Title

Workgroup 2 - **Data Flow** *Recommendations from the Digital Sensing Workshop held at UCLA Feb 28-March 2, 2023*

Author

Digital Sensing Workshop Participants (https://ucla.box.com/v/dig-sensing-wkshp-pubroster)

Publication Date

2023-05-24

DATA FLOW

Data collection from sensors to analysis, including data definition and data standards

WORKGROUP #2 – COMPILED RECOMMENDATIONS

These recommendations were compiled as an output of the Advancing the Utility of Digital Sensing Tools for Mental Health Research workshop ("Digital Sensing Workshop") sponsored by the UCLA Depression Grand Challenge, Wellcome Trust and NIMH. Workshop participant roster may be found at: https://ucla.box.com/v/dig-sensing-wkshp-pubroster.

Please submit feedback at https://bit.ly/dig-sensing-report-feedback by August 31, 2023.

Working group 2: Data Flow (data collection from sensors to analysis, including data definition and data standards)

A list of priorities for each topic area & recommendations for addressing the priorities (as represented in materials in the reporting back sessions)

Resources

Pre-workshop material
Miro board with stickers and upvotes and spreadsheet with the same information: the participant-generated content for discussion is referenced throughout the report

OUTLINE/DRAFT

1. Introduction

To analyze and address the challenge areas identified in the pre-workshop material (data comparison, data & metadata standards, data analysis pipelines and data sharing), the group adopted a timed brainstorming approach using stickers to contribute answers & ideas, and colored dots to upvote content others agreed with.

We started by looking at the positive contributions provided by the development and wide availability of sensor-based digital solutions to address wellness, healthcare and research. The generation of large amount of physiological and behavioral data in everyday settings, close to the person's lived experience has opened possibilities and also brought challenges, which the group identified as problem areas

1. Data privacy, including the risks of re-identifiability
2. Lack of transparency in the way certain sensor-based digital solutions work (black boxes) and limited access to base ("raw") data
3. Paucity of funding (beyond that for digital solution development) particularly for the creation of the infrastructure needed for data analysis and data sharing
4. Adoption of existing standards and development of additional ones for nomenclature, data and metadata representation, provenance, analysis.

Finally, the group identified possible solutions to address the problem areas and considered next steps to devise concrete pathways to start implementing such solutions.
Below we provide additional details regarding the discussion around the problem areas listed above.

2. Key Problem Areas and Proposed Solutions

2.1 Privacy and Re-identification

The properties of digital health technologies that make them great for patient-centric passive sensing also create new challenges for data privacy.

Researchers analyzing data from digital health technologies are not necessarily equipped to adequately de-identify the data and such data is not currently treated as patient health information where more guidance exists in terms of de-identification and management. The risk of re-identifiability of data collected via digital health technologies has been documented. Techniques to reduce such risk tend to make the data less precise and so less useful for research. In contrast, there is the need for more shared datasets.

While it has been shown people are willing to share their digital health data for research purposes, constructing a precise and protective informed consent is a challenge, given the granularity of data, the extent of data collection and also the potential for extraneous people to be included in it (e.g., in case of image or voice recordings). The more complex the informed consent, the more complex its implementation, with the need to keep track of what data has been consented for sharing, for how long, with whom, and what effect the consent has on features derived from a given data type (see also section 2.4 below).

The current approach (public vs. private, sensitive vs. non-sensitive data, etc.) is no longer sufficient in this more nuanced reality. There is a need for inclusion of study participants and patients to highlight and address key concerns, define best practices[1] and ensure that there is trust between the parties involved.

Solution Space

To address the problem space of privacy, proposals include:

- Developing tools to check for re-identifiability of data and models
- Creation of a standard for informed consent (see also section 2.4 below) and of a formal model of informed consent verification
- Tracking of consent and privacy context as metadata
- Investigate scenarios of problems with data leaks from the worst case to the likeliest one
- Soliciting input from other high-security fields like finance, banking, and cybersecurity
- Continuous evaluation of impact of solutions in privacy
- Working with study participants and users to develop privacy practices (privacy by design)
- Meeting with digital health solutions' manufacturers to discuss sensitive data storage

(Additional proposals can be viewed on the spreadsheet.)

1. Shilton, K., et al. (2021). Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211040759

2.2 Devices are black-boxes

Wearable devices and mobile health technologies are being used because of the amazing pace of innovation in the space. The variety and quality of sensors being deployed continue to increase and improve. Consumer wearable devices offer the promise that if we can research the data they generate, findings can be applied to the massive pool of existing users.

However, data generated from consumer-grade devices are often pre-processed, and only derived metrics are reported. These derived metrics can also change over time as the algorithm to create them is updated. While this is useful for consumers by giving them the latest report on their behavior, it poses a serious challenge for researchers attempting to use these data to establish reproducible and consistent findings across devices and over time. Research-grade devices such as the GENEActiv or Empatica watches can report data in more research-friendly formats (e.g., simple CSV tables), have consistent outputs, use published algorithms for processing, and have established protocols for measurement. Much of the existing literature on the processing and utility of actigraphy data, for example, is done using such research-grade devices.

And so we are left with a question as a research community: should we ever use commercial wearables for research, and what are the requirements for devices such that we may use them for scientific research?

It is difficult to entirely discount the use of consumer wearables, as they can capture behavior and interactions in a manner that research products tend not to. For example, consumer devices tend to be used to maximize adherence and utilization with clean and tested user interfaces. They are built to be integrated into people's daily lives in ways that drive continued use through observed utility. In addition, they are constantly being updated and upgraded with new sensors and abilities. For downstream intervention or monitoring purposes, consumer devices are where many results would ultimately be deployed.

With a large variety of sensors, changing algorithms, and sparse documentation, there is a need for a benchmark to compare results from wearable devices to understand their quality and if they are fit for a given research purpose. To enable benchmarking, researchers require a certain level of transparency and access to minimally processed data from wearable device creators. For example, in actigraphy, sleep is measured from algorithms applied to accelerometer data, and so for analysis, access to raw or minimally processed (e.g., resampled) accelerometer data is required. Benchmarking would also require clear documentation on hardware and software versions to enable fair comparisons. Government and in particular the NIMH may have a role to play in leveraging their existing data repositories capable of handling large digital datasets to be the hosts of these benchmarks.

The goals of researchers imply a set of requirements for device developers:

1. Research access to minimally processed data
2. Clear versioning and documentation on devices and software

3. Specification of sensor performance and reliability under certain conditions (i.e., similar to spec sheets from electronic components)

Along these lines, there may be space for a research or medical-grade designation of an application collecting data from a device that might be separate from the device itself. This would be in line with how Apple's ECG application is regulated, but the Apple Watch itself is not a medical device. Certain requirements and reporting standards can then be applied to the data collected via the application (including hardware/software versioning) that wouldn't impose overly restrictive criteria for the whole device. This is an area in which regulatory and government agencies may help develop guidelines borrowing from existing work such as requirements for digital biomarkers in the Digital Medicine Society's (DiMe) EVIDENCE checklist.

Ultimately, however, it is the research team's responsibility to justify their choice of a given device as fit-for-purpose. Research teams should explain why they are using a device and how it matches the data required for their scientific question of interest. Transparency around devices and the existence of benchmark datasets would aid in this justification process. While developing contracts with device companies, it may be clearer to have *data contracts* specifying requirements on the output data rather than purchasing a specific number of devices for a given duration. The reasoning behind device choice can also influence analytical decisions and thresholds set that can massively impact research findings (e.g., multiverse analysis, BOBA). Documentation on best practices for device choice and analysis would aid in designing research studies on mental health that leverage digital sensing.

In summary, consumer wearable devices offer a wealth of data for research, but their pre-processed data and changing algorithms pose challenges for reproducible findings. In contrast, research-grade devices provide more research-friendly formats and established protocols for measurement but are not what the public use. To compare results from wearable devices, a benchmark is needed, requiring transparency and access to minimally processed data. Device developers should provide research access to minimally processed data, clear versioning and documentation, and specification of sensor performance and reliability. Research teams should justify their choice of device, and transparency around devices and benchmark datasets would aid in this process.

Solution Space

To address the problem space of devices as black boxes, proposals include:

- Benchmarking data sets to test device algorithms
- Finding a path to industry's buy-in / create incentives for releasing raw data
- Develop clearer regulatory pathways for devices to become medical/research grade
- Government agencies to require transparency
- Standardize format and exchange formats for raw data
- Availability of more open source algorithms for data processing
- Define openness in terms of fit-for-purpose

(Additional proposals can be viewed on the spreadsheet.)

2.3 Funding and Infrastructure

To address the above challenges, substantial work is needed. This involves both data collection as well as methods development and infrastructure development. Unfortunately, the current funding landscape is not highly amenable to simply supporting these areas[2], but there have been recent advances that are moving in the right direction. For example, the NIH Common Fund's Bridge to Artificial Intelligence (Bridge2AI) has funded 4 sites that are solely centered on data collection and a separate site focused on sharing such data and making it publicly available. Unfortunately, there is no NIH institute or center currently in existence that can sustain such work — perhaps the NLM would be best suited but its extremely modest budget would not sustain such work for the longer term. It is possible that an institute like the National Human Genome Research Institute (NHGRI) should be stood up for leveraging new digital health technologies — may we suggest the National Digital Health Research Institute (NDHRI, pronounced N-dry)? An example of a key programmatic loss which was of great importance in this field was the Big Data to Knowledge (BD2K) initiative, which like Bridge2AI, was also funded by the NIH Common Fund between 2013 and 2020. According to the website, BD2K "has addressed some major big data and data science challenges including lack of appropriate tools, poor data accessibility, and insufficient training in biomedical data science. It facilitated broad use of biomedical big data, developed and disseminated analysis methods and software, enhanced training relevant for large-scale data analysis, and supported efforts toward making data sets "FAIR" (Findable, Accessible, Interoperable, and Reusable). At the time of writing, this work is not completed nor is it ever likely to be, given the continuously changing ecosystems from which data is generated.

Thus, we make the case that, like other ongoing programs supported by the NIH, efforts surrounding data and related tools and methods need a stable funding home. Lastly, one of the five overarching review criteria for NIH grants is innovation, which inherently limits research, development, and sustainability in this space. We are aware that this evaluation criterion is planned to be merged with impact such that it will no longer be necessary to reinvent the wheel and/or domain hop in order to keep critical infrastructure projects on digital health funded long-term. Ultimately, sustained funding mechanisms that can support continuous data collection that keeps up with new technologies and shifting demographics is critical and requires substantial and continued investment from government research agencies.

One key data type that should exist in this space is benchmark datasets. Like in genomics, benchmark datasets can serve to evaluate and compare existing and new technologies. To further support FAIR research, there needs to be more oversight of the code and data that gets published. "Living paper" models would support such a need, with crowd-sourced peer review to ensure that authors continue to be accountable for their past published work (at least the availability and function or acknowledged deprecation of the code and/or data). Another method that could facilitate FAIR practices is standards and automated assessment of protocols related to data and code, for example data management and sharing plans, as well as automated evaluation of published works. Funding that supports long-term maintenance of software and data would be critical.

Further, funding agencies can reward members who uphold the FAIR community standards (e.g., publishers of code that is used widely; contributors to communities like StackOverflow) and withhold funding from those who do not to incentivize FAIR principles in real practice (rather than just in theory). Funding to develop digital health communities for knowledge sharing, such as Biostars for the bioinformatics community[3] supported by an R25 grant from the NIH NHGRI, could be supported for example by NSF's grants for building stakeholder communities. Similar efforts have also seen success in the Open Source Software environment, where initiatives like NSF's POSE may support efforts. Ultimately, however, the short-term funding mechanisms provided by NSF and NIH are inconsistent with the need to sustain and update this key research infrastructure, which involves data storage, computation, and transfer, all of which come with financial costs, over the long term.

Solution Space

To address the problem space of funding and infrastructure, proposals include:

- Public or open-source library of shared methods and ML models and documentation
- Develop a funding model for international sharing of infrastructure
- Engage with companies to support advanced efforts & sustainable business models
- Identify appropriate institutions to provide infrastructure services
- Publication focused on the topic of why current practices are unsustainable
- Engaging the open source software community to support efforts
- Develop a centralized infrastructure for hosting data and methods

(Additional proposals can be viewed on the spreadsheet.)

2. Shandhi MMH, et al. Recent Academic Research on Clinically Relevant Digital Measures: Systematic Review. J Med Internet Res. 2021 Sep 15;23(9):e29875. doi: 10.2196/29875. PMID: 34524089; PMCID: PMC8482196.
3. Parnell LD, et al. 2011 BioStar: An Online Question & Answer Resource for the Bioinformatics Community. PLoS Comput Biol 7(10): e1002216. doi:10.1371/journal.pcbi.1002216

2.4 Standardization, Provenance, and Metadata

Digital sensing gives researchers access to large amounts of data from a host of sensors. However, to leverage these data, researchers must learn the nuances of how each data type is represented, and that representation can change dramatically across studies, devices, and even individual participants over time. To enable consistent and reproducible research there is a need for standards at various levels – including data representation (syntax and semantics), metadata, and data provenance.

The availability of large amount of data at different granularity (from high-frequency, base data, like accelerometry, to calculated measures, like total sleep time) and from different modalities brings the need for standards at various levels, from data representation (syntax and semantics) to metadata and provenance to document data collection and processing (including the study

for which data was collected, demographics represented in that data, and potentially other information critical to FAIR analysis of the data and use of model(s) derived from that).

Even for the same digital health solution, changes in operating system, software and algorithm may affect the reported feature (e.g., heart rate) and need to be considered when pooling data for analysis, comparisons, feature extraction.

Existing data and metadata standards[4] are not well known or not applied[1]. APIs have scant information on the digital health solution generating data, algorithm(s) applied, validation performed. When datasets from different digital health solutions are pooled, researchers need to spend time aligning and harmonizing the data before any analysis can be performed. Insufficient metadata and provenance details also make it difficult to analyze and reuse data. And data heterogeneity impacts the use of digital health data by healthcare entities and the ability of different digital health solutions to use others' data.

Besides syntactic and semantic interoperability (data format and meaning), there is the need to harmonize features extracted from data and to define an ontology to structure such features[5,] which will also clarify and standardize nomenclature.

Connected to the topic of privacy, the person's consent to which of their data to share with whom for which uses and for how long creates a complex and dynamic scenario. The consent needs to be represented in the data to allow users of the data access based on the consent.

There is the need to decouple properties of the digital sensing solution from the person being studied, to avoid proliferation of details that are not as important for data use. The challenge here is that by categorizing information as metadata instead of data, a value judgment is being made as to whether that information is of direct relevance to the analysis or just a factor to be accounted for. For example, the phone hardware being used may be considered metadata when looking at comparing step counts, but is primary data when investigating relationships of phone choice to socioeconomic status. Hence we propose an approach of treating everything as data. Convergence on this topic will lead to agreement on minimal information standards and representation (e.g., separate files per data type, graph models, hierarchical models, etc.)

Analysis of data from digital sensors has a particular need for time synchronization with data at different frequencies. EventDTW can be used for alignment (except for event based data). Sharing real-time is controversial, yet losing that leads to loss of important information, like weekend effects, seasonality, etc. Marking of 'events' as a form of metadata is important for context.

Finally, there is the concept of robustness of digital biomarkers, i.e., what they are sensitive to, and the need for categorizing & describing what is good for what purpose (which brings up the topic of metadata possibly including some value judgment).

Solution Space

There was broad agreement on the need to begin with a minimum agreed-upon metadata set based on the group's initial understanding of what is important.

Considerations informing the decision process may include extending existing standards, like IEEE 1752.1 (Standard for Open Mobile Health Data—Representation of Metadata, Sleep, and Physical Activity Measures) and Open mHealth before developing new ones.

To address the problem space for lack of adherence to standardized data formats, methods, protocols, proposals include:

- Get funders to encourage or mandate the use of existing standards
- Support existing standardization initiatives
- Expand data types considered to include novel types, like voice characteristics
- Benchmark data to use an agreed-upon data format
- Develop / publish National Data Archive (NDA) data constructs for all digital measures
- Form a standards working group/consortium and apply for meeting funds to bring participants together and chart a path forward
- Develop an ontology that captures relationships between digital sensors and features and data streams
- Create a biomarker taxonomy
- Enable linking of sensor data to other data types

To address the problem space for provenance and metadata, proposals include:

- Development of a minimum information standard
- Enforcing metadata collection
- Creation of a standard for informed consent (see also section 2.1 above)
- Hiring of data managers
- Creating templates that make compliance easy
- Metadata indexed to a type (time, speaker, context)
- Survey of common elements in existing datasets
- Creating a standard for what a study protocol is

The following items will help work in the solutions space:

1. Existing data and metadata representation standardization
2. Freely available software for data aggregation & multi-modal visualization
3. Tools that make use of standardized representation of data

(Additional proposals can be reviewed on the spreadsheet.)

4. Digital Medicine Society: Library of Standards, available at: https://dimesociety.org/access-resources/sensor-data-integrations/data-standards/
5. HumanFirst Measure Ontology, available at: https://www.gohumanfirst.com/atlas/measure-ontologies

3. Next Steps

Workgroups

The group identified seven main workstreams and champions for each:

1. Consortium building, including secure funding for additional meeting
2. Participant/patient involvement & concerns
3. Scientific tools & methods, including development of templates and resources (i.e., open source libraries, benchmark datasets, consent forms, privacy tools, algorithms, NDA data constructs for digital measures, digital sensor data ontology, biomarker taxonomy, study protocols)
4. Standards, including development of minimum information standard, CDEs
5. Funding, regulators, & policy, including requirements for transparency, use of standards & metadata collection
6. Outreach to other fields, including industry & open-source software
7. Best Practices, including publications showing why current practices are unsustainable & examining practices that work in other fields

Cross-cutting concerns

were reflected in the structure of this report and for them champions were also identified, except for the first topic, for which the group will reach out to known experts

- Privacy & re-identification
- Devices are black boxes
- Funding and infrastructure
- Standardization, Provenance & Metadata

To prepare for the planned publication detailing the results of the workshop, the group will meet as a team.