Defining Adherence: Making Sense of Physical Activity Tracker Data

LIE MING TANG, University of Sydney, Australia
JOCHEN MEYER, OFFIS Institute for Informatics, Germany
DANIEL A. EPSTEIN, University of Washington, United States
KEVIN BRAGG, University of Sydney, Australia
LINA ENGELEN, University of Sydney, Australia
ADRIAN BAUMAN, University of Sydney, Australia
JUDY KAY, University of Sydney, Australia

Increasingly, people are collecting detailed personal activity data from commercial trackers. Such data should be able to give important insights about their activity levels. However, people do not wear or carry tracking devices all day, every day and this means that tracker data is typically *incomplete*. This paper aims to provide a systematic way to take account of this incompleteness, by defining *adherence*, a measure of data completeness, based on how much people wore their tracker. We show the impact of different adherence definitions on 12 diverse datasets, for 753 users, with over 77,000 days with data, interspersed with over 73,000 days without data. For example, in one data set, one adherence measure gives an average step count of 6,952 where another gives 9,423. Our results show the importance of adherence when *analysing* and *reporting* activity tracker data. We provide guidelines for defining adherence, analysing its impact and reporting it along with the results of the tracker data analysis. Our key contribution is the foundation for analysis of physical activity data, to take account of data incompleteness.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*;

Additional Key Words and Phrases: physical activity trackers, adherence, wear-time, data completeness

ACM Reference Format:
Lie Ming Tang, Jochen Meyer, Daniel A. Epstein, Kevin Bragg, Lina Engelen, Adrian Bauman, and Judy Kay. 2018. Defining Adherence: Making Sense of Physical Activity Tracker Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 37 (March 2018), 22 pages. https://doi.org/10.1145/3191769

1 INTRODUCTION

Millions of people track their physical activity. Eighteen percent of US adult consumers own a wearable fitness tracker, with the majority using their device often [1]. Over 22 million wearable devices were

Authors' addresses: Lie Ming Tang, University of Sydney, School of Information Technologies, NSW, 2006, Australia; Jochen Meyer, OFFIS Institute for Informatics, Oldenburg, 26121, Germany; Daniel A. Epstein, University of Washington, Seattle, WA, 98195-2350, United States; Kevin Bragg, University of Sydney, Sydney School of Public Health, NSW, 2006, Australia; Lina Engelen, University of Sydney, Sydney School of Public Health, The University of Sydney, NSW, 2006, Australia; Adrian Bauman, University of Sydney, Sydney School of Public Health, NSW, 2006, Australia; Judy Kay, University of Sydney, School of Information Technologies, NSW, 2006, Australia.

2474-9567/2018/3-ART37 $15.00
https://doi.org/10.1145/3191769

shipped in the second quarter of 2016 alone [21]. Beyond this, smart phones and watches are increasingly making it possible for more people to track their physical activity.

This means that people are building up vast collections of data about their physical activity. Such data should be valuable for the *individual* who wants to understand their own long term activity. The *aggregated data* can provide a low cost way to collect data about various populations. Table 1 shows examples of the types of questions that such data have the potential to answer. The first row illustrates questions about *activity level*. For example, an individual may want to know how many steps a day they average; an example answer is 10,500 steps. Answering such questions can give the benefits for reflection, self-monitoring and planning as documented in personal informatics research [9, 15, 18, 19, 25, 28, 33]. *Aggregate analyses* can provide a corresponding average daily step count, such as the 7,500 for that population. A large body of health literature has examined such questions, for example, to inform recommendations about levels of physical activity for good health [20, 41]. The second type of question asks whether a target *goal has been met* [38, 41]. For example, did I meet my goal of >30 active minutes a day.

A key challenge for interpreting physical activity data is that it is typically incomplete [13, 26, 30, 34]. Many factors can contribute to gaps in the data, such as forgetting to wear the device, device loss or changes in motivation to track [3, 9, 14, 26, 30]. To answer questions like those in Table 1, it is critical to account for this incompleteness. To see why this is so, consider the first question in the table − determining a person's average step count. Consider the case of Alice, who wears her tracker all day, every day; a reliable answer can be calculated as a simple average over each day's step counts. But consider another person, Bob, who wears his tracker only on 60% of days − but on those days he wears it at all, he wears all day. His average daily step count should be based on *just those days where he has data* (the total step count divided by the days with data). We also need to consider the impact of the wear time within a day. Consider another person, Carol, who wears her tracker every day but only *in the morning on weekends* and *all day on weekdays*. Now a meaningful answer is more complex to determine − it needs to account for the incompleteness of her weekend data.

This paper aims to provide foundations for a systematic process to take account of the incompleteness of personal sensor data for physical activity when answering questions like those in Table 1. We introduce ***adherence***, a notion that reflects the fact that an activity tracker should give accurate answers to questions about activity for people like Alice who has 100% adherence, wearing her tracker all day, every day. We aim to establish adherence measures to account for people with less than 100% adherence, be this like Bob, Carol or the myriad of other wearing possibilities.

While previous work has studied wearing behaviour [3, 9, 13, 14, 26, 30, 34], there has been no research on how to systematically tackle the analysis and reporting of that data to account for its incompleteness.

Table 1. Examples of important questions long term physical activity data can answer, at the level of an individual or aggregate, population levels.

Question Type	Individual	Aggregate
Activity Level	What is my average daily step count? (example answer: 10,500 steps)	What is average daily step count of this population? (example answer: 7,500 steps)
Has a goal been met	How often do I get >30 minutes moderate activity a day? (example answer: yes, on 70% of days)	What proportion of this population gets at least 120 minutes of moderate activity a week? (example answer: 20%)

This is important if people are to *trust* the information that applications report on physical activity, claiming, or appear to claim, that ubicomp sensor data gives objective truths of one's activity. Fogg [17] warns that when systems produce questionable data, people are less likely to trust them and so they are less useful as a behaviour change tool. Bentley et al [3] found that incomplete data led to a loss of trust in their tool. Consolvo et al ([9], page 234) also reported this when missing data affected self-monitoring feedback.

To address these challenges, our work aims to provide systematic foundations for analysis and reporting physical activity data, accounting for data incompleteness. To do this we tackle the following research questions:

RQ1 What is the impact of different adherence measures on data ignored?
RQ2 How can we account for adherence for Activity-level questions?
RQ3 How can we account for adherence for Goal-met questions?

We explored these questions by analysing the impact of different adherence definitions on 12 datasets, with a total of 753 physical activity tracker users, who had more than 77,000 days with data, interspersed with over 73,000 days without data. We analysed them with 4 adherence measures that have been reported in previous literature.

- >0 steps: the least stringent answers activity questions using data from any day that has any data [14, 26, 29, 30, 34],
- >500 uses only days with more than 500 steps [29, 30];
- >10 hours − uses only days with at least 10 different hours with data [6, 26, 31];
- 3-a-day one requires data within 3 time periods of the day [29, 30].

As this is the first work to establish a systematic way to account for adherence, we chose this core set of research questions, 12 very diverse datasets and these 4 adherence measures from the literature.

The next two sections introduce key terminology and related work. Then the core of the paper: the study design, results and their discussion.

2 DEFINITIONS OF ADHERENCE

This section introduces key terms for defining adherence. It begins with ways to describe tracker wear-time. It then defines a *valid day*, one with data of sufficient quality that it is meaningful to include in analyses, along with criteria for assessing the validity of a day and ways to report adherence. It concludes with a review of adherence that goes beyond a single day, to describe adherence for one week and for longer periods.

2.1 Adherence & Wear-time

Table 2 introduces terms that have been used to describe wear-time of trackers. The first row shows how *wear-time* has been expressed as the number of hours of wear-per-day. It also shows non-wear time. For devices like the Fitbit, it is difficult to distinguish inactivity (e.g., 0 or few steps) from non-wearing. However, Migueles et al [31], reviewing use of accelerometer data in physical activity studies, reports that 20 consecutive minutes with very low accelerometer activity reliably indicates non-wear time for adults. The second row shows one common use of the term, adherence, to mean that study participants wore their tracker for the required number of hours per day and days per week. This is used to determine whether each day's data, or a person's full dataset, has good enough adherence to be a reliable measure of their actual activity level. Doherty et al [11] reported the impact of various factors, including age, sex, day, time of day, and season on this adherence measure. We adopted this meaning for adherence. However, our work aims to go beyond analysis of study data where participants are recruited to wear a

tracker as instructed. We want to be able to meaningfully analyse the vast collections of activity data that that people are building up. For this, we need to refine the notion of adherence to provide a conceptual framework and a systematic process for ubicomp researchers, and others, to meaningfully interpret such data.

Table 2. Adherence terminology used in existing literature.

Terminology	Definition	Examples	References
Wear-Time versus Non-wear Time per day	Count of number of hours in a day that the tracker was worn.	A study may report participants had mean weartime of 9.5 hours per day.	[2, 6, 11, 26, 31, 34, 37, 41]
Adherence or Compliance to monitoring protocol	Adherence to instructions on wear time in a study	A study may require participants to wear a tracker for a minimum of 10 hours/day and >= 4 days/week.	[11, 16, 31, 37]
Adherence or Compliance to health recommendations	Adherence to a recommended level of physical activity.	A study may aim to assess if participants achieve the recommended 30 min per day of moderate-to-vigorous physical activity (MVPA) for adults.	[20, 38, 41]

The last row, *adherence to physical activity recommendations* (also called *compliance*), describes how well a person or sample population meets a recommended level of physical activity. Adherence to a recommendation is of enormous importance [20] and the use of trackers to obtain objective measures of activity levels is of intense interest in health literature [11, 16, 31, 38, 41].

2.2 Adherence Measures

Table 3 introduces terms for describing and defining adherence. The key notion is the *Valid day*, a day with enough data to justify including it in analyses. Any data from non-valid days are ignored. The next 3 rows describe valid day criteria that have been used in previous work, listed in the last column. The simplest is based on a step threshold. When the threshold is >0 steps [14, 26, 34] any day with any step data is considered valid. The >500 steps, used in [28, 30] requires at least 500 steps for a valid day.

The second and third rows are ways to determine if there is enough data *through the day* to make that day valid. The 10-hours criterion means that there are at least 10 hours in the day that each have at least one step [16, 31, 34, 37]. The 3-a-day measure, in [30] is another way to ensure data through the day, this time requiring data in each of three parts of a day.

The next part of the table show the cases to consider when answering Goal-met questions. Activity level questions need a minimum wear-time for a valid day, but this may not be needed for a ***Valid-Goal-Day***. For example, with a goal of 10,000 steps and a 10-hours valid day threshold, a valid-goal-day occurs whenever the goal is met:

- *Goal met (valid day)*: i.e. reached 10,000 steps, with 10 hours wear-time,
- *Goal met (non-valid day)*: i.e. reached 10,000 steps, with 3 hours wear-time.

If the goal is not met, the day is only valid if the 10-hour valid day threshold is met:

- *Goal not met + valid day*: e.g., reached 3,000 steps, with 10 hours wear-time,

Table 3. Adherence terms and definitions used in this study.

Terminology	Definition	Key References
Valid day	A day where there was sufficient data for it to be considered valid to include in analysis.	[14, 26, 29 - 31,34]
Valid Day Criteria		
Minimum step	A day is valid only if the step counts is above a set threshold. eg: >0 steps; >500 steps.	[14, 26, 29, 30]
Minimum count of hours with data	A day is valid only if the number of hours with steps is above a threshold. eg. 10 hours	[16, 26, 30, 31, 34, 37]
3-a-day	A day is valid if there is data within 3 time periods: eg. 3am to 11 am, 11am to 3pm and 3pm to 3am	[29, 30]
Valid-goal-day criteria: Additional criteria for to assess if a goal has been met or not		
Goal met	Goal threshold met (even if day is NOT valid) eg had > 10,000 steps	N/A
Goal not met	Goal threshold NOT met and day is valid	
Insufficent data	Goal not met AND invalid day	
Ways to report adherence		
Daily Adherence	Percent of valid days between first and last day of data in the dataset.	[14, 26, 29 - 31]
Weekly Adherence	For a single week, this is the number of valid days. eg 3 days (of 7). For a data set, it is the average number of valid days per week.	[26, 29 - 31, 34]

Then the only days are invalid if there is *insufficient data* to make the day valid if neither the threshold is met, nor the goal e.g., reached 3,000 steps with 4 hours wear-time.

The last two rows introduce terms to use when reporting adherence. **Daily** adherence refers to the percentage of valid days in a dataset, as a description of completeness in terms of valid days. Daily adherence can be calculated for an individual or a population.

We now consider adherence beyond a single day. **Weekly** adherence measures the average number of valid days per week (only calculated during weeks where there is at least one valid day). For example, suppose a person has 50% daily adherence (i.e., having 50% of days valid) and 7 days-per-week weekly adherence. This corresponds to a person who had only 50% of their days valid but this held over every day of the weeks with data. Table 4 summarises other terms that have been used to describe patterns of adherence. A **streak** is an unbroken sequence of valid days. In contrast, a **break, lapse or gap** describes sequences of days that are not valid days. **Phases or trials** describe a series of streaks separated by short breaks and ending with a long break.

In the next section, we review the large body of work reporting physical activity tracker adherence and existing methods to analyse and address incompleteness in physical activity tracker data.

Table 4. Measures for long term adherence patterns.

Terminology	Definition	Reference
Streak	Unbroken sequence of valid days.	[29, 30]
Break / Lapse / Gap	Sequence of days that are not valid.	[14, 29, 30]
Phases / Trials	A series of streaks, each separated by short breaks and ending with a long break.	[14, 29, 30]

3 RELATED WORK

In this section, we first review research on activity tracker wearing behaviour and how this impact data completeness. We then review work that highlights why this is important for ensuring user trust and the ability to reflect on their data. Finally, we review reported methods used to deal with data incompleteness. We describe this work using the term, adherence, as just defined, although these terms were not used by the authors.

3.1 Studies of Wearing Behaviour

Previous work has explicitly studied wearing behaviour and patterns. It indicates wide differences in wearing behaviour, and associated with these diverse levels of data completeness [14, 26, 30]. Epstein et al [14] identified three distinctive groups of tracker users: 1. short term, 2. intermittent and 3. long and consistent. Clearly the intermittent users have incomplete data. Meyer et al [29] also reported wide differences in daily adherence, 20% to 100% of days being valid. At the same time, there are reports of some users who do sustain high adherence over longer periods, months and even years [5, 14, 30, 34]. Some work has studied the factors affecting wear-time, including age, gender and environment [2, 11]; day of week [11, 26, 30]; time of day [11, 30]; and the efforts demanded by some tracking devices (e.g., battery life, water resistance) [9, 13, 18, 34]. This small but growing body of work highlights that there are diverse levels and patterns of wearing behaviour and so diverse levels of data completeness.

3.2 Importance of Accounting for Incomplete Data

This section reviews the work that shows the importance of dealing with the incompleteness of physical activity data for personal health and well-being applications. Consolvo et al (page 211 [9]) identified key design challenges for applications intended to encourage health and well-being. One of these relates to the *perceived accuracy* of trackers. For example, with their trackers, some users were disappointed that their tracker did not measure many activities such as vigorous gardening. They highlight the importance of missing data for self-monitoring feedback especially when graphs are used (page 234 [9]). Such displays tend to overlook missing or incorrect data that can lead to incorrect presentation of trends and patterns. In the Health Mashups system, Bentley et al [3] integrated data from multiple sources to present health behaviour patterns, well-being data and context to promote behaviour change. One of the major challenges they faced was the incompleteness in their source data (*sparseness*) which resulted in inaccuracy in their recommendations. Several users noticed contradicting information and that led to mistrust of the results. They link this result to Fogg's [17] analysis of persuasive systems, warning that systems producing questionable data are less likely to be trusted and thus they are less useful as a behaviour change tool.

Consolvo et al [9] suggested that presenting uncertainty can help make the self-monitoring data appear more accurate and precise. They argued the need to explore how uncertainty can be presented or managed. This view is supported by Kay et al [23] in a study of uncertainty presentations in transport schedules.

Participants reported that uncertainty information helped them make better decisions and alleviate anxiety when the app information did not match their knowledge. In a study of an interface that embeds daily and hourly adherence (wear-time) information in a calendar visualisation, [34] this adherence information helped users reflect on their long term activity, and to link this with their knowledge about the context. Some participants also reflected also on their wearing behaviour as well as factors affecting it. These results indicate the importance of accounting for and presenting information about incompleteness or uncertainty.

To summarize, adherence can impact the *perceived accuracy* of activity tracker data and there is evidence that presenting this may improve confidence and trust in the application as well as to support reflection.

3.3 Adherence Requirements in Studies of Physical Activity

Research on physical activity needs to be based on sufficient data that is of sufficient quality. A body of work has examined how many valid days within a monitoring period are required to provide a confident estimate of behaviour [39]. For example, systematic reviews of accelerometer based physical activity assessments [31, 37] have reported a range of criteria used, with a recommendation for at least 4 days in a 7 day monitoring period. However, they do not offer specific guidelines for longer monitoring periods except to warn that while increasing requirements for *valid-day* or valid-week can improve reliability of data, it also results in greater sample loss [31]. There has been some work on longer duration studies that used statistical methods to estimate missing values. For example, Tudor-locke et al [40] studied 23 participants over 1 year and used the Missing Value Analysis EM function in SPSS to estimate missing values. Similarly, a yearlong study involving 37 males and 44 females [36] also used an estimation method based on linear interpolation. Notably, both studies had very high adherence, 98% and 95% were valid over the 1 year study period. These interpolation methods presume the missing days are like the ones in the dataset, as assumption that is likely to be less reliable in datasets with lower adherence levels.

3.4 Summary

Existing literature highlights that people have varying wearing behaviours, providing diverse levels and patterns of completeness. This poses important problems for ensuring trust needed for people to effectively reflect on their data as a foundation for behaviour change. There has been limited work on how to address the incompleteness that can be expected in many datasets. This is the challenge that our work aims to address.

4 STUDY DESIGN

Our study design has three elements: a suitable collection of datasets; a set of adherence definitions to explore; and a sequence of experiments to perform to gain insights into our research questions. The design process began when the Sydney University authors were analysing several of their datasets and began to appreciate the need to for a systematic approach to taking account of tracker adherence. The team initiated the collaboration with the other authors, Meyer and Epstein, to discuss their insights, based on their work on wearing behaviours and patterns, including streaks, breaks, lapses, phases and trials as just described. The new team then established the design for this study.

4.1 Datasets

The team collected the 12 datasets presented in Table 5. This collection is diverse on many dimensions as we now describe. The first column shows the name we use to refer to the dataset. The second is the

Table 5. The 12 datasets from 9 studies of various lengths and population size. The first column is the identifier we use to describe the dataset. Next is the sample size and average duration in days, the average step count (using only days with >0 steps) and then the recruitment methods. The data source column distinguishes *volunteers* datasets (the first block), from the remainder, being *other study-generated* datasets.

Dataset	Sample Size	Avg Dur	Steps Per Day (Med)	Recruitment	PA Data Source	Focus of published research on these datasets
Volunteer1	113	344	9,025	Forums, mailing groups	User Volunteered	unpublished
Volunteer2 Volunteer3	141	325	7,057 5849	Forums, mailing groups + MTurk	User Volunteered	Studied lapses in tracker use. [13]
Volunteer4	23	523	9,136	Forums, mailing groups	User Volunteered	Participants recruited to report tracker use and see their long term activity data in a new interface. [34]
Volunteer5	33	443	5,549	Newsletter recruitment	User Volunteered	Study of wearing patterns for Vitadock users. [30]
Elder	86	59	7,522	Targetted mailing recruitment	Study Generated	Study of wearing patterns. This is the only dataset with mandatory adherence required. 12-week intervention for participants, aged 65+. [30]
Cardiac	44	194	5,449	Face-to-face recruitment of patients from 2 hospitals	Study Generated	Study of wearing patterns. A multicentric, comparative study with patients, aged 18-75. Starting within 30-days of myocardial infarction, in 12 months of rehabilitation. [30]
Lotus	8	259	6,709	Local participant database	Study Generated	Study of wearing patterns of devices by normal users under real-life over 9 months. [30]
Student1 Student2	97	33	7,519 7,562	On campus recruitment	Study Generated	Study of impact of SMS intervention on tracker adherence for university students Student1 is intervention group, n=49, Student2 is control, n=48. [4]
Student3 Student4	208	65	6,607	On campus recruitment	Study Generated	Study of tracker use and activity levels over a University semester (Student4 - IT students, n=68; Student3 - Medical Science students, n=140). [26]

number of people in the dataset (col 2) and then is the average duration in days (col 3), from the first to the last day with data. The forth column is the median steps per day calculated using the >0 steps threshold. The next two columns (col 5 and 6) indicate the sources of the data in terms of the means used to recruit participants and whether the data was *volunteers* or *other study-generated*. The last column overviews the ways the datasets have been used in previous published work or that it has not been published. The table organised groups as *volunteers* datasets first. The remaining datasets are all *other study-generated* because participants were recruited as part of another study and in these, participants were provided with a tracker.

The first 5 datasets are *volunteers*. In these, people who were already tracking were recruited to volunteer their data. *Volunteer1* consists of 113 Fitbit users who tracked from 18 to 731 days (average 344), with 73% having tracked for >=6 months. The *Volunteer2* and *Volunteer3* Fitbit datasets were used to study gaps and lapses in activity tracker use [13]. These had different recruitment methods: of the 141 users, the 67 of *Volunteer2* were recruited in a similar way to *Volunteer1* and the other 74 via Amazon Mechanical Turk (a popular crowd-sourcing website). The two groups had different Fitbit use patterns, with participants from the snowball recruitment wearing their Fitbits more and walking more each day than those recruited via Amazon Mechanical Turk. *Volunteer4* recruited long term Fitbit trackers via forums and email, for a study that sought to understand how they already used their long term activity

data and then to study their use of a calendar-based interface showing their full record of activity and wearing behaviour [34]. All but 2 had >6 months of data. The last in this group, *Volunteer5*, also a volunteered dataset, involves a different device, the VitaDock [30].

All the remaining datasets are *other study-generated*, meaning that the data was collected as part of another study for which participants were recruited to answer a question unrelated to studying wearing behaviour. The middle block of three datasets and these, along with *Volunteer5* were analysed by Meyer et al [30] to gain understanding of wearing patterns and how to describe them. These three studies used various trackers, including various versions of Fitbit and the Medisana ViFit tracker. The first, *Elder*, is distinctive in that it is the *only dataset in our collection where tracker use was mandatory*, as is typical in medical studies. This means that this dataset only has participants who had recorded tracker data. This dataset is also distinctive as it recruited an older population, aged 65-75. *Cardiac* users were recruited within 30 days of a myocardial infarction as part of a 12-month rehabilitation program. *Lotus* is a very small longitudinal observational study with no explicit intervention. This study's population is closer to the self motivated Fitbit users from *Volunteer1*, *Volunteer2*, *Volunteer3* and *Volunteer4*. These 3 studies used various activity tracking devices including various versions of Fitbit and the Medisana ViFit tracker.

The next 4 data sets came from studies of university students who were lent a Fitbit Zip. *Student2* and *Student1* involved Medical Science students, recruited in a tutorial class and split into control group (*Student2*) and an intervention group (*Student1*) to assess the impact on wear-time from a weekly SMS (text message) on Fridays, reminding the experimental group to wear their tracker. [4]. *Student3* and *Student4* datasets were from an observational study to learn about physical activity levels of undergraduate students [26]. These students were also recruited in a tutorial class.

To summarize, our datasets are diverse in terms of all the dimensions summarised in Table 5 as well as the details above. This makes it a rich collection for exploring our research questions.

4.2 Thresholds

Table 3 introduced several definitions and background for defining the thresholds for a valid day. As this is the first systematic analysis of adherence over a diverse collection of datasets, we restricted our analyses to just four carefully chosen valid day thresholds that have been used in previous research on wearing behaviours and in analysing activity tracking data:

- >0 steps as this is the simplest and least restrictive – used in [13, 26]
- >= 500 steps as another simple step criterion – used by [29, 30]
- 3-a-day, a measure of wear through the day [29, 30]
- >10 hours, the most stringent measure, requiring at least 10 hours, each with at least 1 step. [26, 31, 34]

The 10-hours adherence threshold has been used in health literature [27, 31, 37, 42] - although lesser ones have also been used (e.g. 8 hours [24]).

4.3 Analysis Conducted

We conducted the experiments to compare:

(1) %-age of *days* excluded using each of the 4 thresholds above;
(2) %-age of *people* whose median wear-time (hours per day) exceeds our >10 hours threshold, comparing this with the proportion whose wear-time was 6-9 hours and <6 hours.
(3) how the 4 thresholds affect calculated *activity levels*.

In each of these broad categories, we explored the impact of the across the datasets and tested for significant differences.

In our analysis, wear-time (hours per day) is calculated as the count of hours where at least 1 step was recorded. This approximation is needed because the trackers for our datasets do not distinguish a sedentary person wearing the tracker from a tracker that is not worn.

While our method is not exact for accounting non-wear, it serves as an estimate for comparison at a population level.

5 RESULTS

We now present the results of our analysis. The presentation is organised around our three research questions. First, we consider the impact of different adherence definitions on the data ignore (RQ1). Then we show how these definitions impact on results. In the discussion, we consider how these results point to ways to account for these results (RQ2 and 3).

5.1 Impact of Thresholds on Valid Days and Valid Weeks (RQ1)

At a high level, we would expect that the more restrictive a threshold is, the more data would be excluded from analysis of physical activity data. However, it is not obvious what differences will follow from different adherence definitions used to analyse datasets.

Differences in the Days Ignored for Different Adherence Measures

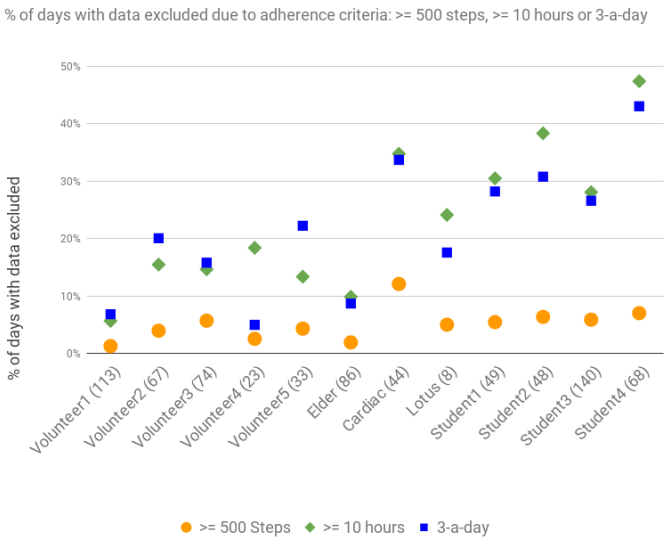

Fig. 1. Comparing %-age of days discarded, with valid day thresholds: >=500 steps, >=10-hours and 3-a-day, against >0 steps − days with no data

Figure 1 compares the %-age of days of data that are ignored for dataset. It takes the >0 steps measure as a baseline, since it is the least stringent measure. It shows how the other three measures compare against it. This and subsequent graphs order the datasets as in Table 5. This groups them according to

the broad characterisation of the datasets in terms of whether they were volunteered or part of a study that influenced participants to track. We include the sample size in the labels so the reader can see where the small size of a dataset may help explain the results.

First consider the >500 minimum step criterion (yellow circles). Overall, the graph shows that this criterion discards between 5% and 10% of days that are valid on the >0 steps criterion (mean: 5.1%, SD: 2.8%, 95% CI: ±1.8%). These are days with exceedingly modest use of the tracker (i.e., between 1 and 500); if a dataset has many of these, the reasons for this may deserve exploration to understand why people would often make so little use of a tracker or if it indicates problems with the device (such as problems with device calibration for people with specific mobility problems, and using a walking frame).

We now consider the *through the day* thresholds (green diamond for >= 10 hours and the blue square for 3-a-day). There are three striking trends here. First, the levels of data loss now have a far wider range, from 6% (*Volunteer1*) to 47% (*Student4*). Secondly, the data loss is always higher than the >=500 step threshold; these differences are significant (one-way ANOVA $F(2,33)=5.64$, $p<0.001$). Thirdly, both *through the day* thresholds are strikingly similar to each other for most datasets, also reflected in the mean %-age of days of data discarded showing no significant differences.

- mean: 21.5%, SD: 11.7%, 95% CI: ±7.4% - 3-a-day threshold
- mean: 23.4%, SD: 12.7%, 95% CI: ±8.1% - 10-hour threshold

One clear outlier is *Volunteer4*, the only case where 3-a-day is around 5% and much closer to the >= 500 steps and the 10 hours threshold is almost 20%. shown in Figure 1. This may be due to the small sample size (N=23) or variations at the individual level. As we would expect, these results show that the *through the day* thresholds, being stricter, may discard far more data. However, the differences varied widely across the datasets. The figure indicates much smaller differences for the *Volunteers* datasets, compared with the other datasets (Students plus Others - *Elder*,*Lotus*,*Cardiac*) and the difference is significant (2 sample t-test, $t_{10}=3$, $p=0.01$, 95% CI: ±12.6%):

- mean data loss: 13.5%, SD: 4.8%, 95% CI: ±5.9% for the *Volunteers* datasets;
- mean data loss: 30.4%, SD: 11.8%, 95% CI: ±10.9% for the other datasets (Students plus Others - *Elder*,*Lotus*,*Cardiac*).

The example in Figure 2 illustrates that this issue is also relevant when answering Goal-met questions. In this example, using a valid-day threshold of >10 hours, we show the percent of days that would be excluded (or considered insufficient data days) for each of our datasets. The green bars represent percentage of days where participants meet the 30 active minutes goal (i.e., defined as goal-met in Table 3). The yellow bars represent percentage of days where participants did not meet goal and recorded 10 hours or more data (i.e., goal-not-met). The grey bars represent the percentage of days where participants did not meet the goal but also did not record 10 hours of data (i.e., insufficient data). The figure shows very wide differences in the percentage of insufficient data days across different datasets from 6% for *Volunteer1* to over 52% for *Student4*.

Differences in the People Ignored for Different Adherence Measures

We now move from analysis of the *days discarded* and consider the effects for the %-age of *people* whose data is discarded. Figure 3 delves into this for the >=10 hour threshold, considering three bands of median wear-time (the count of hours with at least one step). The bottom green part of each bar is the %-age of people with median >=10 hour threshold. The next, orange part enables us to see the potential impact of a less stringent threshold, as it is the %-age of people with median wear-time >= 6 and < 10 hours and the top, grey part is for < 6 hours. Overall, our analysis shows a very different profiles of wear-time between different datasets. For example, a dataset like the *Student4*, where only about half

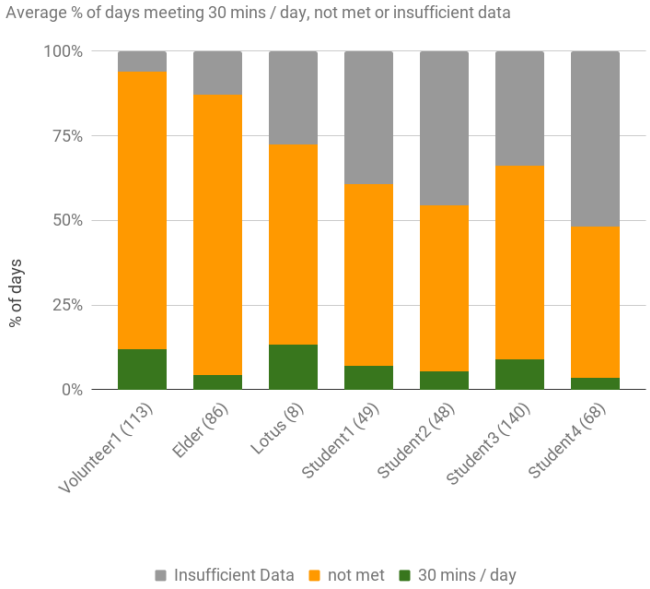

Fig. 2. An example of a Goal-met report showing percentage of days that met 30 active minutes, not met or had insufficient data. Note: using 10 hours valid threshold to determine insufficient data days. Note: datasets *Volunteer2*, *Volunteer3*, *Volunteer4*, *Cardiac* and *Volunteer5* are not included due to lack of per minute data. N of each sample included in X-axis label.

the participants have wear-time above the threshold of 10 hours, it may be worthwhile exploring less stringent thresholds, and so include more of the population.

Continuing our focus on *people* whose data is discarded, Figure 4 shows an analysis of weekly adherence, the number of days a week that were valid (using the 10 hours valid day threshold). The *Volunteers* datasets were dominated by people with high weekly adherence, with almost half of them averaging 7 valid days a week. The Others datasets (*Elder*, *Cardiac* and *Lotus*) have a flatter distribution but still have 30% of people with 7 valid days a week. By contrast, the Students (red squares) have a very different profile of weekly adherence, with very few reaching 7 valid days a week. This is reflected in the weekly adherence across the groups:

- *Volunteers*: mean 5.5 days, SD:1.8, 95% CI: ±0.21.
- Others (*Elder*,*Cardiac*,*Lotus*): mean 4.7 days, SD:2.2, 95% CI: ±0.38.
- Students: mean 2.8 days, SD:1.9, 95% CI: ±0.21.

We repeated this analysis for the least restrictive >0 steps threshold. We found that overall, this more relaxed threshold gave 1.5 (21%) more valid days per week.

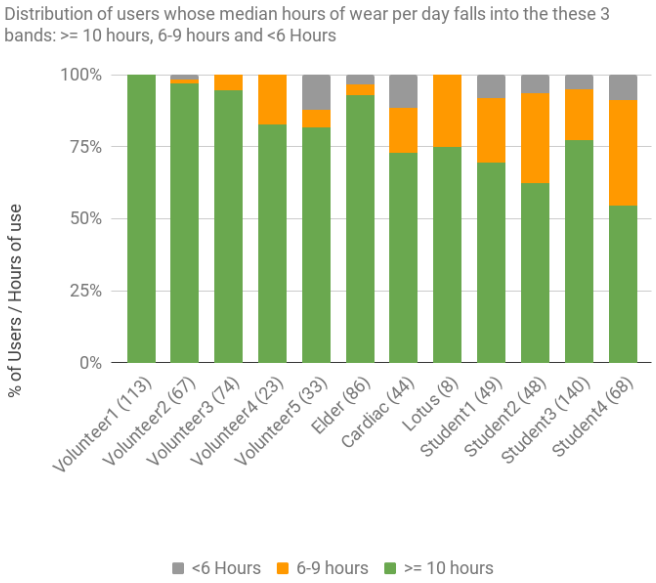

Fig. 3. Comparison of %-age of users with median wear-time >=10 hours, >= 6 and < 10 hours and < 6 hours. N included in X-axis label.

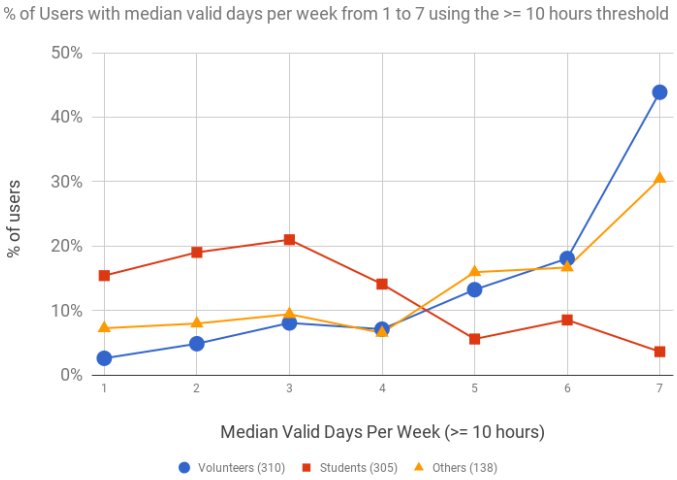

Fig. 4. Comparison of users median weekly adherence (no. of valid days per week). Grouped as follows: *Volunteers* (*Volunteer1*, *Volunteer2*, *Volunteer3*, *Volunteer4*, *Volunteer5* N=310), Students (*Student1*, *Student2*, *Student3*, *Student4* N=305), Others (*Elder*, *Cardiac*, *Lotus* N=138)

5.2 Exposing Uncertainty: The Impact of Threshold Methods on Activity Level Reporting (RQ2 and 3)

In this section, we show how the adherence measure can affect interpretation of the activity data. We do this in two activity levels, average step counts per day and average active minutes per day.

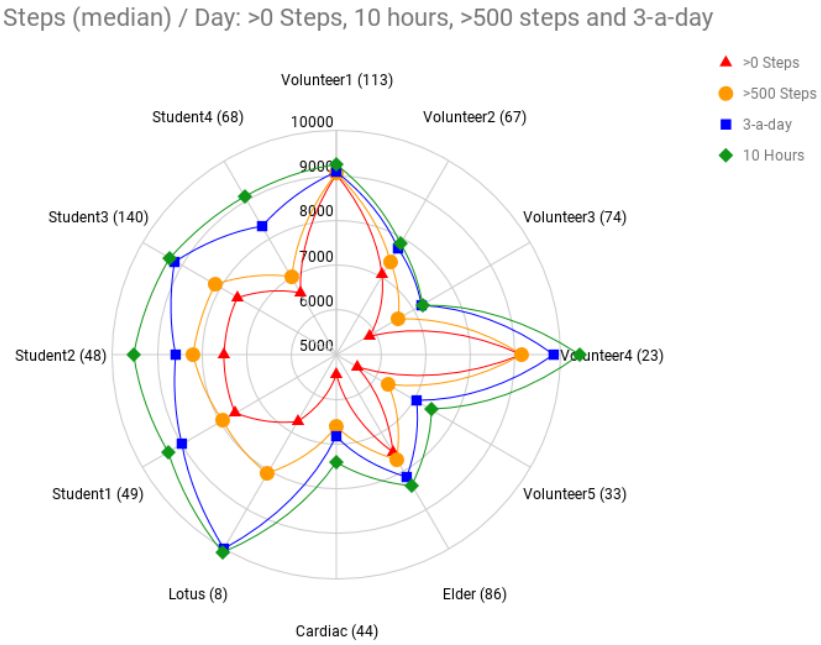

Fig. 5. Comparison of median steps across populations, showing impact of different valid day thresholds. N included in dataset labels.

Figure 5 shows the median steps per day when calculated against the 4 adherence thresholds. For each dataset, it shows, from the top, 10 hours (top end of line), 3-a-day (top of box), >=500 steps (bottom of box) and >0 Steps / day (lower end of line) [1]. This figure indicates how results might differ depending on the adherence definition used. The clear picture that emerges is that the impact varies considerably across the datasets.

For example, across all datasets, the mean steps for each adherence definition are:

- 7133 steps, >0 steps (SD: 1197, 95% CI: ±761)
- 7682 steps, >=500 steps (SD: 926, 95% CI: ±588).
- 8415 steps, 3-a-day (SD: 1060, 95% CI: ±673)
- 8779 steps, >10 hours (SD: 1090, 95% CI: ±692)

[1] We used a candlestick-like visualisation (or box plot) to convey the spread between the datasets. While it is theoretically possible for steps count for the 3-a-day threshold to be higher than 10 hours or lower than the 500 steps, this is not the case in any of our datasets. So, this format gives a compact summary of our analyses.

There was no significant difference between the two step-count thresholds (>0, >=500), nor between the *through the day* thresholds (3-a-day and >10 hours). However, the step count using the least stringent threshold (>0 steps) is 1,646 less steps (23%) than using the most stringent (>=10 hours). This significant difference is quite large in terms of absolute activity level difference (paired t-test, t_{11}=6.9, p<0.0001, 95% CI: ±527).

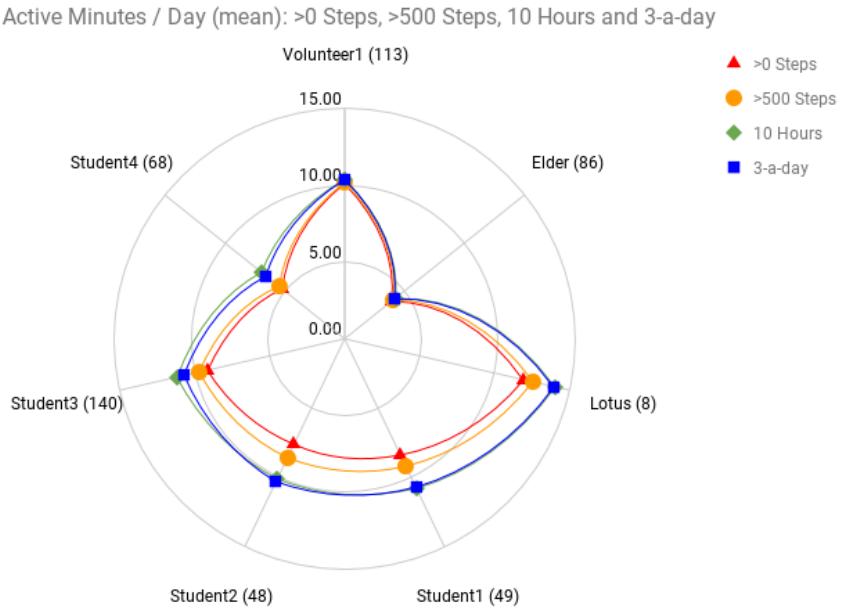

Fig. 6. Comparison of active minutes results across populations, showing the impact of different valid day thresholds. Datasets *Volunteer2*, *Volunteer3*, *Volunteer4*, *Cardiac* and *Volunteer5* are not included due to lack of per minute data. N included in dataset labels.

We also compared >0 steps and 10 hours thresholds for *Volunteers* datasets and found no significant differences. However, when we examined this with others datasets (i.e., Students plus Others -*Elder*,*Cardiac*,*Lotus*), there was an average of 2,020 steps difference between step counts using the 2 thresholds (paired t-test, t_6=6.9, p<0.001, 95% CI: ±718).

Figure 6 shows a similar analysis, now for active minutes per day [2]. Similar to our analysis of steps counts, Students and Others (*Elder*,*Lotus*) datasets had significantly higher active minutes (1.9 minutes more) using the 10 hours threshold compared to the >0 steps threshold (paired t-test, t_5=5.9, p=0.002, 95% CI: ±0.81).

[2]To calculate active minutes calculations, we used 120 steps per minute, commonly used to calculate moderate-to-vigorous physical activity (MVPA) [41]. Some datasets were excluded from this analysis due to lack of per minute activity tracker data (i.e., *Volunteer2*, *Volunteer3*, *Volunteer4*, *Volunteer5* and *Cardiac*)

6 DISCUSSION

In the introduction, we presented examples of core Activity-level and Goal-met questions, both for individuals and in aggregate. As a foundation for our discussion, we now introduce the following questions – these refer to two hypothetical long-term data sets, called *Dataset1*, *Dataset2*.

(1) In *Dataset1*, what were people's average daily steps counts in 2014 and 2017?
(2) Were people in *Dataset1* more active than those in *Dataset2* in 2017?
(3) For the goal of 120 minutes a week of moderate activity, what percentage of *Dataset1* people met the goal in 2014 and 2017?

The first is an Activity-level question, to compare activity level *within* a dataset. The second question is similar to the first, but involves comparisons *between* datasets. The third is similar to the first, but it is for a Goal-met question. These go beyond the analyses we have reported but are useful for broadening the scope of our discussion, building on the reported work. We refer to these questions in the discussion, which starts with the key insights for our three research questions. Building from this, we present a set of recommendations for analysing physical activity data. We then briefly discuss the diverse goals for activity tracking. Finally, we discuss the limitations of our work, along with future directions for research to support systematic and effective accounting for adherence in physical activity data and other personal sensor data.

6.1 Key Insights for the Research Questions

Our study showed very diverse impacts of the four core adherence measures across the 12 datasets, with *significant* and *important* differences in *%-age of days discarded* and *%-age of people ignored* (RQ1).

In terms of the number of days of data discarded by the thresholds (Figure 1):

- both minimum step-measures, >0 days and > 500 steps were similar for most (but not all) datasets;
- the through-the-day measures, >10 hour and 3-a-day, had diverse impacts – similar on some datasets, quite different for others;
- both through-the-day measures ignored more data than the minimum step-measures.
- these through-the-day measures had quite diverse effects across the datasets.

Some of our datasets had consistently high adherence on all these measures. This was the case for *volunteers* datasets and *Elder* (shown in Figures 1 and 3). For these populations, strict adherence thresholds cause very little loss of days or people.

Our 12 datasets represent considerable diversity in terms of many factors, including the way the data was acquired, the purposes for which it was collected, the ages of participants, the size of the cohort and the duration. Some trends in the adherence levels are:

- The volunteers datasets tended to have higher adherence than the others. This seems likely to be due to the recruitment methods, tending to attract committed trackers.
- The student datasets tended to have the lowest adherence, perhaps because of the complex factors affecting the students' enthusiasm for tracking and interest in their physical activity levels: they were invited to participate as part of class practical work and the Fitbits were on loan, just for the study period.

Based on the diversity of results across the datasets, and within the groupings (volunteered, students and others), the main observation is that one needs to analyse any new dataset, using all four adherence criteria, to see the impact of each. Our results cannot be used to predict the adherence to be expected for a new dataset.

This impacts calculations of Activity-level questions (RQ2). For example, in the combined student datasets, the >0 steps adherence measure gave a daily step count of 6,952 where the >=10 hours gave 9,423, 35% higher (see Figure 5). From a health perspective, this is an important difference. Corresponding to this, the different adherence measures had very different impacts on the data ignored (RQ1). The %-age of *days ignored* moved from 32% to 60% (see Figures 1 and 5). In terms of people ignored, the >10 hours per day measure excludes more than a third of students (34%), compared with days with >0 steps (see Figure 3). For weekly adherence, less than a third of students (32%) averaged 4 days or more per week with >10 hours (see Figure 4). This highlights the challenge in interpreting such data to draw conclusions about Activity-level and Goal-met questions. It indicates the potential for introducing *bias* by using an adherence measure that excludes many low adherence users. There is a similar picture for Goal-met questions (RQ3), For example, for the combined students dataset, the >10 hours threshold would ignore 43% of days. our work suggests that there are no easy answers to interpreting such questions but it does point to the importance of reporting adherence measures and their impact along with inferred answers to questions.

6.2 Recommendations for Systematic Analysis of Activity Data

1: Establish a suitable set of adherence measures to consider. This first step in analysing physical activity data calls for identifying potential adherence definitions. This paper focused on these:

- >0 steps − the least stringent measure;
- >500 steps − to exclude days with very little data;
- >10 hours − a very stringent measure requiring wear for many hours of the day;
- 3-a-day − similar to >10 hours, but requiring wear in the 3 time periods.

We recommend that these four be considered because they range from minimal to quite stringent and these have been used in previous work, providing an initial set of comparative data. There are many other possibilities, and we have touched on some which should also be considered, such as weekly adherence >= 6 days a week, more complex weekly adherence of at least 1 weekend-day and at least 4 weekdays.

2: explore the impact of the adherences measure(s). For each adherence measure considered, replicate our analyses to assess its impact in terms of:

- the %-age of *days* ignored
- the %-age of *people* ignored.

Then follows the actual analyses of the data, based on the chosen set of adherence measures, to determine the answers to the core questions, such as the examples at the beginning of this section.

3: Report adherence along with results of data analysis. For many contexts, analysis of tracker data needs to provide a single answer to a question. Even in these cases, we recommend that this result is reported along with:

- the adherence measure used;
- the results of the analyses for this measure in Step 2;

These recommendations provide a way to enhance trust and confidence in information by presenting information about accuracy. In the case of research reports, where more information can be provided, we also recommend providing an explanation for the choice of the adherence measure as well as details of the fuller analyses from Step 2. Over time, this would make it easier for researchers to compare reported results across the literature.

Examples of Accounting for Adherence in Reporting Personal Informatics Results

As discussed earlier, failure to properly account for incompleteness can lead to a loss of trust and ultimate usefulness of such systems [3, 9, 17]. We now consider how our work can help address this problem by enabling a user to consider the impact of adherence when they try to interpret their physical activity data. For this case, we recommend making it possible for the user to:

(1) see the adherence measure used;
(2) have the opportunity to select the adherence measure the user prefers, based on their own knowledge of their adherence.

Here are two examples of reporting activity levels along with adherence measures.

(1) You had 4,500 steps yesterday **in 6 hours with data.**
(2) You averaged 9,500 steps a day in the last month **based on the 17 days that you had 10-hours with step data.**

The first example illustrates a way to report steps when a user did not have enough data to qualify as a valid-day (e.g., only 6 of 10 hours). By adding the adherence information, a user can see an indication that the step count may be an under-estimate. The second example illustrates how to report data over a period of time. The next two examples apply these principles for comparisons.

(3) You were less active this summer than last year. **Based on days with at least 10 hours of wear. This is 20% of all your summer days.**
(4) You are in the top 20 percentile of your peer group, by age and gender, **Based on days with at least 10 hours of wear. This is 50% of your days and 30% for the peers**.

These examples enable a user to decide whether they trust the result, based on the %-age of days included and their beliefs about their tracker use.

The next example illustrates a goal-met result:

(3) You met your goal of at least 30 active minutes a day on 3 days last week. **You were under it on 2 days that had 10-hours of data and the other 2 days had less than 10-hours of data.**
(4) You are in the top 20 percentile of your peer group, by age and gender, **Based on days with at least 10 hours of wear. This is 50% of your days and 30% for the peers**.

6.3 Reasons for Collecting and Analysing Activity Data and Implications for Accounting for Adherence.

Motivations for Collecting Tracking Data

There are many possible reasons for a person to track physical activity, producing a dataset. Tracking may be *initiated by the individual*. Increasingly, smart-phones automatically collect activity data, often without the owner being aware. But much tracker data is purposefully collected. This form of tracking has been the focus of Ubicomp's *Personal Informatics*, as well as the Quantified Self communities [7–9, 12, 25, 32, 33]. That research reports various motivations for such tracking, such as self-monitoring, self-reflection and conducting n-of-1 experiments [7, 10] and with suitable interfaces, to support behaviour change [9]. (Our volunteer data is most similar to this work.) Beyond this, tracking may be initiated as *part of an intervention* (*Elder*, *Cardiac* and *Lotus*) perhaps on advice of a medical professional or in a *study of a population* (like our student datasets). People may use trackers for short periods, for example just for a week to establish a baseline and then at later points, to assess the effects of the intervention, as in [35]. Others track consistently over long period. In all these cases, the answers to questions such as the examples at the beginning of this section demand a solid adherence analysis such as we have

recommended. This can be the basis for assessing the accuracy of the results and if multiple adherence measures are used and reported on, this will support comparisons in the literature.

Using Adherence Measures to Change Wearing Behaviour Versus Using Adherence Measures to Make Sense of Available Data

If adherence measures are reported along with results, as in the examples above, this may give individuals information to help them re-consider their adherence levels. For example, Tang et al [34] reported that some of their participants said that they planned to be more adherent. But a core goal of our work is to provide foundations to harness available activity data, regardless of whether the adherence is low, intermittently high, consistently high or any other pattern. Even with lower adherence, if the adherence is reported, people may well gain valuable answers to their questions, along with information to assess the reliability.

6.4 Limitations and Future Work

The driver for our study design was to provide foundations for accounting for adherence and this drove decisions about each element of the study design. We now discuss the limitations of our work. If our recommendations are followed, we hope to see future work adding to our results for each of these aspects and we give pointers to these. We conclude with future directions for smart-phone data, drawing on our work.

Questions explored. Our three research questions explored the impact of adherence measures and how to account for adherence when answering questions about physical activity. We began with questions about pure Activity-level and Goal-met questions. At the beginning of this section, we introduced similar questions for comparisons *within* and *between* datasets. These are basic questions, representing just a starting point in analysing physical activity data.

Adherence definitions and analyses explored. To manage the complexity of results to present, we carefully selected just 4 valid day definitions, two each, *minimum-step-count* and *through-the-day*. In the case of minimum-hours-in-the-day, we focused on 10-hours a day because it is common in literature, but we did explore the impact of considering <6 hours and 6-9 hours. Similarly, the set of experiments we report was carefully chosen to explore our research questions and enable a reader to see the key results showing the importance of definitions of adherence.

Datasets: participants. while our datasets are large and diverse, they all do come from authors of the paper. It will be valuable to see this work replicated on other data sets. Our results show that our volunteer and student categorisation does show some commonalities within these categories. There are many dimensions that may be important for describing key characteristics of the people in a dataset, such as gender, age, health status and importantly motivation for collecting the data. Associated with these are characteristics such as the duration of data collection.

Datasets: activity tracking devices. Our datasets all come from similar trackers, although people had different variants on these trackers. None of our devices can distinguish inactivity from non-wear. Many current tracker devices can do this (for example, [22]). Even in our datasets, some participants may have changed devices through the time. We ignored these issues in our analyses. To answer the three questions at the beginning of this discussion, device differences need to be considered. There is also a broader discussion around accuracy of trackers for reliability and comparability [16, 31, 37, 39].

Using Smart-phone physical activity data. Another important *wearable* (or carried) device is the mobile phone. Our work provides a foundation for identifying meaningful measures of adherence for tracking physical activity as captured by a mobile phone. For people who always wear/carry their phones when awake, and so having high adherence, the phone could provide a reliable way to track activity. However, many people do not do this. Yet a recent large study assumed it was reliable to compare populations, based on assessing daily wear-time from the first to last phone use in the day [2]. On this basis, they reported a mean step of 5,039 across 111 countries and 717,527 users and compared step counts by country. There is no indication of incompleteness in their data. Are comparisons between different countries based on comparable daily adherence (e.g., 80% US versus 70% Australia or 80% US versus 20% China)? Our work suggests that it would be valuable to also calculate the step counts for other adherence measures that are well suited to a mobile phone. An adaptation of our >10 hours measure seems a good starting point.

Our work has implications for other inferences from wearable devices that measure many other things, such as heart-rate, stress, sleep quality, air-quality. We see an important role for adherence measures in the ubicomp field when applications seek to combine and provide self monitoring using data from different classes of devices and over time. As tracking capabilities, devices and sources of data change, so too would the wearing and adherence patterns of the data people have collected. This is true for both aggregate data and personal informatics.

7 CONCLUSION

We reviewed both health and computing literature to establish definitions of *adherence*, a measure of incompleteness, to help answer two important classes of health questions: Activity-level and Goal-met.

We also introduced a new adherence measure *Valid-Goal-Day* needed when answering Goal-met based health questions. Our analysis of 12 large and diverse physical activity tracker datasets showed that previous threshold based methods of addressing incompleteness is not appropriate for large scale *volunteers* datasets such as those collected through personal use e.g., Fitbit users. When dealing with *volunteers* datasets, we recommend to analyse and also to report adherence measures for individual applications. Further research is needed on how the information can best be delivered. For aggregate reports, it is also important to report adherence criteria used along with the adherence measures and their impact on the activity level results.

REFERENCES

[1] Laura Albert. 2017. The Surprising Potential Fitness Tracker Buyer. (Aug 2017). https://civicscience.com/surprising-potential-fitness-tracker-consumer/ [Online; posted 21-July-2016 CivicScience].

[2] Tim Althoff, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp, and Jure Leskovec. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* (2017). https://doi.org/10.1038/nature23018

[3] F Bentley, K Tollmar, and P Stephenson. 2013. Health Mashups: Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. *Tochi* 20, 5 (2013), 1–27. https://doi.org/10.1145/2503823

[4] Kevin Alexander Bragg. 2015. *Does The Quantified Self Equal Quantified Health?* Ph.D. Dissertation. University of Sydney.

[5] Lisa Cadmus-Bertram, Bess H Marcus, Ruth E Patterson, Barbara A Parker, and Brittany L Morey. 2015. Use of the Fitbit to Measure Adherence to a Physical Activity Intervention Among Overweight or Obese, Postmenopausal Women: Self-Monitoring Trajectory During 16 Weeks. *JMIR mHealth and uHealth* 3, 4 (2015), e96. https://doi.org/10.2196/mhealth.4229

[6] Lisa A. Cadmus-Bertram, Bess H. Marcus, Ruth E. Patterson, Barbara A. Parker, and Brittany L. Morey. 2015. Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine* (2015). https://doi.org/10.1016/j.amepre.2015.01.020

[7] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'17). ACM, New York, NY, USA*, Vol. 10.

[8] Eun Kyoung Choe, Nicole B Lee, Bongshin Lee, Wanda Pratt, and Julie A Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1143–1152.

[9] Sunny Consolvo, Predrag Klasnja, David W McDonald, James A Landay, et al. 2014. Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness. *Foundations and Trends® in Human–Computer Interaction* 6, 3–4 (2014), 167–315.

[10] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons Learned from Two Cohorts of Personal Informatics Self-Experiments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 46.

[11] Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T van Hees, Michael I Trenell, Christoper G Owen, et al. 2017. Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank Study. *PloS one* 12, 2 (2017), e0169649.

[12] Chris Elsden, David S Kirk, and Abigail C Durrant. 2015. A Quantified Past: Toward Design for Remembering With Personal Informatics. *Human–Computer Interaction* (2015), 1–40.

[13] Daniel A Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A Munson. 2016. Beyond Abandonment to Next Steps: Understanding and Designing for Life After Personal Informatics Tool Use. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), 1109–1113. https://doi.org/10.1145/2858036.2858045

[14] Daniel A. Epstein, Jennifer Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the Device in the Drawer: Lapses as a Design Opportunity in Personal Informatics. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 829—-840.

[15] Daniel A Epstein, An Ping, James Fogarty, and Sean A Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 731–742.

[16] Kelly R Evenson, Michelle M Goto, and Robert D Furberg. 2015. Systematic review of the validity and reliability of consumer-wearable activity trackers. *The international journal of behavioral nutrition and physical activity* 12, 1 (2015), 159. https://doi.org/10.1186/s12966-015-0314-1

[17] B J Fogg. 2003. *Persuasive Technology: Using Computers to Change What We Think and Do*. The Morgan Kaufmann series in interactive technologies, Vol. 5. Morgan Kaufmann. 283 pages. https://doi.org/10.4017/gt.2006.05.01.009.00

[18] Thomas Fritz, Elaine M Huang, Gail C Murphy, and Thomas Zimmermann. 2014. Persuasive technology in the real world: a study of long-term use of activity sensing devices for fitness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 487–496.

[19] Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2015. How do we engage with activity trackers?: a longitudinal study of Habito. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1305–1316.

[20] William L Haskell, I Lee, Russell R Pate, Kenneth E Powell, Steven N Blair, Barry A Franklin, Caroline A Macera, Gregory W Heath, Paul D Thompson, Adrian Bauman, and Others. 2007. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Medicine and science in sports and exercise* 39, 8 (2007), 1423.

[21] IDC. 2017. Worldwide Quarterly Wearable Device Tracker. (Aug 2017). https://www.idc.com/tracker/showproductinfo.jsp?prod_id=962

[22] Hayeon Jeong, Heepyung Kim, Rihun Kim, Uichin Lee, and Yong Jeong. 2017. Smartwatch Wearing Behavior Analysis : A Longitudinal Study. *ACM Ubicomp 2017 / Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (2017). https://doi.org/10.1145/3131892

[23] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.

[24] K Konstabel, T Veidebaum, V Verbestel, L A Moreno, K Bammann, M Tornaritis, G Eiben, D Molnar, A Siani, O Sprengeler, N Wirsik, W Ahrens, and Y Pitsiladis. 2014. Objectively measured physical activity in European children: the IDEFICS study. *Int J Obes (Lond)* 38 Suppl 2, S2 (2014), S135–43. https://doi.org/10.1038/ijo.2014.144

[25] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2012. Using context to reveal factors that affect physical activity. *ACM Transactions on Computer-Human Interaction* 19, 1 (2012), 1–21. https://doi.org/10.1145/2147783.2147790

[26] Judy Tang Lie Ming and Kay. 2016. Daily & hourly adherence : towards understanding activity tracker accuracy. *CHI '16 Extended Abstracts on Human Factors in Computing Systems* (2016).

[27] Charles E. Matthews, Kong Y. Chen, Patty S. Freedson, MacIej S. Buchowski, Bettina M. Beech, Russell R. Pate, and Richard P. Troiano. 2008. Amount of time spent in sedentary behaviors in the United States, 2003-2004. *American Journal of Epidemiology* 167, 7 (2008), 875–881. https://doi.org/10.1093/aje/kwm390 arXiv:NIHMS150003

[28] Jochen Meyer, Wilko Heuten, and Susanne Boll. 2016. No Effects But Useful ? Long Term Use of Smart Health Devices. *Ubicomp/ISWC'16 Adjunct* (2016), 516–521. https://doi.org/10.1145/2968219.2968314

[29] Jochen Meyer, Jochen Schnauber, Wilko Heuten, Harm Wienbergen, Rainer Hambrecht, Hans-Jürgen Appelrath, and Susanne Boll. 2016. Exploring Longitudinal Use of Activity Trackers. *Procedings of IEEE ICHI - International Conference on Healthcare Informatics* (2016), 198–206. https://doi.org/10.1109/ICHI.2016.29

[30] Jochen Meyer, Merlin Wasmann, Wilko Heuten, Abdallah El Ali, and Susanne Boll. 2017. Identification and Classification of Usage Patterns in Long-Term Activity Tracking. *CHI '17 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2017). https://doi.org/10.1145/3025453.3025690

[31] Jairo H. Migueles, Cristina Cadenas-Sanchez, Ulf Ekelund, Christine Delisle Nystrom, Jose Mora-Gonzalez, Marie Lof, Idoia Labayen, Jonatan R. Ruiz, and Francisco B. Ortega. 2017. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Medicine* (2017), 1–25. https://doi.org/10.1007/s40279-017-0716-0

[32] Amon Rapp and Federica Cena. 2016. Personal Informatics for Everyday Life: How Users without Prior Self-Tracking Experience Engage with Personal Data. *International Journal of Human-Computer Studies* 94 (2016), 1–17. https://doi.org/10.1016/j.ijhcs.2016.05.006

[33] John Rooksby, Mattias Rost, Alistair Morrison, and Matthew Chalmers Chalmers. 2014. Personal tracking as lived informatics. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (2014), 1163–1172. https://doi.org/10.1145/2556288.2557039

[34] Lie Ming Tang and Judy Kay. 2017. Harnessing Long Term Physical Activity Data: How Long-term Trackers Use Data and How an Adherence-based Interface Supports New Insights. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (jun 2017), 26:1—-26:28. https://doi.org/10.1145/3090091

[35] Anne Tiedemann, Leanne Hassett, and Catherine Sherrington. 2015. A novel approach to the issue of physical inactivity in older age. *Preventive medicine reports* 2 (2015), 595–597.

[36] Fumiharu Togo, Eiji Watanabe, Hyuntae Park, Akitomo Yasunaga, Sungjin Park, Roy J. Shephard, and Yukitoshi Aoyagi. 2008. How many days of pedometer use predict the annual activity of the elderly Reliably? *Medicine and Science in Sports and Exercise* 40, 6 (2008), 1058–1064. https://doi.org/10.1249/MSS.0b013e318167469a

[37] Stewart G. Trost, Kerry L. Mciver, and Russell R. Pate. 2005. Conducting accelerometer-based activity assessments in field-based research. *Medicine and Science in Sports and Exercise* 37, 11 SUPPL. (2005), 531–543. https://doi.org/10.1249/01.mss.0000185657.86065.98 arXiv:arXiv:1011.1669v3

[38] Jared M. Tucker, Gregory J. Welk, and Nicholas K. Beyler. 2011. Physical activity in U.S. adults: Compliance with the physical activity guidelines for Americans. *American Journal of Preventive Medicine* 40, 4 (2011), 454–461. https://doi.org/10.1016/j.amepre.2010.12.016

[39] Catrine Tudor-locke. 2016. The Objective Monitoring of Physical Activity: Contributions of Accelerometry to Epidemiology, Exercise Science and Rehabilitation. (2016). https://doi.org/10.1007/978-3-319-29577-0

[40] C Tudor-Locke, D R Bassett, A M Swartz, S J Strath, B B Parr, J P Reis, K D Dubose, and B E Ainsworth. 2004. A preliminary study of one year of pedometer self-monitoring. *Ann Behav Med* 28 (2004). https://doi.org/10.1207/s15324796abm2803_3

[41] Catrine Tudor-Locke, Cora L Craig, Wendy J Brown, Stacy A Clemes, Katrien De Cocker, Billie Giles-Corti, Yoshiro Hatano, Shigeru Inoue, Sandra M Matsudo, Nanette Mutrie, Jean-Michel Oppert, David A Rowe, Michael D Schmidt, Grant M Schofield, John C Spence, Pedro J Teixeira, Mark A Tully, and Steven N Blair. 2011. How many steps/day are enough? for adults. *International Journal of Behavioral Nutrition and Physical Activity* 8, 1 (jul 2011), 79. https://doi.org/10.1186/1479-5868-8-79

[42] Catrine Tudor-Locke, Yoshiro Hatano, Robert P. Pangrazi, and Minsoo Kang. 2008. Revisiting "how many steps are enough?". *Medicine and Science in Sports and Exercise* 40, 7 SUPPL.1 (2008). https://doi.org/10.1249/MSS.0b013e31817c7133

Received August 2017; revised November 2017; revised January 2018; accepted January 2018