant features from analyzing raw sensor data.
- **Valid Range** — For numerical variables (*i.e.*, sensor signals and dBMs), a valid range is indicated by minimum and maximum values that can be measured. For enumerated variables, it is a list of predefined categorical values. One example is the rest classification biomarker

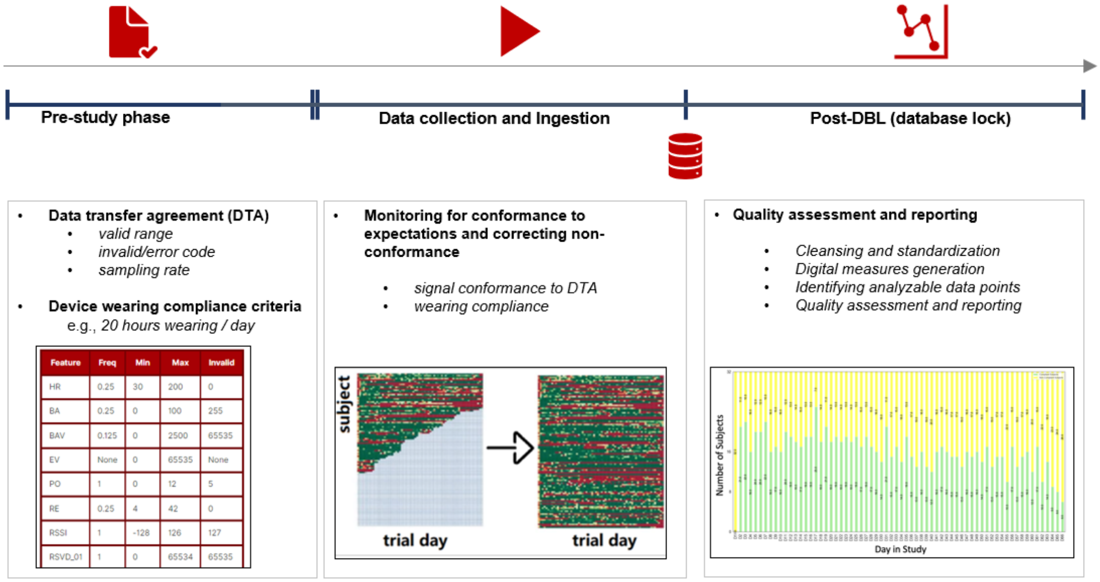

Pre-study phase
Data collection and Ingestion
Post-DBL (database lock)
Data transfer agreement (DTA)
valid range
invalid/error code
sampling rate
Device wearing compliance criteria
e.g., 20 hours wearing / day
Feature | Freq | Min | Max | Invalid
HR | 0.25 | 30 | 200 | 0
BA | 0.25 | 0 | 100 | 255
BAV | 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
subject
trial day
trial day
Quality assessment and reporting
Cleansing and standardization
Digital measures generation
Identifying analyzable data points
Quality assessment and reporting
Number of Subjects
Day in Study

Figure 1: The overall data quality assessment scenario — from establishing DTA in the pre-study phase, to compliance monitoring in the live phase, and finally to the quality assessment and reporting in the post-Database Lock (DBL) phase.

TABLE I: EXAMPLE OF A SIGNAL DATA QUALITY METRICS TABLE FOUND IN A TYPICAL DTA DOCUMENT.

Channel	Description	Units	Min Value	Max Value	Invalid Value	Sampling Frequency (Hz)
$accel_x$	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
po	Posture • Laying Down = 0 • Standing = 2 • Walking = 3 • Running = 4 • Unknown = 5 • Leaning = 11	Enum	0	11	5	1

which has the following classes: “awake”, “sleep”, “toss and turn” and “interrupted”.

- **Invalid Value/Error Code** — In addition to the valid range, devices often provision specific invalid values or error codes to indicate different statuses of malfunctioning, which helps pinpoint the underlying issue.

B. Signal Data Quality Assessment

Connected clinical trials for dBM research often are conducted under a free living condition, *i.e.*, participants wear sensor devices on a best effort basis using instructions communicated during study enrollment. Inevitably, the free living conditions, device wearing compliance, potential device failure, or device malfunction introduce data issues such as missing data or invalid data collected when participants do not wear or incorrectly wear the devices. Figure 2 illustrates how valid signals (*i.e.*, correctly worn signals) can mix with invalid signals (*i.e.*, incorrectly or not worn signals) in the data collection and how they differ when plotted. Therefore, a *qualitative* means is needed to tell whether a device was operating normally and worn correctly (*i.e.*, **data usefulness**).

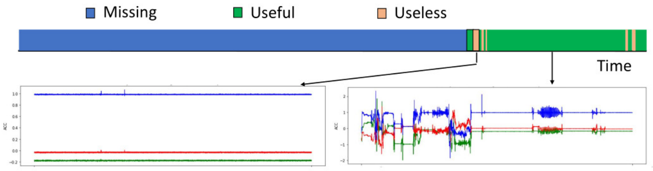


Figure 2: Illustration of sensor signal data issue. Visualized sensor data show different patterns when correctly versus incorrectly worn.

To fulfil this goal, the quality assessment is performed in two stages, as discussed in the following.

- **Validity Check**. Data validity check leverages signal data metrics, as discussed in Section III-A. We immediately know how many valid data points we expect to receive for a sensor signal or dBM using its pre-configured sampling frequency. We can filter out invalid values with a valid value range to get valid data coverage, *i.e.*, coverage of valid data points.
  Since raw sensor signal directly correlates with derived dBMs, we can perform a validity check against the two independently and then align their valid data coverage to check the consistency. We may further overlay device incident events to understand the root cause of observed issues better.
- **Non-wear Detection**. After dropping out invalid data through the validity checking process, the subsequent task is to detect moments when the devices were not correctly worn. The non-wear detection can be challenging as data

from such moments can be entirely valid in terms of falling within its valid data range. Instead of reinventing the wheel, we rely on Biobank [11] [12], an accelerometer data processing pipeline whose non-wear detection module is widely adopted as a standard. Below are two key concepts in non-wear detection.

- *Epoch* — Although data points are collected initially at a high resolution, *e.g.*, 50Hz sampling frequency, the processing is conducted on aggregated values (*e.g.*, 1 or 5 second *short epochs* or 15 minutes *long epochs*) due to the following reasons: (1) collapsing data to epoch summary measures helps to standardize differences in sample frequency across studies; (2) there is little evidence that raw data is an accurate representation of body acceleration, and all scientific evidence so far has been based on epoch averages; (3) collapsing data to epoch summary measures also helps to average out different noise levels making results more comparable across sensor brands.
- *Non-wear Detection* — Accelerometer non-wear time is estimated based on the standard deviation and the value range of the raw data from *each* accelerometer axis. Classification is done per 30-second epochs based on the characteristics of a larger window centered at these 30-second epochs. Specifically, Biobank identifies stationary periods in 10-second windows where all three axes have a standard deviation of less than $13.0mg$ ($1mg = 0.0098\, m \cdot s^{-2}$). These stationary periods are then used to define whether a window is stationary or not.

C. Signal Data Quality By Granularity

In addition to *qualitative* assessment as discussed in Section III-B, *quantitative* measures that define how much usable data is in a specific period (*i.e.*, **data quality** at different levels) are required before statisticians can begin analysis.

The Data Quality Model. Based on Biobank's non-wear classification on 30-second epoch level, we can further generate data quality that can be used for analysis at different time resolutions. Each phase in our data quality derivation flow is illustrated in Table II to Table V and expanded upon below. Column name "Cvge." is the abbreviation for "Coverage in Minutes".

- **Epoch Level** — This table is generated from Biobank's 30-second epoch classification. It serves as the working basis for subsequent data quality tables. Note that we have one additional column, "Subject," to indicate participant ownership of an epoch.
- **Hourly Level** — From the epoch quality table, we can apply a filter to only keep correctly worn epochs and in turn infer hourly data coverage in terms of compliant minutes. This hourly data quality table is the source for data quality reporting at the finest granularity.
- **Daily and Intraday Window Level** — From the hourly data quality table we can summarize the total coverage for each day and produce daily level data quality tables. In addition, for analysis purposes, we are often interested in specific intraday windows from which digital endpoints are derived — for instance, walking time or step count during the daytime (*i.e.*, daily **p**hysical **a**ctivity) and sleep hours during the nighttime. Thanks to the "Hour" column in the hourly quality table, intraday window coverage can be easily derived by applying filters.
- **Extended Quality with External Mappings** — We can further extend the data quality table with additional mappings when they become available as the study progresses, for instance, mapping between patients and sites/visits, as reported from the clinical operation site. These extra fields allow analysis-specific filtering and aggregation, *e.g.*, to find out which participants have sufficient data and set up individual baselines. We use this table to look for the patients with at least three valid days (>= 20 hours of data for a day to be qualified as a valid day) during a pre-treatment visit.

D. Representing Digital Data Quality

Fully understanding the quality of a large dataset, especially one that contains data from wearable device sensors, is not always a trivial undertaking. With numerous considerations to be cognizant of, as discussed in Section III-C, the most logical first step is to present the data with visualizations. Thoroughly understanding the data coverage and quality requires more than one visualization, simply because there is more than one aspect to check. This section presents a family of commonly used visualization examples in our data quality strategy.

- **Identifying Outliers and Missing Data.** Certain metrics must fall between threshold ranges depending on the study and associated data sources. One example is heart rate, which falls within a specified range of 30 to 200 beats/minute for one study. This range is outlined in the DTA for the study and must be applied to all heart rate data points collected. By plotting these signals against the specified thresholds, outliers can be immediately detected by viewing a plot. If outliers exist, further investigation will be completed for that participant's data to see if there are outliers for other metrics. Further, gaps in data can be identified within the same visualization, as demonstrated in Figure 3(a). Detailed data quality reports are generated in conjunction with the visualizations created for displaying outliers and missing data. For example, we convert the signal data from 3(a) to a sequence of colored blocks in Figure 3(b), with green blocks indicating valid sensor signal value in the corresponding period and red indicating missing or invalid signal value identified. In Figure 3(c), we compute the valid data ratio, and therefore can represent the data quality with a numeric value, or with a color from the color palette keyed to the valid data ratio (see e.g., Figure 3(d)).
- **Data Quality Map with Levels of Detail.** The quality of sensor signal data must be examined on various levels, each offering a specific level of detail. While certain levels are more useful for identifying distinct patterns, we will

TABLE II: EPOCH LEVEL QUALITY.

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

TABLE III: HOURLY LEVEL QUALITY.

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

TABLE IV: DAILY AND INTRADAY LEVEL QUALITY.

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

TABLE V: EXTENDED QUALITY WITH EXTERNAL MAPPINGS.

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

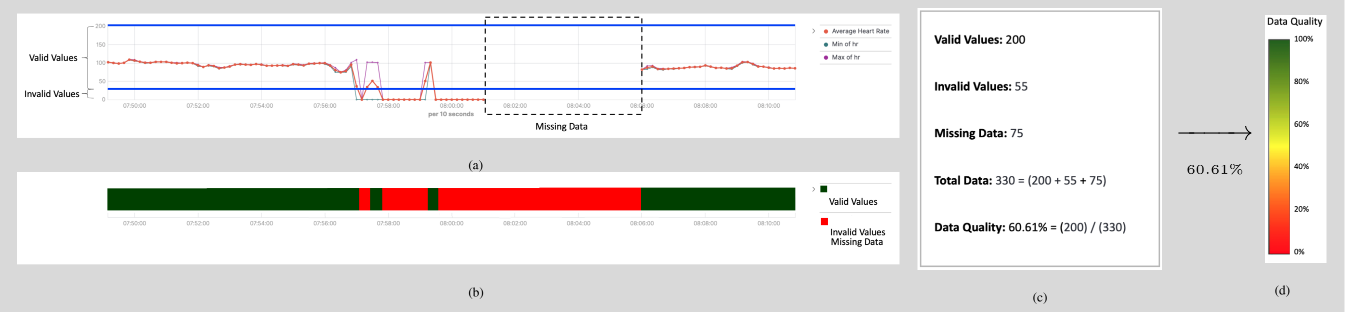
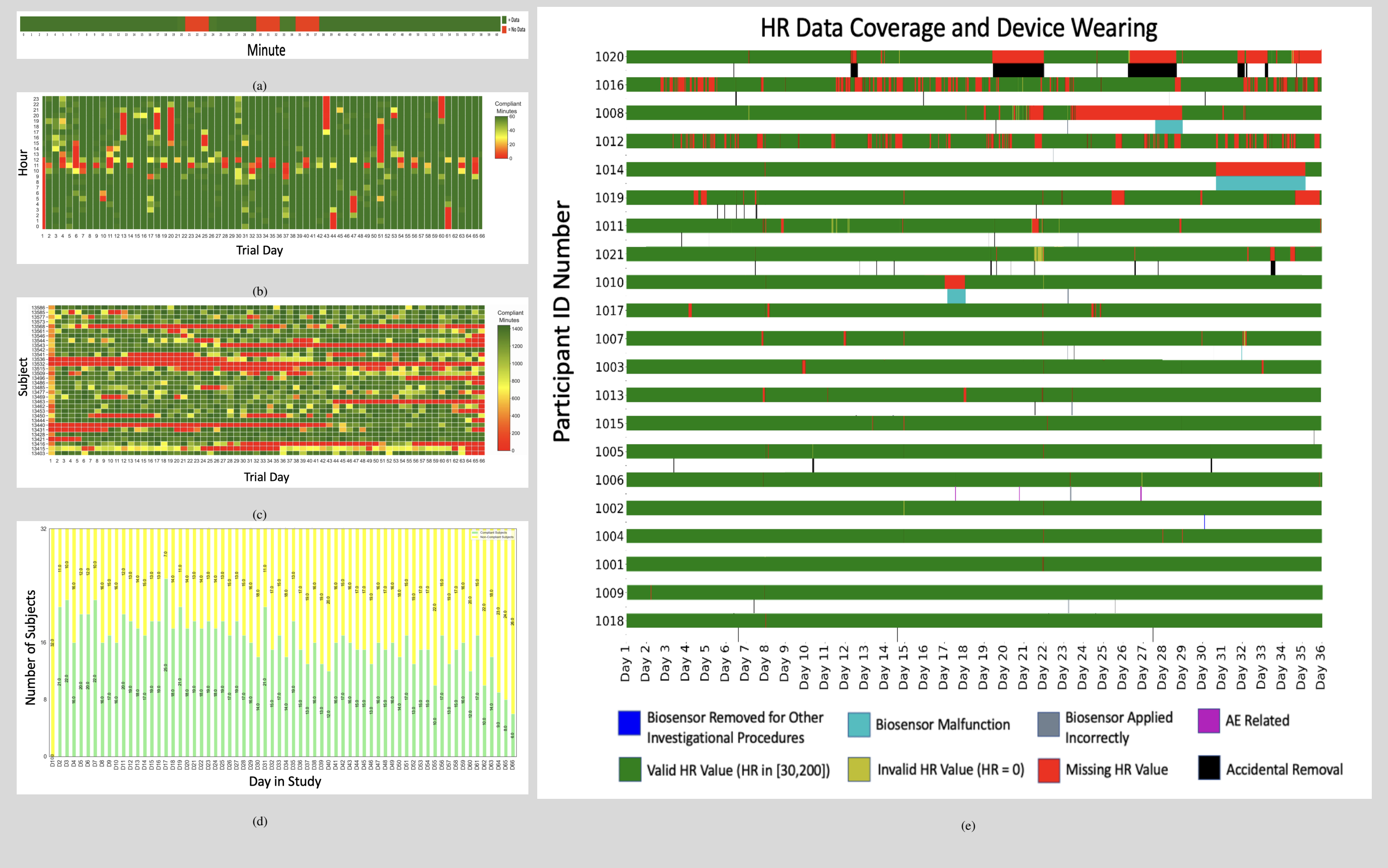


Figure 3: Visualization for sensor data quality. (a) Heart rate data (beats/minute) observed for one participant between 2021-02-15 07:49:00.000 and 2021-02-15 08:11:00.000. Valid range between 30 - 200 beats/minute, as denoted by threshold lines. Invalid data was observed multiple times. Missing data was observed between 2021-02-15 08:01:08.994 and 2021-02-15 08:06:09.000 with nearly 5 minutes of no data. (b) Use colored blocks to represent sensor signal data quality. (c) Deriving numeric representation of the data quality, *i.e.*, valid data ratio. (d) Interpreting data quality with color.

focus on the hourly, daily, and study levels on both a patient and population level:

- *Minute-by-Minute Quality Map for a Day* — Examining signals on a minute level can help to identify the minutes where a device may have intermittent connectivity, or more minor issues can be identified and further inspected, as seen in Figure 4(a).
- *Hour-by-Hour Quality Map for a Trial* — Zooming out, we can look at each hour across all days in the study. The hourly level aggregation mentioned in Section III-C is used to configure the day level plot, shown in Figure 4(b). This figure shows minutes of data coverage for each hour across all study days. This type of visualization allows us to look at compliance trends for a patient that may persist during certain hours of each day. Figure 4(b) shows an interesting device wearing pattern for the participant — taking off the wearable device to charge the battery for a couple of hours in the middle of each day of the trial has resulted in *missing data*, visualized as a sequence of red blocks in the center area of the map.
- *Day-by-Day Population-level Quality Map for a Trial* — Plotting data quality for all hours, days, and participants in a study yields the observation of data quality patterns seen in Figure 4(c). This study-level visualization can help us gain insights into the overall data quality at the population level and the compliance trends at the participant level throughout the trials.
- *Compliant Days Throughout a Trial* — In addition to the number of hours per day, it is also useful to view the number of *compliant* In addition to the number of hours per day, it is also useful to view the number of *compliant* days throughout the study, with a definition of compliance dependent on a study's protocol. One can recognize device-wearing patterns by plotting the number of patients compliant daily in a given study. As seen in Figure 4(d), the number of compliant days in a study decreased due to reduced device wearing as the study progressed.

- **Identifying and Aligning Data Issues.** In many clinical trials, it is a requirement that patients visit a site periodically. Whether it be for receiving dosing of a



(a)

(b)

(c)

(d)

(e)

Figure 4: Plots showing (a) minute-level quality representation throughout a participant day, (b) hourly-level quality representation for a participant throughout an entire trial, (c) daily-level quality for a population throughout the entire trial, (d) number of compliant days across all days in a study and (e) data coverage and device wearing issues observed throughout a study.

drug, having their vitals checked, or obtaining a device, information is collected by the sites and stored in various reports. One type of report, device reports, are used during data processing and can help understand the device's overall performance, specifically if any device issues exist. Additionally, information derived from these reports can be used to populate visualizations such as Figure 4(e). By combining this visualization with the information received in site reports, patterns specific to potential device issues and wearing patterns can be derived.

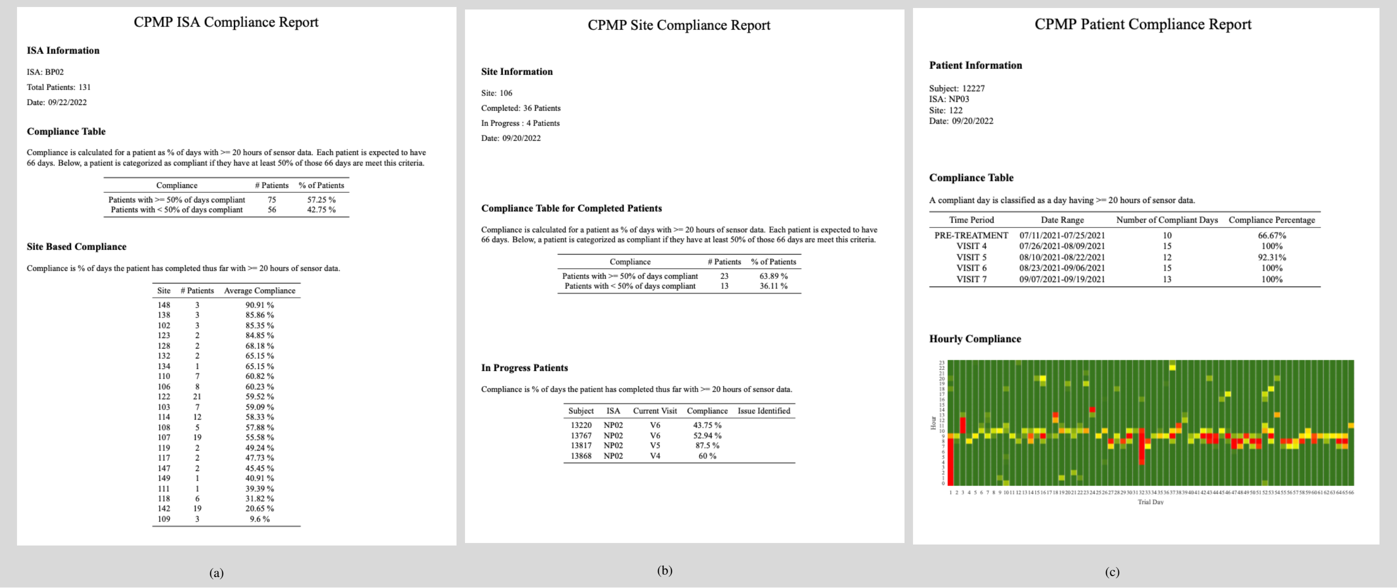
From the aforementioned data visualizations, various issues and patterns can be identified. When these are paired with actionable recommendations and delivered to the study team promptly, the study team can notify the corresponding site and participant to ensure the issue is rectified. This process leads to a quick turnaround time for potential improvements to data collection and can resolve the challenges that create low compliance in studies.

E. Generating Compliance Reports

Visualizing data is key to understanding data quality, as discussed in Section III-D. However, it is equally important to have a standardized reporting system for compliance to distribute quality and compliance information. Such systems generate reports that outline compliance on three levels: trial, site, and patient. In addition, automated generation allows systems to be configured at the start of a trial and run at set cadences to produce consistent quality assessment reports efficiently.

For each report, regardless of the level or contents, the thresholds used to configure and derive data metrics and visualizations are based on the expectations outlined in the study protocol. Each report aims to give insights into the population's compliance behavior:

- **Trial Summary**: A single comprehensive trial report can be generated and contains metadata regarding the number of patients, sites, and overall compliance percentages.
- **Study-Level Compliance**: A study-level report, such as Figure 5(a), will typically contain metrics displaying overall enrollment and compliance on a site level. These can allow a clinical trial team to gauge the progress of a specific study easily, *i.e.*, the number of patients who have completed their time in the study and the number of patients still in progress.
- **Site-Level Compliance**: Generating reports based on sites, as seen in Figure 5(b), allows clinical teams to efficiently



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)

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)

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)

Figure 5: Putting together compliance reports for Intervention-Specific Appendices (ISAs) under Chronic Pain Master Protocol (CPMP). (a) Generated compliance reports on the patient level. (b) Compliance by visit. (c) Customizable compliance report at patient level.

identify which sites may be experiencing issues regarding low compliance across their assigned patients. Typically, site reports contain information for overall performance, with specifics for patients that may fall below a set compliance threshold. The patients with low compliance are labeled with a potential issue- such as low compliance during the nighttime. The potential issues are derived from the hourly compliance for that patient. From here, sites can identify which of their patients contribute most to low compliance and attempt to resolve the issues linked to the low compliance.

- **Patient-Level Compliance**: Reports on a patient level can give insight into their specific patterns of device wearing. In these reports, as seen in Figure 5(c), the number of visits, compliant days within each visit, and compliance percentage per visit are displayed. In addition, an hourly compliance heatmap is visible, allowing for further understanding of when patients wear their devices across the study duration.

F. Data Quality in Novel Digital Endpoint Development

For novel digital endpoint development, raw sensor signals are collected along with annotations or labels, considered the ground truth. Annotations describe events explaining the status of the patient. As such, it is critical to assess the data quality of annotations and sensor signals to identify and address as many defects as possible.

Assessing Annotation Quality. Annotations are typically collected through patient reporting via a survey system or are labeled via software by trained clinicians who observe patient behavior. We first check for defects in the annotations. Defects may include improper data structure, invalid label categories, incomplete annotations, duplicates, and impossibly overlapping annotations. Defects could be caused by bugs in the annotation software or improper training on how to label.

Assessing Annotation Quality with Sensor Signals. Evaluating annotation quality in isolation is insufficient because digital endpoint development requires both annotations and raw sensor signals. So, we must also assess the data quality of annotations and raw sensor signals in conjunction. Therefore, we plot annotated time segments along with raw sensor signals (e.g., Figure 6) to facilitate the data quality assessment.

Discrepancies in the alignment of annotations and raw sensor signals can vary considerably due to time tracking configurations and device properties in each step of the data collection process. Misalignment between annotation and raw sensor signals can be caused by improper device time configuration or the precision of the sensor device's initial time configuration. In addition, if the sensor device's time tracking is not periodically synced, the device's internal Real-time clock (RTC) will slowly drift over time. We measure drift using the sensor signal overlaid with annotation plots. Once the misalignment from the initial configuration time and RTC drift are measured, we align the raw sensor signals to the annotations.

After the annotations and sensor signals have been properly aligned, we observe the plots to identify possible defects in annotation quality. Defects could include improper labels, annotated events that are not apparent in the sensor signals, and time segments that appear to be missing annotations or sensor signals. Specific time segments of concern are selected and validated with the source to determine if further action is needed.

Lastly, depending on study-specific requirements, we may apply other methods to assess data quality. For example, output from movement detection algorithms can be compared to

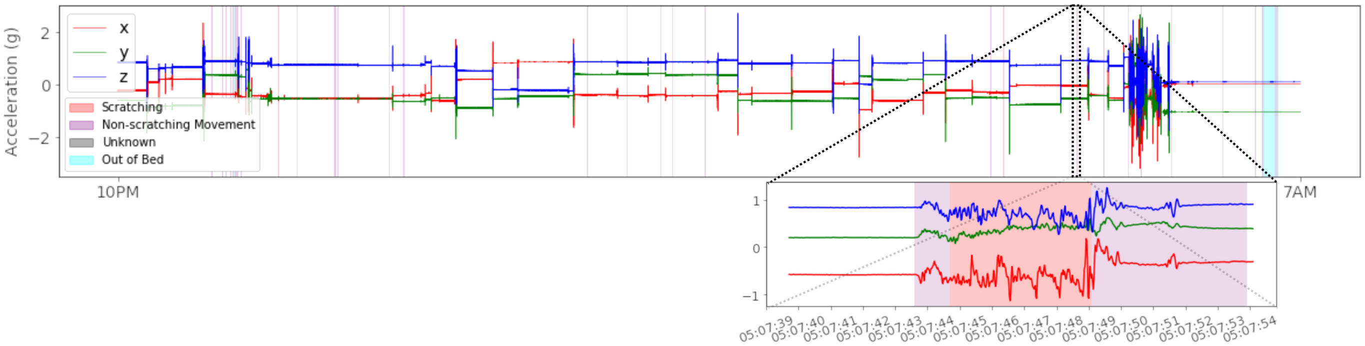
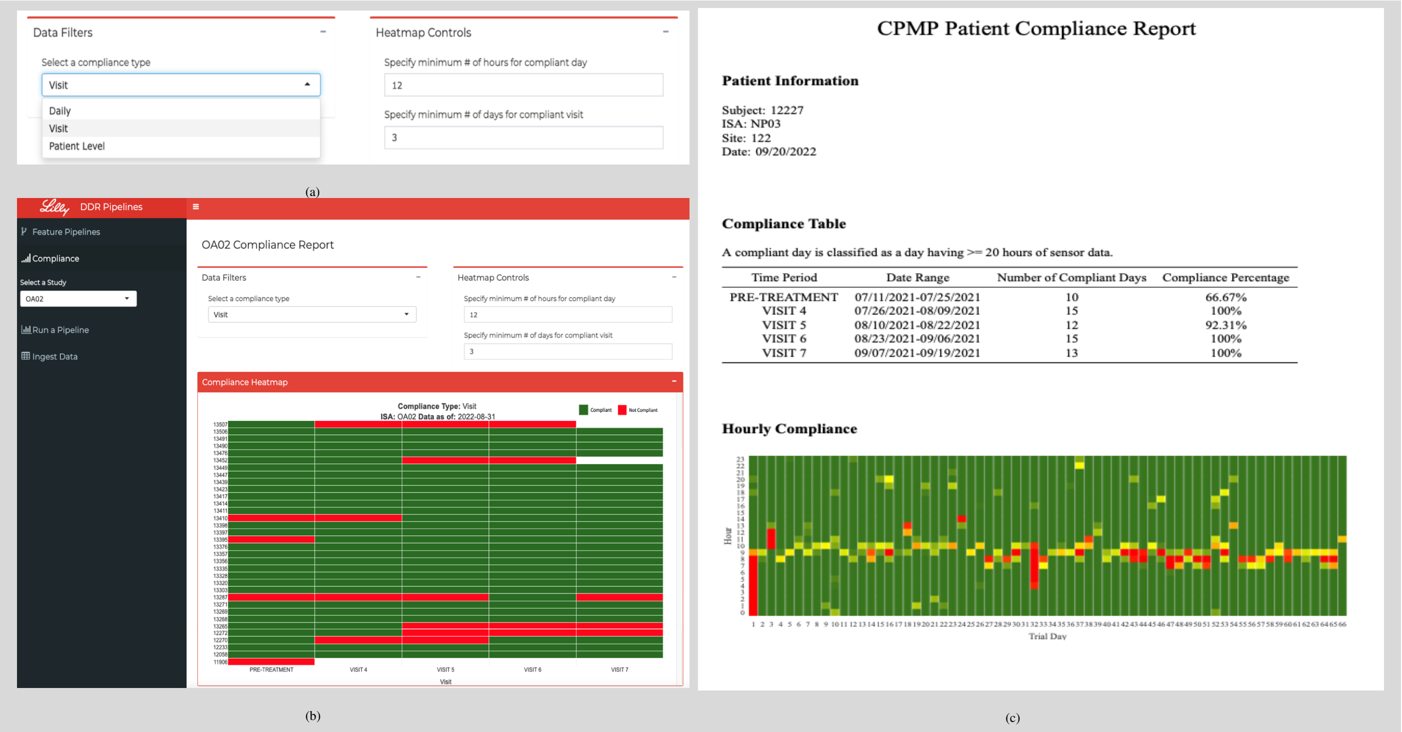


Figure 6: A plot of sensor signals overlaid with annotation labels is used to assess the data quality of annotations in conjunction with sensor signals.



Data Filters

Select a compliance type: Visit

Heatmap Controls

Specify minimum # of hours for compliant day: 12

Specify minimum # of days for compliant visit: 3

(a)

Select a Study: OA02

OA02 Compliance Report

Data Filters

Select a compliance type: Visit

Heatmap Controls

Specify minimum # of hours for compliant day: 12

Specify minimum # of days for compliant visit: 3

Compliance Heatmap

(b)

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)

Figure 7: The platform features displaying (a) filters for customizable compliance reports, (b) compliance by visit, and (c) generated compliance reports on the patient level.

annotated time segments that describe the movement to check annotation validity and coverage. Using various methods to assess data quality from different approaches is essential to maintain the data quality needed for novel digital endpoint development.

IV. The Data Quality Assessment Platform

Throughout a clinical trial, accessing data quality metrics is critical to upholding our outlined principles. Therefore, in addition to the compliance reports generated, an interactive data quality assessment platform is used to monitor data quality throughout a trial continuously.

The platform design allows users to customize the plots and view data quality through various lenses, utilizing filters and user controls. For example, users may want to view compliance on a day, visit, or patient level. As seen in Figure 7(a), they can select the level and the metric for which the visualization will show, as discussed below.

Let us take configuring and viewing compliance visualizations as an example. A user wants to view compliance for all patients in a study on the visit level, as seen in Figure 7(b). They define *compliance* as having at least 12 hours of data daily, with 3 days each visit comprising a compliant visit. By selecting the compliance type, which in this case is visit, and inputting the number of hours and days for defining compliance, the user can see the population’s compliance for these specific thresholds, as seen in Figure 7(a). Additionally, they can easily compare and contrast different levels and compliance thresholds

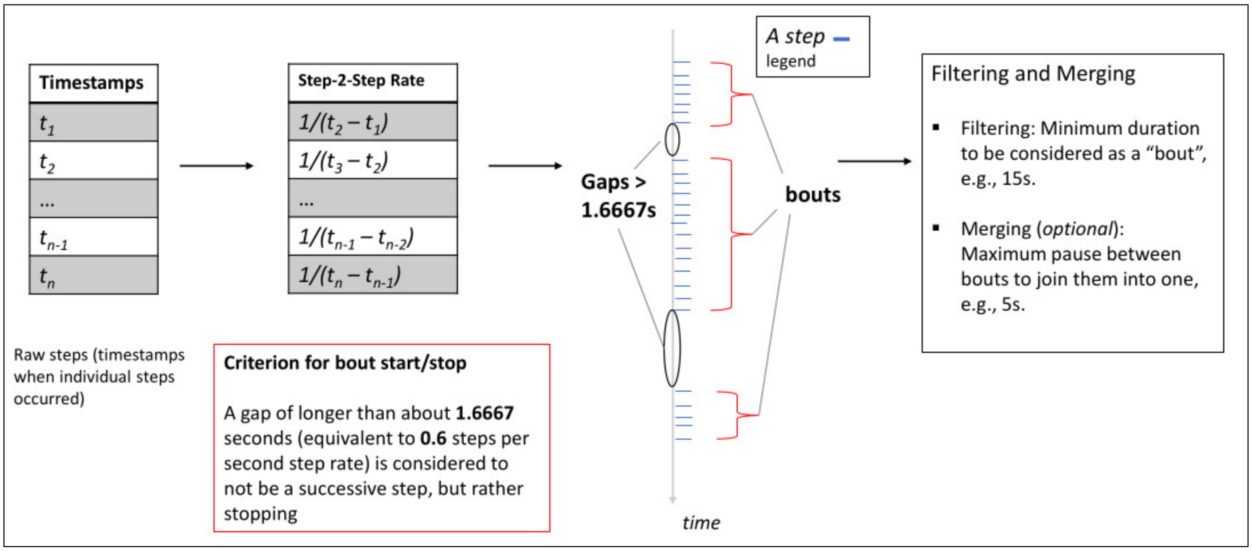

Timestamps
t_1
t_2
…
t_{n-1}
t_n
Step-2-Step Rate
$1/(t_2 - t_1)$
$1/(t_3 - t_2)$
…
$1/(t_{n-1} - t_{n-2})$
$1/(t_n - t_{n-1})$
Raw steps (timestamps when individual steps occurred)
Criterion for bout start/stop
A gap of longer than about **1.6667** seconds (equivalent to **0.6** steps per second step rate) is considered to not be a successive step, but rather stopping
Gaps > 1.6667s
A step
legend
bouts
time
Filtering and Merging
Filtering: Minimum duration to be considered as a "bout", e.g., 15s.
Merging (*optional*): Maximum pause between bouts to join them into one, e.g., 5s.

Figure 8: Process of transforming raw steps into bouts.

within the data quality assessment platform.

In addition to the compliance assessment, data quality visualizations, such as Figure 4, are created and customized within the platform. For example, as seen in Figure 7, a user can select a specific time range or time level to view the data. This zoom in and out can be used to identify and trace patterns of device wearing.

The data quality assessment platform allows for customizable, real-time, informative visualizations that enable insights into patient compliance and device-wearing data patterns. The study team can process and act upon these key insights with these visualizations housed in a centralized, consistent, and efficient platform.

V. Digital Endpoints

With out data quality assessment platform, we are able to derive digital endpoints from two categories: **P**hysical **A**ctivity (PA) [13] [14] and sleep [15] [16]. Typical PA features include duration of daily light/moderate/vigorous activities, steps count and gait features. For sleep features they are night sleep duration and **W**akeup **A**fter **S**leep **O**nset (WASO).

Gait features are a unique set of physical activity endpoints that unveil fine-grained walking characteristics, for which we see a significant distinction between health and chronic pain cohorts. Due to their importance, we detail our effort in deriving gait features in this section.

Determining bouts is the most fundamental step since all gait features are based upon bouts. Figure 8 illustrates this process: (1) raw individual steps with their timestamps are obtained from an open source step detection algorithm; (2) derive step rate for every two consecutive steps; (3) since bout by definition is a short period of intense walking activity with less than 1.6 seconds of stop between two steps, we can apply this gap threshold to detect individual gaps; and (4) depending on specific settings of a study (*e.g.*, profile of participating cohorts), we apply a constraint on minimum bout duration (*e.g.* filtering to keep $>= 15s$ bouts) and optionally merge bouts with small gaps in between into a single bout.

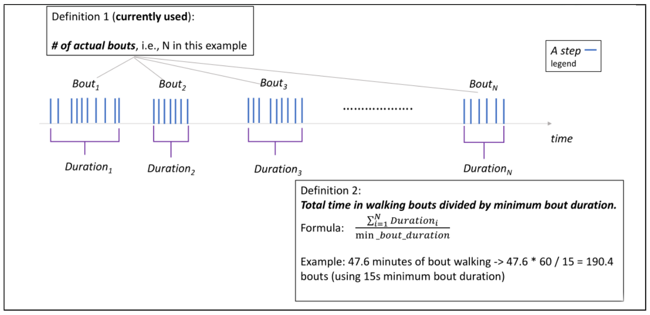

Definition 1 (**currently used**):
# of actual bouts, i.e., N in this example
A step
legend
$Bout_1$
$Bout_2$
$Bout_3$
$Bout_N$
………………..
time
$Duration_1$
$Duration_2$
$Duration_3$
$Duration_N$
Definition 2:
Total time in walking bouts divided by minimum bout duration.
Formula: $\frac{\sum_{i=1}^{N} Duration_i}{\min\_bout\_duration}$
Example: 47.6 minutes of bout walking -> 47.6 * 60 / 15 = 190.4 bouts (using 15s minimum bout duration)

Figure 9: Bout count.

Once bouts are identified, we can derive bout and gait-related features. Below we summarize the derivation process.

- **Bout Count**. We currently use the definition of actual number of identified bouts. Another meaningful definition is the count in terms of minimum duration bout , *i.e.*, $\sum_{i=1}^{N} \frac{Duration_i}{15} = 1$, where 15s is used as minimum bout duration. Figure 9 illustrates the two definitions.
- **Bout Duration**. Bout duration is the average duration across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} Duration_i}{N}$.
- **Steps per Bout**. Steps per bout is average of the count of steps across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} StepCount_i}{N}$.
- **Cadence**. A single $bout_i$'s cadence is the number of steps per its duration, *i.e.*, $\frac{StepCount_i}{Duration_i}$, we can then use the averaged cadence across all bouts for the cadence feature, *i.e.*, $\frac{\sum_{i=1}^{N} Cadence_i}{N}$, as shown in Figure 10.
- **Gait Rate**. For a single $bout_i$ with $M + 1$ steps, its mean step rate is defined as $\frac{\sum_{i=1}^{M} stepRate_i}{M}$, where $stepRate_i = \frac{1}{t_{i+1} - t_i}$ is the step rate between $step_{i+1}$ and $step_i$, whose occurring timestamps are t_{i+1} and t_i respectively. The gait rate feature is then derived as the average of the

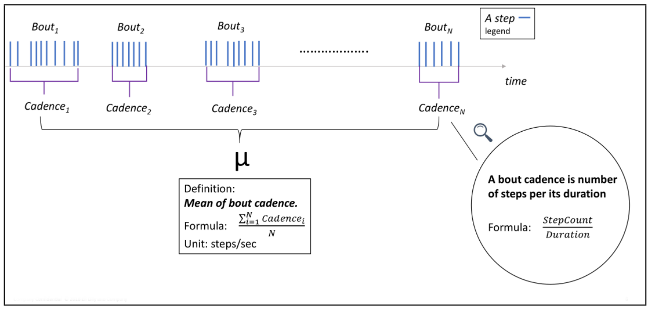


Figure 10: Cadence.

mean step rate across all bouts, *i.e.*, $\frac{\sum_{i=1}^{N} MeanStepRate_i}{N}$. Figure 11 illustrates this process.

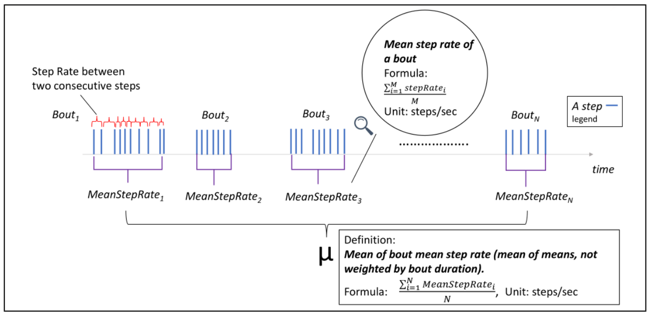

Step Rate between two consecutive steps
Bout$_1$
Bout$_2$
Bout$_3$
Bout$_N$
Mean step rate of a bout
Formula: $\frac{\sum_{i=1}^{M} stepRate_i}{M}$
Unit: steps/sec
A step
legend
time
MeanStepRate$_1$
MeanStepRate$_2$
MeanStepRate$_3$
MeanStepRate$_N$
μ
Definition:
Mean of bout mean step rate (mean of means, not weighted by bout duration).
Formula: $\frac{\sum_{i=1}^{N} MeanStepRate_i}{N}$, Unit: steps/sec

Figure 11: Gait rate.

- **Gait Rate Standard Deviation**. Similar to mean step rate, for a single $bout_i$ with $M + 1$ steps, we can calculate standard deviate over the M steps rates, *i.e.*, $\sigma(StepRate_i)$, $i = 1 \cdots M$. The feature is then derived as the mean of standard deviation in step rate from each bout, *i.e.*, $\frac{\sum_{i=1}^{N} StepRateStd_i}{N}$.
- **Step Rate Change**. As shown in Figure 12, a bout's step-to-step rate change is the difference of step rate from the first set of steps (*i.e.*, steps 6 to 8) to steps 23 to 25 on any period of walking with at least 25 steps long. Therefore for $bout_i$ with 25 or more steps, its step to step rate change can be calculated as $\mu(\sum_{i=23}^{25} StepRate_i) - \mu(\sum_{i=6}^{8} StepRate_i)$. In turn, the feature is the mean of step rate change from each eligible bout ($\geq$ 25 steps), *i.e.*, $\frac{\sum_{i=1}^{N} StepRateChange_i}{N}$.

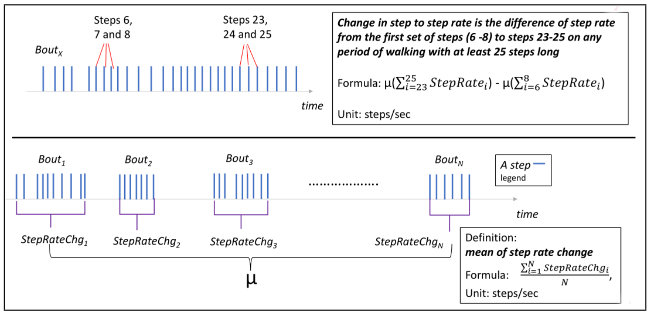

Steps 6, 7 and 8
Steps 23, 24 and 25
Bout$_X$
time
Change in step to step rate is the difference of step rate from the first set of steps (6 -8) to steps 23-25 on any period of walking with at least 25 steps long
Formula: $\mu(\sum_{i=23}^{25} StepRate_i)$ - $\mu(\sum_{i=6}^{8} StepRate_i)$
Unit: steps/sec
Bout$_1$
Bout$_2$
Bout$_3$
Bout$_N$
A step
legend
time
StepRateChg$_1$
StepRateChg$_2$
StepRateChg$_3$
StepRateChg$_N$
μ
Definition:
mean of step rate change
Formula: $\frac{\sum_{i=1}^{N} StepRateChg_i}{N}$,
Unit: steps/sec

Figure 12: Step rate change.

VI. Conclusion and Future Work

As DHT continues to evolve and collect more complex digital data in clinical trials, the need for a digital data quality assessment platform is increasing. By defining and implementing the fundamentals of data quality into the digital data quality framework and platform, we can generate automated compliance reports, customizable visualizations, and real-time quality metrics. In addition, the methods for facilitating dBM research have been simplified with the centralized digital data quality assessment platform. As dBM research continues, so will the use of the digital data quality assessment platform. 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.

References

[1] J. M. Wright *et al.*, "Evolution of the digital biomarker ecosystem," *Digital Medicine*, vol. 3, no. 4, pp. 154–163, 2017.

[2] R. Y. Wang, V. C. Storey, and C. P. Firth, "A framework for analysis of data quality research," *IEEE transactions on knowledge and data engineering*, vol. 7, no. 4, pp. 623–640, 1995.

[3] A. Sharma *et al.*, "Using digital health technology to better generate evidence and deliver evidence-based care," *Journal of the American College of Cardiology*, vol. 71, no. 23, pp. 2680–2690, 2018.

[4] R. Lyons, G. R. Low, C. B. Congdon, M. Ceruolo, M. Ballesteros, S. Cambria, and P. DePetrillo, "Towards an extensible ontology for streaming sensor data for clinical trials," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2021, pp. 1–6.

[5] C. M. Rey, "Wearable data revolution: Digital biomarkers are transforming research, promising a revolution in healthcare," *Clinical OMICs*, vol. 6, no. 2, pp. 10–13, 2019.

[6] I. Clay, "The future of digital health," *Digital Biomarkers*, vol. 4, no. 1, pp. 1–2, 2020.

[7] M. Chen and M. Decary, "Artificial intelligence in healthcare: An essential guide for health leaders," in *Healthcare management forum*, vol. 33, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 10–18.

[8] S. M. Hossain *et al.*, "Mcerebrum: A mobile sensing software platform for development and validation of digital biomarkers and interventions," ser. SenSys '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–14.

[9] A. Dillenseger *et al.*, "Digital biomarkers in multiple sclerosis," *Brain Sciences*, vol. 11, no. 11, pp. 1519–1544, 2021.

[10] M. M. Rahman *et al.*, "Towards reliable data collection and annotation to extract pulmonary digital biomarkers using mobile sensors," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 179–188.

[11] A. Doherty *et al.*, "Large scale population assessment of physical activity using wrist worn accelerometers: the uk biobank study," *PloS one*, vol. 12, no. 2, p. e0169649, 2017.

[12] C. Sudlow *et al.*, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS medicine*, vol. 12, no. 3, p. e1001779, 2015.

[13] V. T. Van Hees *et al.*, "Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity," *PloS one*, vol. 8, no. 4, p. e61691, 2013.

[14] S. Sabia *et al.*, "Association between questionnaire-and accelerometer-assessed physical activity: the role of sociodemographic factors," *American journal of epidemiology*, vol. 179, no. 6, pp. 781–790, 2014.

[15] V. T. van Hees *et al.*, "Estimating sleep parameters using an accelerometer without sleep diary," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[16] A. Doherty *et al.*, "Gwas identifies 14 loci for device-measured physical activity and sleep duration," *Nature communications*, vol. 9, no. 1, pp. 1–8, 2018.