Understanding disease through remote monitoring technology
A mobile health perspective on disease and diagnosis in three conditions: stress, epilepsy, and COVID-19

Stewart, Callum

Awarding institution:
King's College London

Understanding disease through remote monitoring technology

A mobile health perspective on disease and diagnosis in three conditions: stress, epilepsy, and COVID-19

Callum Stewart

Supervisor: Prof. R.J. Dobson
Dr A. Folarin

Department of Biostatistics and Health Informatics
King's College London

This dissertation is submitted for the degree of
Doctor of Philosophy

February 2024

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. The contents of this dissertation are the result of my own work, and are not the result of collaborations with others, unless specified in the text and acknowledgements in accordance with King's College London regulations. This PhD Thesis contains fewer than 55000 words excluding appendices and bibliography and has 34 figures.

Callum Stewart
February 2024

Acknowledgements

I would like to thank my supervisors, Richard Dobson and Amos Folarin, for introducing me to the field of digital health and giving me their guidance and support. It has been a great pleasure to be part of the Precision Health Informatics Data Lab at the IoPPN through which I have met and worked with many wonderful people. The RADAR-CNS consortium provided an expansive introduction to mobile health studies and set of talented researchers. This work would not have been possible without the support of the NIHR Maudsley BRC studentship, and I am grateful for the opportunity I would also like to express my thanks to Nic Lane and Abhinav Mehrotra for examining this thesis and for their thoughtful suggestions. I'd also like to thank Björn Canbäck and Tobias Ambjörnsson, who ran the bioinformatics Master's programme at Lund and supervised my Master's project respectively, for providing my first taste of academia. Finally, I want to especially thank my family for their constant love and support.

Abstract

Mobile systems and wearable technology have developed substantially over the last decade and provide a unique long-term and continuous insight and monitoring into medical conditions in health research. The opportunities afforded by mobile health in access, scale, and round-the-clock recording are counterbalanced by pronounced issues in areas like participant engagement, labelling, and dataset size. Throughout this thesis the different aspects of an mHealth study are addressed, from software development and study design to data collection and analysis. Three medically relevant fields are investigated: detection of stress from physiological signals, seizure detection in epilepsy and the characterisation and monitoring of COVID-19 through mobile health techniques.

The first two analytical chapters of the thesis focus on models for acute stress and epileptic seizure detection, two conditions with autonomic and physiological manifestations. Firstly, a multi-modal machine learning pipeline is developed targetting focal and general motor seizures in patients with epilepsy. The heterogenity and inter-individual differences present in this study motivated the investigation of methods to personalise models with relatively little data. I subsequently consider meta-learning for few-shot model personalisation within acute stress classification, finding increased performance compared to standard methods.

As the COVID-19 pandemic gripped the world the work of this thesis reoriented around using mHealth to understand the disease. Firstly, the study design and software development of Covid Collab, a crowdsourced, remote-enrollment COVID-19 study, are examined. Within these chapters, the patterns of participant enrolment and adherence in Covid Collab are also considered. Adherence could impact scientific interpretations if not properly accounted for. While basic drop-out and percent completion are often considered, a more dynamic view of a participant's behaviour can also be important. A hidden Markov model approach is used to compare participant engagement over time.

Secondly, the long-term effects of COVID are investigated through data collected in the Covid Collab study, giving insight into prevalence, risk factors, and symptom manifestation with respect to wearable-recorded physiological signals. Long-term and historical data accessed retrospectively facilitated the findings of significant correlations between development of long-COVID and mHealth-derived fitness and behaviour.

Table of contents

List of figures

List of tables

Acronyms

ACC accelerometry

API application programming interface

BMI body mass index

CNN convolutional neural network

ECG electrocardiography

EDA electrodermal activity

EMG electromyography

FAR false alarm rate

HF high frequency

HMM hidden Markov model

KCH King's College Hospital

KM Kaplan-Meier

LF low frequency

mHealth mobile health

MLP multilayer perceptron

PPG photoplethysmography

PPV positive predict value

RADAR-CNS Remote Assessment of Disease and Relapse—Central Nervous System

SVM support vector machine

TPR true positive rate

UI user interface

UK United Kingdom

ULF ultra low frequency

URL Uniform Resource Locator

VLF very low frequency

WESAD Wearable Stress and Affection Detection

Chapter 1

Introduction

1.1 Context and motivation

The intersection of medicine and technology has revolutionised the way healthcare is delivered, disease is diagnosed, and how new biological insights are made. Broadly, mobile health (mHealth) refers to the application of mobile technology within the medical or healthcare domain, and includes sensing technology that can monitor data streams from our personal devices, such as smartphones and wearables.[1]

The proliferation of consumer-level wearable devices containing physiological sensors offers an unparalleled degree of continuous and pervasive health state monitoring. Additionally, these technologies are driving greater self-management and self-ownership of a person's own health and data. A recent YouGov poll shows 31% of the adult UK population currently uses a wearable device.[2] These vast repositories hold potential as a trove of unique health data, but have not yet been leveraged to a great extent by health researchers. Issues remain on how to effectively access and use shared personal health data. Privacy concerns, bias due to ownership patterns of wearable devices[3] and the quality of data from commercial devices are some of the largest barriers to clinical and research use.

New insights into disease, progression, diagnosis, and treatment can be formed through the unprecedented perspective that remote monitoring technologies (RMT) provide.[4] Biomedical research and health informatics has historically been bottlenecked by the expense of data collection. As next generation sequencing transformed genomics to a field awash with data,[5] mHealth is driving the development of massive biomedical datasets[6,7] and the development of software platforms to support them.[6,8] There have already been important contributions to medical understanding in several fields, including cardiology,[9–11] epilepsy,[12] psychology, [13] pain management,[14] and recently COVID-19.[7,15,16]

1.2 Evaluation and limitations

1.2.1 Opportunities of mobile health

Mobile and remote sensing technology has the potential to fill a role that traditional medical research lacks. Outcome measures are often taken at a single point, and even where a longitudinal study design is used, measures are repeated at a coarse temporal resolution.[17] Mobile health can provide pervasive, continuous, and objective monitoring of a person's health state. [18] There is therefore interest in using mHealth to measure long-term trajectories and to fill in the gap in the time between traditional biomarker measurements.[19] A move towards personalised or person-centric medicine and clinical trials[20] may also be well-supported by the deep level of data available through wearable devices.[18]

The real-time perspective of remote sensors also make them an attractive target for health monitoring.[21] Particularly with increased prevalence of chronic health conditions[22] and an older population, monitoring disease progression and responding to relapses or acute events are becoming more important.

Finally, as smartphones and remote sensors become ubiquitous, they are a vector through which under-served and hard-to-reach groups could be included in research[23] and to interface with people outside the typical medical institutions and pathways.[24]

1.2.2 Limitations and current challenges

Even as the field of mobile health grows, scepticism remains.[25] There are a number of well-founded concerns and limitations that have limited impact on medical delivery and healthcare.[26] Two broad problems that affect mobile health research are firstly, the quality of mobile health data,[27] and secondly, how to interpret or analyse the complex data collected.

Data quality and bias

The objectivity of passive remote sensing is often touted,[28,29] but in practice data is often plagued by data quality issues, such as missingness[30] and sensor accuracy.[31,32] Passive metrics may be objective in the sense that they are free from the beliefs and attitudes, but not in the sense that they are free from bias. Additionally, there is still often a requirement for comparison against subjective labels in the absence of gold-standard outcomes.[33]

Bias in healthcare datasets can exacerbate existing inequalities[34] and limit any inferences to the study population, rather than more broadly.[35] Remote sensing and mobile health studies are not uniquely effected, models and findings based on relatively homogeneous and biased datasets are a problem across healthcare,[36,37] but specific aspects of mHealth studies,

such as recruitment strategies, participant adherence, and use of secondary data sources can introduce bias. On the other hand, remote patient monitoring has unique reach and could be a tool for reducing inequalities in healthcare and increasing the representativeness of data.

Secondary data sources, such as those collected by commercial wearables,[38] are a rich and vast source of digital health data, but should be used carefully. Data is likely to be biased towards sociodemographic factors that influence wearable ownership.[39] Many of the people who could most benefit from remote monitoring are the least likely to already have wearables, such as older adults with cognitive difficulties.[40] Additionally, overreliance on single manufacturers or device may produce algorithms that do not perform well out-of-sample data.

Issues in remote sensing analysis

A majority of studies still follow expert feature engineering followed by traditional machine learning algorithms. Mohr et al. put forward a hierarchical model to making sense of remote sensing data, where input from raw sensors are gradually built up to low-level features (e.g. activity recognition, sleep times, semantic location), higher-level behavioural markers (e.g. fatigue, circadian rhythm, stress), and finally to a clinical state such as depression.[1] It is a conceptually satisfying framework because there are similarities to the way that psychological disorders are built on behavioural, psychological, or biological criteria,[41] and it lends itself to interpretable findings.

However, to compound the above data issues, medical conditions are often heterogeneous in aetiology, symptoms, and presentation. Moreover, human behaviour and response to disease or mental health states can be complex and context dependent, context that can often be hard to determine. A combination of sparse labels, missing data, variable modalities, biased datasets, and heterogeneous outcomes make standard machine learning approaches difficult to apply in a way that will generalise well. The complexity of remote sensing data arguably lends itself to deep learning approaches, which have been increasingly used in the last decade as a data-driven approach to learning representations of data,[42] as opposed to the classic feature engineering approach. More recent developments in few-shot learning[43] and self-supervised learning[44] seem like natural fits to the problems of inter-individual variability, creating cross-domain or generalisable models, and making use of sparsely labelled data.

Deep learning techniques are generally accepted as being data-hungry.[45] While there are potentially huge quantities of data, privacy concerns and commercial interests mean a lot of it is hard to access. In practice, many existing datasets are small. Additional, label sparsity is very common. Health outcomes are often time-consuming and expensive to record, and

are often rare events that are not guaranteed to occur during the course of a study. Even large unlabelled repositories that may be suitable for self-supervised learning, such as those collected by Fitbit[46] or Apple Health,[47] are typically tied to a specific device or wearables company and so may not generalise well to other and future devices.

The multimodal nature of remote sensing offers many opportunities in the breadth of psychological or physical states that are identifiable.[48] For example, in seizure detection movement sensors, such as an accelerometer or electromyogram, are vital for identifying motor symptoms, while autonomic symptoms are detected through pulse monitoring, temperature, or skin conductivity sensors.[49] It also brings additional complexity to analysis. Balancing between modality-specific and cross-modality learning,[50] the point at which modalities should be aggregated,[51,52] irregular sampling,[53] and how to deal with the variable inclusion of modalities[51,54] have all received attention. However, differences in included sensors and the future development of wearable devices could mean inflexible multimodal models act as a barrier to generalisability.

1.3 Scope and research questions

This thesis concerns mHealth as it refers to the use of mobile systems to understand health, rather than its use in the delivery of healthcare. I consider aspects from the full spectrum of a mobile health study: study protocol design, software development, enrolment, adherence, data processing, and two types of analysis, a machine learning approach and a statistical risk factor analysis.

Mobile health is a broad field, covering a wide variety of technologies and data streams. A limited set of modalities is considered here, each study uses typical worn devices with photoplethysmography and accelerometry. Research-level devices are used in first two studies, which contain additional modalities (electrodermal activity, electrocardiography). The latter chapters concern a bring-your-own-device study which asked for the donation of commercial device wearable data.

1.3.1 Research Questions

Considering the full study pipeline reduces the depth that focusing on one aspect delivers, but allows the illustration of the impact of study design and adherence on downstream analysis and how particular outlooks can be applicable across disparate aspects of mHealth. In a sentence, the overarching theme can be described as *How can mobile health studies be used to make generalisable medically-oriented inferences?*. Whether that is through reducing or

understanding bias caused by study design, or trying to model inter-individual variation, I attempt to address this through the following research questions:

Can a multi-modal remote sensing system be used to detect focal seizures?

Seizures are characterised by bursts of electrical activity in the brain, but the physical manifestations exhibit great variation, making a widely applicable seizure detection algorithm difficult to create. One approach is to use the multiple modalities available from wearable devices to target different potential seizure symptoms.

How can contextual baseline data be used to make more accurate predictions?

Time-consuming and costly data collection often rule out person-specific models. However, depending on the task, small amounts of baseline and even labelled data often exist. This research question asks how that data might be used efficiently to improve individual-level inferences and approaches it in two ways. Firstly in a few-shot machine learning stress classification task in Chapter 4, and secondly using long-term commercial device heart rate recordings to predict counterfactual heart rate estimates in COVID-19 positive participants in Chapter 7.

Can a citizen science study with 'opportunistic' historic wearable sensor data provide novel insights into COVID-19?

Apart from providing contextual data to condition models, historic wearable data can also provide an objective marker of fitness or health that would otherwise be vulnerable to recall bias and subjectivity. Within the Covid Collab study, I look at the potential utility of historic wearable fitness data.

What are the implications of participant engagement on analysis in citizen science and mHealth studies?

There are potentially huge quantities of available personal health data collected through commercial devices or that could be collected as part of routine healthcare. Citizen science initiatives and data donation drives have started to make some of this data available to researchers. However, care must be taken to ensure that data quality and bias do not undermine outcomes or generalisability.

1.4 Chapter outlines and contributions

This thesis covers the aspects from the full lifetime of a mobile health study and the research questions therefore cover particular aspects over different parts of a study with the overarching goal of effectively using mobile health data. That is, I attempt to opportunistically use data to reduce bias and increase generalisability of results. For example, use of historic personal health data in the long-term sequelae following COVID-19 infection, the formulation of stress detection models as a few-shot learning problem using small amounts of baseline data, or considering how study protocol can reduce bias. Several chapters are included as papers with a preamble to provide context within the thesis and a chapter summary section following the paper to set the contributions of the paper within the theme of the thesis.

Chapter 2: Background and Methodology

Chapter 2 aims to give an overview of mHealth data, the biological background of the conditions studied, and the datasets and methodology used throughout the rest of the thesis.

Chapter 3: Seizure classification

The first analysis chapter looks at the classification of epileptic seizures in the RADAR-CNS epilepsy study.[55] Many existing seizure detection algorithms and studies focus on generalised tonic-clonic seizures and use a single signal modality. Here I use multiple modalities from a wearable device to classify multiple types of motor seizure. The main contribution is in the multi-modal model performance across both general and focal motor seizures.

The work was important to me specifically because it demonstrated the heterogeneity of medical conditions and the problems of generalisability. Focal motor movements across seizure types can be very different, and therefore it can be challenging to fit a model that generalises well. However, the seizures of a single person are often similar. A small amount of wearable data could be easily collected as part of a routine clinical visit. While not enough to train a traditional machine learning classifier, efficient use of personal labelled data could lead to more accurate, personalised models. My ability to test personalisation in this dataset was limited due to a lack of repeat seizures, but a physiological few-shot learning approach is considered in the following chapter.

Chapter 4: Stress classification

The work in seizure detection made clear the potential for and importance of personalised models. Motivated by the desire to efficiently use small amounts of an individual's data, in

Chapter 4 I look at how a probabilistic meta-learning neural network approach to adapt a model to a particular participant's data in a stress classification task.

Essentially, the problem of personalisation is posed as a few-shot learning problem and approached using a model from the neural process family in participants from two public datasets, *WESAD* and *DriveDB*. Neural processes (NPs) are a family of latent variable neural networks models introduced in 2018[56] that are designed to be capable of rapid adaption to new data.

I compare NPs to machine learning models that are traditionally used in stress and biosignal classification problems. The NPs are adapted using either baseline negative-class data, the baseline data and the first instance of a positive stress label, or random points from the test participant. In each case the adapted model outperforms the traditional machine learning algorithms while a NP conditioned on another participant's data performs similarly to the traditional methods.

Chapter 5: Mass Science and Engagement

When the COVID-19 pandemic swept the globe a lot of my work was refocused on setting up, and later analysing the data from, a remote citizen science study aiming to learn about the disease through historic and prospective wearable data. This chapter details the development of the Mass Science application and backend infrastructure which were produced for the study. Subsequently, the app has also been used in a national core study on the long term impacts of COVID-19 and could be further used in RADAR-base studies.

I also look at the engagement and adherence of participants in the study in this chapter. I used hidden Markov models to cluster participant's based on their self-reported questionnaire completion rates over time. This dynamic view of adherence shows bias in the patterns of self-report response that may not be adequately captured by a typical dropout or missing rate.

Chapter 6: Covid Collab study protocol paper

A protocol paper for the Covid Collab study forms the content of Chapter 6. Largely this chapter sets the scene for the following analysis chapter, but there are a couple of contributions from this work. A distinction between Covid Collab and the similar COVID-19 citizen science wearable studies that were run around the world is the inclusion of long-term historic data, and the succeeding chapter will help illustrate the importance of this data. Secondly, the data from Covid Collab is in the process of being made publicly available and therefore may be a useful source for secondary research.

Chapter 7: Long COVID presentation and risk factors using long term wearable data

The last analysis chapter uses statistical methods to investigate long COVID in the Covid Collab study. The presence of long COVID itself is considered in two ways: through persistent self-reported symptoms and through persistent changes to wearable-measured resting heart rate following COVID-19 infection. Uniquely, we include historic wearable data taken from up to several years before enrolment. This long-term data was important not only as an objective measure of prior physical fitness, but also as a source to fit time-series models to better estimate changes to resting heart rate.

We found significant persistent changes to the passive wearable signals and self-rated mental health scales following COVID-19 infection in a case-control comparison. Regressions showed significant protective effects against long COVID from increased historic activity levels.

The final chapter reflects in greater detail on the contributions of the thesis, how they fit into the wider research landscape, the limitations, and future direction.

1.5 Output and Publications

Throughout my PhD I was fortunate to work on a range of projects and software, some of which don't form part of the thesis but are within the realm of mobile health. Outside the work included in the thesis, they particularly focus on mobile health analysis in major depressive disorder and the development of aspects of the RADAR-base software platform. A list of publications follows.

1. C. L. Stewart et al.: RADAR-base: Major Depressive Disorder and Epilepsy Case Studies. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM. 2018, 1735–1743.
2. Z. Rashid et al.: RADAR-base: Epilepsy Case Study. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM. 2018, 227–230.
3. Y. Ranjan et al.: RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. *JMIR mHealth and uHealth* **7**(8) (2019), e11734.

4. Y. Ranjan et al.: Challenges & solutions in a hybrid mHealth mobile app. *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 2019, 183–186.
5. C. L. Stewart, A. Folarin, and R. Dobson: Personalized acute stress classification from physiological signals with neural processes. *arXiv preprint arXiv:2002.04176* (2020).
6. S. Sun et al.: Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *Journal of medical Internet research* **22**(9) (2020), e19992.
7. P. Laiou et al.: Home stay reflects symptoms severity in major depressive disorder: A multicenter observational study using geolocation data from smartphones. *medRxiv* (2021).
8. Y. Zhang et al.: Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. *JMIR mHealth and uHealth* **9**(4) (2021), e24604.
9. S. Sun et al.: The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions. *arXiv preprint arXiv:2112.11903* (2021).
10. C. Stewart et al.: Investigating the use of digital health technology to monitor COVID-19 and its effects: Protocol for Covid Collab, an observational study. *JMIR Research Protocols* (2021).
11. S. Liu et al.: Fitbeat: COVID-19 estimation based on wristband heart rate using a contrastive convolutional auto-encoder. *Pattern recognition* **123** (2022), 108403.
12. F. Matcham et al.: Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC psychiatry* **22**(1) (2022), 1–19.
13. Y. Zhang et al.: Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. *JMIR mHealth and uHealth* **9**(7) (2021), e29840.
14. Y. Zhang et al.: Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR mental health* **9**(3) (2022), e34898.
15. P. Laiou et al.: The association between home stay and symptom severity in major depressive disorder: preliminary findings from a multicenter observational study using geolocation data from smartphones. *JMIR mHealth and uHealth* **10**(1) (2022), e28095.
16. R. Dobson et al.: Long-term Participant Retention and Engagement Patterns in an App and Wearable-based Multinational Remote Digital Depression Study (2022).

17. P. Laiou et al.: Temporal evolution of multiday, epileptic functional networks prior to seizure occurrence. *medRxiv* (2022).

Chapter 2

Background and Methodology

2.1 Introduction

An idea of the breadth of the mHealth field was given in the previous chapter. Necessarily, the methods used and analysis that can be undertaken within mHealth research are equally broad. This chapter aims to give an overview of the medical areas and corresponding datasets that are used in this thesis and the analytical techniques that are either used in the succeeding analysis chapters, or else commonly used within closely related research. More detailed explanations of specific algorithms or datasets are given in the appropriate chapters.

2.2 Medical and biological background

2.2.1 Autonomic Nervous System

Before giving an overview of the three medical conditions relevant to this thesis, it is worth briefly considering the autonomic nervous system (ANS). Several of the common physiological signals collected in mHealth studies are indirect measures of autonomic function. Skin conductivity, or EDA, is largely driven by one of the two autonomic pathways. The cardiac signals, ECG and PPG, measure heart rate, which is itself partially controlled by the ANS.[57] Many of the cardiac-specific features derived from those signals are directly motivated as a biomarker for autonomic responses.[58] Moreover, autonomic dysfunction or particular signatures in the ANS are extremely common in a wide variety of medical conditions, including the three here: COVID-19;[59,60] epilepsy[61,62] and stress.[63]

The ANS is one of the two major components of the peripheral nervous system and is involved in unconscious control of many of the body's systems.[64] It is one of the crucial systems responsible for maintaining homeostasis, including digestion, blood pressure, and

kidney function.[65] It has three major subdivisions, two of which are visualised in Figure 2.1. The enteric nervous system is one of the components of the ANS, it deals with gastrointestinal function, but it is not considered in any further detail here.

The sympathetic nervous system (SNS) governs what is called the *fight or flight* response and is highly activated in stressful or emergency situations. The parasympathetic nervous system (PNS), on the other hand, has the epithet *rest and digest* and is predominant in relaxed conditions. The effect that each component has on other organs naturally follow these two goals in an obvious way. Importantly, several of these functions can be directly observed externally by physiological sensors. The SNS increases heart rate, causes blood vessel contraction, pupil dilation, contraction of piloerection muscles, and causes secretion from sweat glands. The PNS reduces heart rate, constricts the pupils, and causes secretion from the lacrimal and parotid glands.

Although not included in Figure 2.1 for visual clarity, all of the thoracic and lumbar spine segments shown involved with organs through the sympathetic pathway are also involved in sympathetic control of the blood vessels, hair follicles, and sweat glands.[64]

The ANS is a key mediator in many normal and abnormal physiological functions. It would be a Herculean task to list all the ways the ANS is affected by or causes disease and dysfunction. Suffice to say, it is involved across a wide range of conditions, including direct disorders of the ANS;[66] cardiac diseases;[67] mental health conditions;[68–70] neurological conditions[71] including epilepsy;[72,73] and many more. Certain manifestations of the autonomic nervous system can be monitored through wearable sensors including electrocardiography (ECG), electrodermal activity (EDA), and photoplethysmography (PPG). Long-term pervasive monitoring through these sensors could help unlock new understanding of many of the diseases and conditions that affect or are affected by the ANS, especially over longer time periods than have previously been possible.

2.2.2 Stress

Stress is a person's reaction to pressure or threat, it is also generally defined as something that alters the homeostatic balance of the body. It is common, it can be motivating, but it is also linked with seven of the top ten leading causes of death in the developed world and it is a major burden to health and wellbeing.[74] Stress has physiological, cognitive, emotional, and behavioural manifestations.[75] Physiologically, it is intrinsically linked to the ANS and to the hypothalamic-pituitary-adrenal (HPA) axis. As detailed above, the ANS can be monitored through wearable sensor signals. The increased concentration of cortisol produced by the HPA axis is a common biomarker for stress. There are research-level devices that can monitor cortisol levels non-invasively;[76,77] but these are not widespread. Most real-time wearable

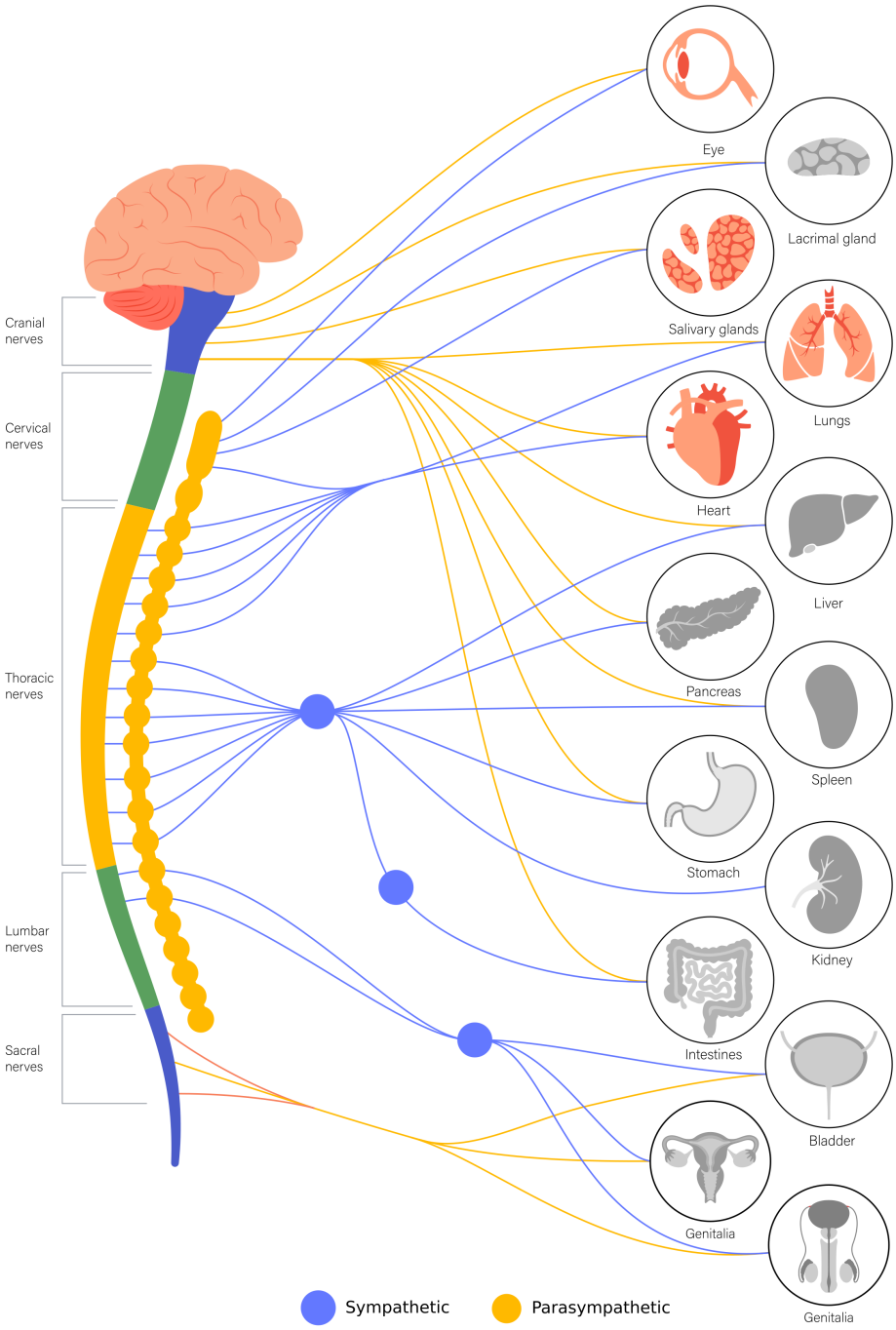


Fig. 2.1 Autonomic nervous system
The highlighted nodes represent the ANS-related physiological processes that are frequently or occasionally monitored in mHealth studies. The figure is an adaption of an image licensed from Adobe Stock.

stress detection research and datasets have, therefore, focused on electrodermal activity and heart rate variability,[78–81] as they are biomarkers of the ANS.

Stress is interesting as an intermediate measure because it is influential in and involved with many diseases. Accurate detection of stress could lead to better diagnostic or detection models for other diseases. For example, there is a strong correlation between stress and the triggering of epileptic seizures.[82] Automatic wearable-based recognition of when a person is stressed could therefore be an important part of understanding when somebody has a higher likelihood of having a seizure. A similar point could be made for other diseases, such as depression,[83] or risk of infection.[84]

2.2.3 Epilepsy

Epilepsy is a neurological condition which affects an estimated 65 million people around the world. It is characterised by the episodic appearance of seizures, and excessive or abnormally synchronised brain activity which can cause the presentation of a variety of psychological and physical symptoms. Despite advances in treatment, medication, and surgery, around one quarter to one third of patients contend with treatment-resistance epilepsy..[85]

Seizure classification

Seizures are perhaps most popularly imagined as the intense generalised tonic-clonic seizure (previously known as a grand mal); loss of consciousness followed by tonic contraction (stiffening of the muscles) and then clonic movement (rapid, repetitive jerking). However, many types of seizures exist. Advances in the conceptualisation of epilepsy has led to recent changes to the classification system of the aetiology epilepsy and the types of seizures presented.

The International League Against Epilepsy's (ILAE) 2017 classification system for seizures[86] splits seizures according to three main criteria: (1) the location of onset, (2) whether there is a motor component, and (3) level of awareness in focal onset seizures. Further levels of classification are possible through identifying specific motor or non-motor characteristics.

Location of onset is broadly split into generalised, the seizure originates in both sides of the brain and causes unconsciousness; focal, the seizure originates in a particular region of the brain; or unknown. More specific regions of onset in the brain can be determined. Awareness in a focal seizure is broadly split into *aware* and *impaired awareness*, but again more specific descriptions can be given.

Motor components include **automatisms**, somewhat coordinated motor movement which may resemble a normal activity; **atonic**, loss of muscle tone; **clonic**, a regular repetitive

jerking motion; **epileptic spasms**, a sudden extension or flexion of (typically truncal and proximal) muscles; **hyperkinetic**, intense, complex movement of the limbs and trunk; **myoclonic**, sudden short-duration muscle contraction; **tonic**, sustained muscle contraction; and **tonic-clonic**, a tonic contraction followed by clonic movement.

There is also an autonomic link, both during the seizure (ictal period) and at other times (post- or inter-ictal).[87] All aspects of the ANS can be involved. Ictal changes include changes to heart rate, blood pressure, pupil dilation, gastrointestinal, diaphoresis and flushing, and breathing rate, among others.[88] Other ictal symptoms can include cognitive and emotional manifestations.

Classifications are useful to understand shared behaviour and patterns between seizures, but seizures are unique. More specific descriptions of the presentation of a seizure are often given. An important caveat is that while a person with epilepsy can have multiple types of seizure, often seizure episodes in the same person are similar to each other.

Seizures and mHealth

Automated detection and classification of seizures through recorded signals is dealt with in more detail in Chapter 3. Hopefully it is clear that many of the physical manifestations of a seizure could theoretically be captured by wearable device sensors; inertial sensors could capture many of the motor components, while ECG, PPG, and EDA sensors may provide a way of monitoring autonomic changes.

Long-term monitoring may also illuminate relationships between stress, lifestyle, or particular activities and the likelihood of having a seizure. In an optimistic case, this may even extend to forecasting seizures or the likelihood of having a seizure ahead of time.[89]

2.2.4 COVID-19

Originating in December 2019 and quickly sweeping the globe, coronavirus disease 2019 (COVID-19) developed into a pandemic which defined healthcare, medical research, and people's lives for years following[90] and has gradually become endemic. There have been 6.54 million deaths and 616 million confirmed cases of COVID-19 as of September 2022. [91] The disease is caused by the Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus.[92] It is a respiratory infection with a wide range of symptoms and multi-organ effects. Common symptoms include fever, cough, breathing difficulties, fatigue, anosmia, sore throat, congestion, and nausea,[93] although different variants have different reported rates of symptoms.[94] The majority (80%) of cases with symptoms are mild, but there moderate (15%) and severe (5%) cases, which particularly affect older populations, and a mortality

rate of about 1%. In addition, around 25% of people with COVID-19 are estimated to be asymptomatic.[95]

Long-term sequelae following acute COVID-19 infection was brought to wider attention by patient advocacy groups in early 2020.[96] Particularly as the impact of the acute infection has faded and public health and safety measures have decreased, the longer term impact to health has been put in to focus. Popularly known as *long COVID* or *post-acute COVID-19 syndrome*, a wide range of symptoms and affected functions have been reported, perhaps most commonly fatigue,[97] but additionally chronic pain,[97] neurological conditions,[98] including to the ANS,[60,99] reduced mental wellbeing,[100,101] cognitive decline,[102] and long-term decreased respiratory[103,104] and cardiac[105,106] function. Many studies note the continued presence of symptoms and reduced function even at the end of the study,[107] suggesting an unknown or unbounded duration of long COVID symptoms for some people.

2.3 Mobile health data

This section will give an overview of mobile health datatypes, nomenclature, and the datasets used in this study.

Common modalities

Common types of data, which can be termed a modality, in remote monitoring mHealth studies include the following:

Physiological sensors

PPG An optical sensor which measures the blood volume pulse. Additionally, a low frequency baseline includes respiration and SNS activity components.[108] They are widely used as heart rate monitors on consumer wearable devices.

ECG A measurement of the electrical activity of the heart. Most wearable devices with ECG require electrodes placed on the skin, reducing their viability in consumer products. Recently ECG sensors have been included in some wearable wrist devices, requiring the person to form a circuit by touching the device with both hands.[109] Because it requires contact with both limbs, it can not continuously monitor heart rate.

EDA A sensor to measure the electrical activity of the skin. EDA is thought to be largely driven by SNS arousal.

Electromyography (EMG) A sensor to measure the electrical activity of a muscle. When implemented in a wearable, it is typically surface-EMG - an electrode in surface contact with the skin

Inertial sensors

Accelerometry A sensor which measures acceleration in 3 dimensions and often used in activity detection or classification. It is very common in both smartphones and wearable devices.

Gyroscope A gyroscope measures angular velocity, giving the orientation of a device. While common in smartphones, a gyroscope typically has much higher power usage than an accelerometer, so their use in wearables is restricted.

Magnetometer A compass, it measures the strength and direction of a magnetic field.

Location Precise geolocation can be determined by Global Navigation Satellite System (GNSS). The most well known satellite system is the Global Positioning System (GPS), but smartphones and some wearables often work with multiple systems.

Surveys and exercises Often mHealth studies use smartphones as a delivery mechanism for surveys, questionnaires, and exercises. Exercises can include cognitive tests[110] or physical tests, such as a six-minute walk test.[111]

Audio Audio can be recorded passively or as part of an active survey or exercise. Often audio recordings are of a person's voice, which has relationships to many diseases and mental states[112–114]

Images and video It has already found many uses in mHealth research, including melanoma risk in a mole mapping study,[115] burn severity diagnosis,[116] anaemia diagnosis.[117]

Smartphone usage How a person interacts and uses their phone can be measured in several ways, for example the types of apps they use, screen time, battery usage, nearby Bluetooth devices, phone or text communications, ambient light, and keyboard use.

Processed data

In addition to the above modalities, which can more-or-less be termed *'raw'* signals, combinations of signals, processed signals, and contextual information can be calculated from them. Especially when using commercial devices, it is more common to have access to these

derived or processed signals than to the underlying raw signal. For instance, a commercial wearable device may include a PPG sensor and an accelerometer. From these two signals, the device or backend infrastructure belonging to a company may classify sleep periods, level of activity, heart rate, and steps. It is often these processed signals that are available to a researcher.

Passive and active data

An important distinction between mHealth modalities is whether they require active attention from a participant, or whether they can be collected passively without significant burden. Passive RMT (pRMT) modalities include data such as the physiological sensor signals collected from wearable devices, background geolocation collection, smartphone use, and in some cases audio and video recordings. Active RMT (aRMT) include surveys, audio recording tasks, taking a photograph, or taking part in an exercise or cognitive test. Often, but not always, an active task forms part of an outcome measure in an mHealth study.

2.3.1 Datasets

Four datasets are used in this thesis. The epilepsy data comes from Remote Assessment of Disease and Relapse—Central Nervous System (RADAR-CNS), an international research project that had initiated data collection in 2017. There are two stress datasets, both are publicly available. The COVID-19 data comes from Covid Collab, an mHealth study we set up towards the beginning of the pandemic. A brief overview of each dataset follows. A fuller description will be given in the appropriate chapters, particularly for the Covid Collab dataset because the development of the study and data collection were direct parts of this thesis.

RADAR-CNS

The RADAR-CNS study was a large European-wide research project assessing the clinical use of smartphones and wearable devices in three disorders of the central nervous system: major depressive disorder, multiple sclerosis, and epilepsy. Data from the epilepsy study (RADAR-EPI) is used in this thesis. Participants were inpatients on the epilepsy monitoring units at two hospitals, King's College Hospital (KCH), London, and the Universitätsklinikum Freiburg. Continuous video-EEG monitoring with clinician-labelled seizure events was combined with the collection of wearable sensor data through the RADAR-base platform.

In total, 190 participants were enrolled at Freiburg and 72 were enrolled at KCH. One, or a combination, of three main wearables were used. The Empatica E4 wristband[118] and Biovotion vsm1 armband are wearable devices which measure PPG, EDA, acceleration, and temperature. They are used at both study sites. In addition, some participants at KCH used a prototype device developed by IMEC. Throughout the thesis, this device will often be referred to as the 'IMEC device'. Rather than measuring the heart through PPG, it uses ECG. In addition to EDA, it has an EMG sensor. The analysis in this thesis is carried out on the IMEC dataset.

WESAD

The Wearable Stress and Affection Detection (WESAD) dataset is, as the name suggests, a public dataset for stress and affect detection.[78] Physiological sensors monitor 15 participants under certain conditions designed to elicit an affective response or relaxation. Each recording is roughly two hours in length, and covers a baseline period, amusement, stress, meditation, and recovery conditions. Self-reported measures of affect are completed between tasks. Two devices were used, a RespiBAN with ECG, EDA, EMG, and temperature sensors worn on a chest strap, and an Empatica E4 with PPG, EDA, accelerometry, and temperature sensors worn on a wrist. The original paper considers classification with decision trees, random forest, AdaBoost, linear discriminant analysis, and k-Nearest Neighbours. The authors emphasised the future need for model personalisation because of inter-individual differences.

DriveDB

The Stress Recognition in Automobile Drivers dataset[79] (drivedb) is a public collection of physiological recordings of 17 participants driving a car along a route designed to elicit stress. Each participant is monitored through ECG (496Hz), EMG (15.5Hz), EDA (31Hz), and respiration(31Hz). The route protocol begins in a rest phase without driving and then alternate between city driving (3 instances) and highway (2 instances), and a final rest phase. The city driving segments are assumed to be more stressful. The original study included 24 participants, but only 17 are available in the public repository.[119]

Covid Collab

Covid Collab is a citizen science project. Participants could sign up, donate wearable data, and fill in regular COVID-19 and mental health related surveys, but there was little to no direct contact between researchers and participants. Since study set up as part of PhD, study

design and software development are elaborated on in future chapters before analysis. Over 17,750 participants had enrolled as of August 2022.

2.4 Analysis pipeline

The following two chapters in this thesis concern stress classification and seizure detection. Both rely heavily on electrical biosignals and inertial sensors. These are high frequency signals, between 32Hz and 2048Hz in the datasets used here, and require a typical digital signal processing pipeline to reduce noise and produce meaningful features from the raw signal. On the other hand, data collection is under fairly controlled conditions, with smaller cohorts, more participant oversight, and directly observed or induced outcome measures. The reasons for incomplete data are typically known.

The latter chapters are based around Covid Collab, a remote citizen science project. Data is collected opportunistically through mobile-based surveys and commercial wearable fitness devices and is largely captured in a processed or high-level form. Therefore, pre-processing and feature extraction is often less necessary or already done by third parties. There is a high degree of control over what is shared is given to participants, minimal oversight over individual participants, a large cohort size, and a longitudinal study design with high attrition rates, which all lead to a heterogeneous dataset with respect to data availability and completeness. The two types of dataset used here, therefore, lie far apart on the spectrum of mHealth data characteristics and the analytic techniques that will need to be used. However, there are commonalities and shared issues that should become apparent throughout the thesis.

2.4.1 Cleaning and pre-processing

Data collection

Data collection relies on a software platform to support it, particularly for large studies. The data collection apparatus can determine how the initial dataset is stored, formatted, and in the case of a live study, arrives. The storage and layout of a dataset is important from a FAIR perspective — how it is made findable, accessible, interoperable, and reusable — but also for performance. Mobile health datasets can be very large, long-term recordings of high frequency physiological signals for example, and so data format and locality can make order of magnitude differences in the time it takes to perform analyses. The data used in this thesis is initially stored in a variety of formats, but is typically converted into compressed N-dimensional arrays using the zarr library and format.[120]

Missing data

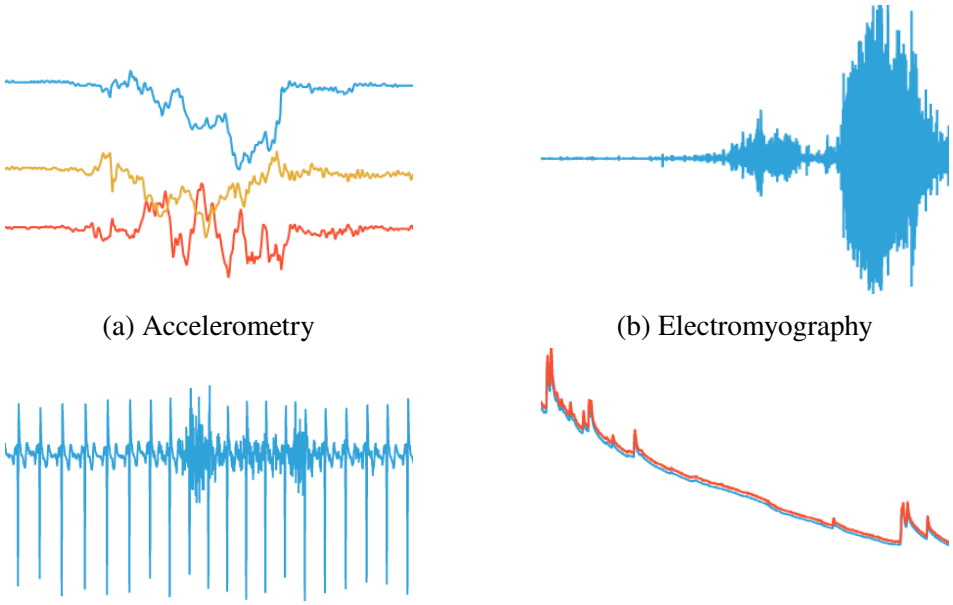
Data completeness and missingness is a problem in all studies, but exasperated in longitudinal mHealth studies and particularly citizen science projects. Missingness has implications on what analysis techniques can be applied, interpretation, and can cause bias.

It is common to describe missingness as either *Missing at random* (MAR), *Missing completely at random* (MCAR), or *Missing not at random* (MNAR). The differences between them can be subtle. MAR is not truly random, it refers to missingness which can be explained by other observed variables, for instance a younger person may be more likely to randomly miss a prompted-for survey. MCAR refers to missing data where the missingness does not have a relation to any of the observed data, for instance due to a random technical failure. MNAR refers to missing data with a systemic relationship to observed data even when taking into account other observed variables. For example, in a study of depression a person in a depressive episode may be less likely to complete a survey measure depressive symptoms.

If it is reasonable to assume that data is missing at random, certain techniques can be used to increase the power of the analysis compared to only using complete data. In certain circumstances it is possible to impute missing values. Multiple imputation[121] is a popular technique in epidemiological and clinical research.[122] Other techniques include coding category indicators for missingness in a regression, replacing missing values with mean or previous values, or by jointly modelling the reason for missingness and the outcome variable.[123]

Typically, these techniques envisage a matrix of covariates with individual missing points. Mobile health studies often record data persistently and so missing data can correspond to periods of time. Taking the case of a wearable that measures ECG at 256Hz, if a participant does not wear the device for several hours there will be millions of sequential unknown values. Imputation is clearly not possible on the actual signal, and whether it is possible on higher level features will depend on factors like the window length used to generate the feature of interest.

Missingness may also be informative. Above an example of a depressed participant not completing an outcome measure on depression was given, but it has also been suggested that that missingness could also be used as part of a predictive model, if active tasks were being collected.



(a) Accelerometry (b) Electromyography

(c) Electrocardiography (d) Electrodermal activity

Fig. 2.2 Example recordings of physiological signals

Implementation

To promote re-usability many of the features and preprocessing steps above were implemented in a Python library named pymhealth.[124] It uses Numba,[125] an LLVM compiler for Python, to produce much quicker feature extraction pipelines than were otherwise available in Python.

Signal processing

Digital signal processing is important in studies where raw high frequency signals are collected, such as ECG or accelerometry. Many of these physiological signals are well studied and with a large body of work on their processing steps and analysis. However, typically they were developed under very controlled circumstances with high quality sensors. Mobile health adds a dimension of complexity because the sensors are typically constrained in size and therefore quality, participants are often in free living or ambulatory conditions rather than confined to a ward, and data must be collected remotely rather than offline. These difference often lead to signals with more noise, increased motion artefacts, and potential missingness due to technical errors, all of which can require modification to the typical processing steps.

An important step in signal processing is often the removal of certain frequency bands outside the informative frequencies, or in which noise is expected. Different signals will have different frequencies with useful information. Unless otherwise stated, all filters used in this thesis are 5th order Butterworth filters implemented in the scipy package.[126]

The initial processing of ECG signals in the analyses of this thesis follows a standard pipeline. Baseline drift and interference from electromyography are removed with a 0.5Hz high-pass and 100Hz low-pass filter respectively. Power-line interference is removed with a 50Hz band-stop filter. The sinus rhythm produces a characteristic QRS-complex on an ECG, where the R-peak is detected to calculate the inter-beat-interval between heartbeats. The Hamilton-Tompkins algorithm[127] is implemented in the previously mentioned pymhealth analysis library for R-peak detection.

EDA is measured through applying two small electrodes to the skin, applying a very low voltage, and measuring the current between them. The conductance of the skin (the reciprocal of the resistance) is linked to the sympathetic nervous system and EDA is often used as a marker for changes in sympathetic arousal.[128] It is formed from two components, a slowly varying baseline *tonic* component and transient peaks which last several seconds in the *phasic* component. The peaks in the phasic component are often called the skin conductance response (SCR), and are in reaction to sporadic or event-driven sympathetic stimuli. Although sophisticated optimisation algorithms exist to separate the tonic and phasic components in EDA,[129] specific and irreducible types of noise in some of the data used ruled out their use. The tonic component is estimated using a 0.2Hz low-pass filter and the phasic component by a 0.5Hz-2Hz bandpass filter.

Accelerometry is also formed of two components. The *linear* acceleration caused by movement of the sensor, corresponding to whatever part of the body it is worn on, and the constant *gravitational* acceleration pulling towards the ground. The gravitational component can be roughly estimated by applying a 0.5Hz low-pass filter, while the linear component is a high-pass filter at the same value. It is possible to roughly calculate roll and pitch, but now yaw, from the acceleration in the x, y, and z plane of the gravitational component, the equations are given below.[130] A simple approach is taken with surface EMG signals. Only power-line interference is removed using a bandpass filter at 50Hz.[131]

$$roll = arctan2(y, z) \tag{2.1}$$

$$pitch = arctan2(-x, \sqrt{y^2 + z^2}) \tag{2.2}$$

2.4.2 Feature extraction

A feature is a higher level representation or characteristic of an underlying signal. Sometimes they are basic summary statistics, creating single values from a longer sequence: minimum, maximum, mean, and so on. Often they are biologically motivated or designed to reflect or capture certain physiological behaviour. Features are often task specific, even if the underlying signal is the same. For example, low frequency heart rate variability might be important if you were looking at depression,[132] while short-term heart rate changes may be important in a seizure with a tachycardia.

Typically a feature is calculated over a window, not the entire recording. Because we are usually interested in detecting an event without prior knowledge of the time or duration, features are extracted in a sliding window. Features are calculated i a window of length t_w and shifted along in steps of length t_s. Often the windows are overlapping. Calculating features in a window repeatedly over the length of a signal is a problem that is easily solved by looping over the indices of the windows, but it is a problem for which Python is notoriously slow. The windowing loop and below features are therefore implemented in pymhealth and just-in-time compiled to a LLVM representation using Numba, which greatly speeds up processing.

Common features that are applied to a variety of signals include summary statistics (mean, standard deviation, minimum, maximum, range, median, skew, kurtosis), zero-crossing rate, statistics on the derivatives of a signal, Hjorth parameters, and entropy based features. Hjorth parameters are a set of metrics developed to describe the characteristics of an EEG trace,[133] but have since found wider application. Entropy is a measure of the complexity or information content of a signal. Approximate entropy and sample entropy are often used as features in physiological signal analysis.[134]

Cardiac

Heart rate can be estimated as the reciprocal of the inter-beat-intervals estimated as the distance R-peaks. Often basic summary statistics are calculated directly on heart rate. Heart rate variability refers to the variation in inter-beat-intervals, the duration between heart beats. Many features are derived from the normal R-peal intervals (NNI). A list of the time domain and some non-linear recurrence-based features is given below. Recurrence refers to features based on the Poincaré plot — a plot of NNI_n against NNI_{n+1}. Other non-specific features are also often calculated for heart rate, including detrended fluctuation analysis and sample entropy. A more detailed overview is given in a paper by Shaffer & Ginsberg.[135] The window size used for cardiac features vary. For time-domain features and high-frequency frequency

features, a window length of 60s can be adequate. Lower frequency features require a longer time window, from 5 minutes to 24 hours.

SDNN Standard deviation of normalised R-peak intervals. Typically taken on a period of 5 minutes or 24 hours.

SDANN The standard deviation of the mean R-peak interval in each window (typically 5 minutes) over a longer period (typically 24 hours).

SDNNI The mean of the standard deviation of the R-peak intervals in each window (typically 5 minutes) over a longer period (typically 24 hours).

pNN50 The proportion of R-peak intervals that differ from the previous value by more than 50ms.

RMSSD The root-mean-square of successive differences.

SSD Sum of successive differences

SD1 The width of the ellipsis on a Poincaré plot. Equivalent to the standard deviation of successive differences times some factor.

SD2 The standard deviation of the longitudinal length of the Poincaré plot.

Frequency domain features are based around the power, peak, and relative power of different frequency bands in the HRV sequence. The four bands are the ultra low frequency (ULF) band (≤ 0.003 Hz), very low frequency (VLF) band (0.003 - 0.04Hz), low frequency (LF) band (0.04 - 0.15Hz), and high frequency (HF) (0.15-0.4 Hz) band.

Electrodermal Activity

Along with standard statistical features on the phasic SCR and tonic level, some specific features are generated, typically in the time domain.[136] Features associated with the SCR peaks include rise time, peak amplitude, recovery time, area under the SCRs, number of SCRs in a window, and the mean magnitude of the SCRs.[137] Window length can vary, from as little as around 10s in the literature if only single SCR are of interest. A window length of 40s is used in the stress detection classification task in this thesis, while in seizure detection the long-term changes to tonic EDA following a seizure take several minutes to appear, and so a window of 5 minutes is used.

Electromyography and accelerometry

The electromyography and accelerometry features used here are fairly straightforward and not specific to the signal. They include measures like the zero crossing count, line length, and Hjorth parameters.[133] Window lengths as low as 2s are common for electromyography, and around 10s for accelerometry.

2.4.3 Visualisation

Visualisation is an important step in an analytic process. Data exploration can help form impressions of new data, give an idea of the patterns present, and also help communicate results. It is also useful as an overview of incoming data while study is running, giving a better chance to react to missing data or technical errors.

2.4.4 Modelling

Machine learning modelling is the process of generating an algorithm or program that can perform a task from a set of data. There are various ways to categorise types of machine learning models and the tasks that they perform. The choice of a machine learning model is on the basis of the type of task, the estimated or empirical performance, and the assumptions and requirements the model has. Common tasks include:

Classification Predicting a categorical label.

Regression A regression is a prediction of a continuous outcome variable through explanatory variables. Forecasting is a subset of regression, in which the future of a continuous time-series is predicted.

Clustering Group data without explicitly given labels.

Supervised and unsupervised learning

Supervised learning refers to training a model where the training instances have a known outcome, for instance class labels. Unsupervised learning refers to training a model that does not have a known output. For example, clustering into groups on the basis of similarities in the input data.

Interpretability and feature importance

There is often a trade-off between the complexity of a model and how easy it is to interpret the results. In an ordinary least squares regression, the relationship between the explanatory variables and the outcome is very clear - each variable has an associated coefficient, and they are linearly related to the outcome. On the other hand, a deep neural network can model complex non-linear behaviour, but how the inputs relate to the output in the model is hard to determine. While there is work done to interpret neural networks,[138] it is intrinsically harder than many other models.

Evaluation

There are several metrics that can help evaluate the performance of a model. Below the performance of a binary classification task is considered, which is the most commonly evaluated task type in this thesis.

		True condition		
		Positive	**Negative**	**Total**
Predicted condition	**Positive**	TP	FP	TP + FP
	Negative	FN	TN	FN + TN
	Total	TP + FN	FP + TN	N

TP: True positive FP: False positive FN: False negative TN: True negative N: Total number of classifications

Table 2.1 Contingency Table

Sensitivity, specificity, and precision

Sensitivity, also known as the true positive rate (TPR) or recall, is the number of correct positive predictions divided by the total number of all positive predictions.

$$Sensitivity = TPR = Recall = \frac{TP}{TP + FN}$$

Specificity is the number of correct negative predictions divided by the number of all negative predictions.

$$Specificity = \frac{TN}{TN + FP}$$

Precision, also known as the positive predict value (PPV), is the proportion of positively classified instances that were correct.

$$Precision = PPV = \frac{TP}{TP + FP}$$

Accuracy is the proportion of correctly predicted class labels

$$Accuracy = \frac{TP + TN}{N}$$

The harmonic mean of the sensitivity and specificity, which is called the F1-score, is often given as a performance metric.

Receiver Operator Curve and Precision-Recall

In addition to point estimates, it is common to visualise how a model's performance varies as its discrimination threshold is changed. A Receiver operator curve (ROC) is a plot of the senstivity against the false positive rate (1 - specificity). It visualises the model performance and can be used as a tool to compare models without making a prior assumption about the relative importance of the classes (A TP may be more or less important than a FP). The area under the curve (AUC) of a ROC plot is often used as a performance metric in its own right. A precision-recall plot is similar, but plots precision against sensitivity (recall). It can be a more useful visualisation of model performance where there are large class imbalances.

How a model should be evaluated depends on the goal or task. For example, a seizure is a rare event and a detection model with clinical or real life use would require a very high sensitivity. The number of false alarms should be reduced, but a lower specificity could be accepted up to a point.

Training and cross-validation

While there are methods to reduce overfitting and increase generalisability, a model is likely to perform better on the training dataset than unseen samples from the same task. Certain train/test methodologies can be used to estimate generalised performance. The most basic is to hold back a proportion of 'test' data instances which are not used during training, and to then perform model evaluation on them. Additionally, a test dataset formed from a different cohort or sample to the training dataset can provide a more accurate view of how the model

will perform on new data. Cross validation is a resampling method which randomly splits the dataset into train/test set at each iteration. It is often used either in conjunction with the standard train/test set to fit hyperparameters or select a model before final testing, or instead of the standard train/test if a dataset is too small to set aside a portion for testing. Leave-one-participant-out (LOPO) or leave-n-participants-out is a modification to cross validation in longitudinal or time-series based studies. Because data points belonging to a single participant are not independent, the dataset can not be split randomly. Instead, all data instances belonging to a single participant are included or not included in the training set.

2.5 Models

2.5.1 Machine and Statistical learning

Linear models

There is an expansive field of generalised linear models. Only multiple linear regression and logistic regression are used in this thesis. The statsmodels library implementation of both linear models is used in this thesis.[139]

A GLM consist of random component (distribution of outcome variable), systematic component (explanatory variables), and a link function (a function that relates the outcome variable to the explanatory variables).

A linear regression is a regression model which takes a linear combination of weighted explanatory variables to predict a continuous outcome variable.

A logistic regression is a 2-class classification model (which can be extended to multi-class). The outcome variable (random component) is assumed binomial, the link function is the logit.[140]

$$logit(\theta_i) = log\left(\frac{\theta_i}{1-\theta_i}\right)$$

A related technique is the proportional hazards regression, used for survival analysis with right-censored data. A more thorough explanation is included in the appropriate chapter.

Nearest neighbours

A simple supervised algorithm that can either be used in either regression or classification by comparing a point to the nearest labelled points according to some distance measure. For an unlabelled point, the closed K points are taken. In a regression the K points are averaged. In a classification the most common label is taken.

Support vector machine

The support vector machine (SVM) is a common clustering and regression machine learning algorithm.[141] For classification, the aim is to fit a decision boundary (or hyperplane) such that the margin (distance between the separating hyperplane and the closest points) is maximised. For a one-dimensional variable, the decision boundary would be a point, for two dimensions a line, for three a plane, and so on. The boundary fit is linear, but often a boundary between classes will not be. To account for that fact, an SVM can apply a kernel function to the data. The kernel function transforms, or projects, the data into a higher dimensional space in which the SVM can fit a linear boundary that can appear non-linear in the original input dimension. The SVM can therefore be used in classification problems with a non-linear separation in the explanatory variable space with clever choice of a kernel function. When used in this thesis, SVMs are fit with a radial basis function kernel using the scikit-learn python package, [142] which wraps the libsvm library.[143]

Decision tree

A decision tree is essentially a flow chart for classification or regression. It successively splits data according to certain rules until the data at a split is all of one type. There are several types of decision tree and training algorithm.[144]

Ensembles

An ensemble is a collection of weaker estimators, or models. The predictions from each model are combined to produce a better estimate. One of the most popular ensembles is the Random Forest (RF),[145] a collection of decision tree predictors.

Hidden Markov models

The hidden Markov model (HMM) is a model in which 'hidden' states are responsible for generating the observed sequence of variables. As the name implies, the model assumes the system is a Markov process and so the probability of being in a state is only dependent on the previous state and the observed outcome at that time. Consider a model with N_S hidden states and a categorical outcome variable with an alphabet of length N_O. At each time point t_i, the probability of moving to a state is defined by an $N_S \times N_S$ transition matrix. Traditionally the outcome sequence is a categorical variable and so the emission probability is defined by an $N_S \times N_O$ matrix, giving the probability of emitting each category for each state. A vector gives the probability of starting the model at t_0 for each state. The typical problems solved are:

1. Estimating the optimal sequence of hidden states given an observed sequence and certain model parameters. Solved by the Viterbi algorithm.

2. Estimating the likelihood of a model given certain parameters and an observed sequence. Solved by the Forward-Backward algorithm.

3. Estimating the model parameters. Estimated with the Baum-Welch algorithm, a particular implementation of the expectation-maximisation algorithm.

2.5.2 Neural networks and Deep learning

Introduction and MLP

A neural network is a model inspired by biological neural networks. They are non-linear models which learn to approximate a function by building together layers of connected nodes. Each node has a set of weights w_i, a bias b, and activation function. It transforms an input vector x_i into a scalar output y. The output is the result of applying the activation function to the sum of the weighted input plus the bias. An activation function is typically non-linear. The three most common are the sigmoid, TanH, and RelU functions.

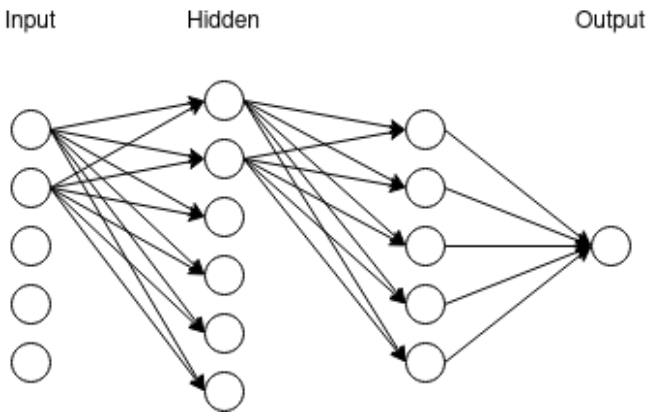


Fig. 2.3 An example neural network
A neural network with two hidden layers. The connections between each layer are only visualised for at most 2 nodes. Typically, a node in one layer connects to every node in the previous and subsequent layer.

A loss function is a measure of how well a model predicts the data and is minimised when fitting the model. The loss function used is task dependent. In a regression it is common to use the mean squared error (MSE), and in a binary classification problem the log loss is used (as in a logistic regression).

Common structures

In the introduction to neural networks the fully connected layer and multilayer perceptron (MLP) were briefly explained. The parameter space of a large or deep fully connected neural network would become very big. There are several over structures that are commonly used in neural networks and are relied upon in most successful modern models.

Fully connected A fully connected or linear layer was described above. It is a layer in which all nodes are connected to all the nodes in the previous layer. A fully connected layer is commonly used as the last layers in a model. For example, a final fully connected layer will be used in a classification task, where each node corresponds to a class label.

Convolutional A convolutional layer is a main component of a convolutional neural network (CNN). It consists of a learnable filter (or kernel) which is convolved with the input.[146]

Pooling A convolutional layer is often combined, or interspersed, with a pooling layer. The pooling layer is a non-trained layer which downsamples the feature maps output by a convolutional layer. A region of a certain size and sliding across the input with a certain stride takes an aggregation at each point. The most common aggregation is to take the maximum value (Max pooling).

RNN Recurrent neural networks were developed to address arbitrarily length sequences that contain a hidden unit which allows the output at a current time point to affect the input of the same node at a future time point.

Autoencoder An autoencoder is a particular structure that learns a reduced representation of an input. It is composed of an input layer, a hidden layer, and an output layer. Typically, the hidden layer is composed of fewer nodes than the input and output layers. The input layer(s) take an input sequence, and gradually reduce the number of nodes at each step until the hidden layer. The output layer reverses the process and is trained to reconstruct the original sequence. A schema for an autoencoder is given in Figure 2.4.

Meta-learning

Meta-learning is a term which is conceptually concerned with *learning to learn*, or how the process of learning a model can be improved. It refers to multiple different specific areas or implementations within machine learning. Here we focus on the use of meta-learning within neural networks. There are several ways in which the learning process can be improved, but

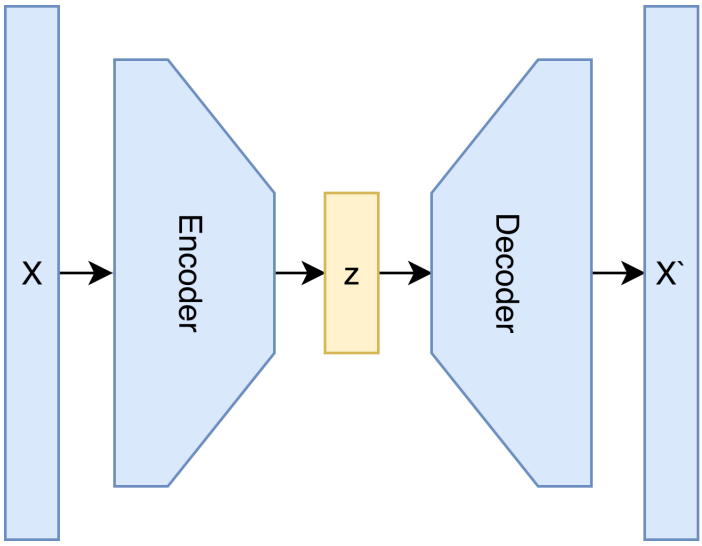


Fig. 2.4 Autoencoder
An autoencoder learns a reduced representation z of the input X.

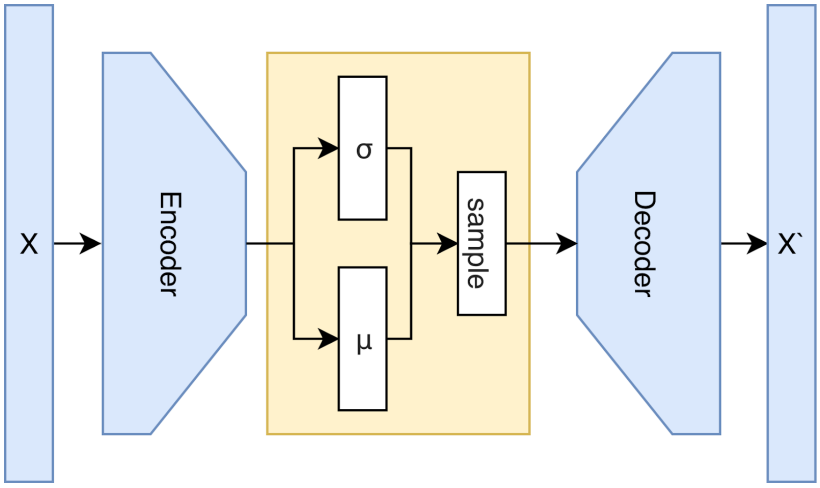


Fig. 2.5 Variational autoencoder
A variational autoencoder is a generative model. Rather than a reduced representation of fixed values, a latent probability distribution that can be sampled from is learnt.[147]

in this thesis we are primarily concerned with reducing the amount of training data required to produce a model, or in other words few-shot learning. Conceptually, few-shot learning is attractive because many physiological processes, personal behaviours, or manifestations of a disease can be very different depending on the individual, but can be similar over time when considering only a single person. Therefore, a general model trained on a particular task may not be accurate enough for certain (or any) individuals. For a medical application it is not viable to collect enough data from every person to provide individual models. Therefore it would be attractive if a model that could be quickly trained on only a few training instances because a meta-learner had learnt how to quickly adapt on the basis of many similar tasks. There are several meta-learning methods. Only the neural process,[56] the implementation of which is explained in Chapter 4, is used.

Chapter 3

Multi-modal motor seizure detection

3.1 Introduction

Epilepsy is a common neurological condition, affecting 65 million people worldwide.[85] Epilepsy is characterised seizures cause by transient episodes of excessive activity in the brain. Around two-thirds of people with epilepsy respond well to treatment with drugs. A small proportion of those remaining can benefit from surgery, leaving around 25-30% with treatment-resistant epilepsy and at the risk of seizures. In this context, there are a few ways in which automatic detection of seizures through physiological signal monitoring can be useful. Firstly, in those people with drug-resistant epilepsy, seizure detection could alert a caregiver that a seizure is happening to provide timely intervention and a reduction in the risk of Sudden Unexpected Death in Epilepsy (SUDEP). Secondly, video-EEG systems demonstrate that patients under-report seizures[148] and can be themselves unaware of a seizure.[149] A detection device could provide an objective measure of seizure frequency, which can be important in the delivery and evaluation of treatment.

The most common and accurate seizure detection methods rely on direct recording of brain activity by an electroencephalogram (EEG) device. However, EEG relies on either the surgical implanting of electrodes at the surface or within the brain,[150] or else attaching multiple electrodes in proximity to the scalp.[151] However, scalp EEG is not suitable for long-term every day use because patients find them stigmatising, uncomfortable, and awkward. [151,152] Implanted EEG requires surgery, which carries risks and may not be tolerated by patients.[153] Despite the block in long-term monitoring, video-EEG monitored by a clinician, typically in a clinical setting, still provides the gold standard in seizure detection against which other methods are compared.

Non-EEG monitoring covers anything from radar to sensor-equipped mattresses or seizure alert dogs.[154] Since the early 2010s there has been an increasing amount of seizure detection

work using wearable physiological sensors; typically one or a combination of accelerometry (ACC), electromyography (EMG), photoplethysmography (PPG), electrocardiography (ECG), and electrodermal activity (EDA). These sensors are widely available, often even on consumer wearable fitness trackers, and can be integrated into discrete or every-day devices.

Accelerometry and EMG are both sensors that could identify motor components of a seizure. An accelerometer could theoretically recognise movements that are characteristic of a seizure so long as the accelerometer is attached to a part of the body that is affected by the particular seizure. Certain common types of motor patterns, e.g. tonic, clonic, and myoclonic, leave characteristic accelerometer traces. Other types could also be recognised, but the particular pattern may differ. An EMG is a measure of electrical activity in a muscle. So long as the EMG sensor is in proximity of a muscle that is noticeably activated during a seizure, it could be useful in detection.

The PPG and ECG sensors are two methods of measuring heart beats. Autonomic symptoms are common in seizures and those affecting cardiac function could be picked up.[87] EDA is also linked to autonomic symptoms through the sympathetic nervous system, and could therefore provide another view of autonomic changes. Seizures are made up of various symptoms or components that can occur at different points in the seizure.

Certain sensor modalities are likely to pick up different symptoms. Figure 3.1 shows the ACC, EMG, ECG, and EDA recordings of a seizure in a participant in this study with the components in blocks above. It is clear that different signals become recognisably different from normal behaviour at different points in the seizure. Additionally, signal quality, outside of ACC, can be adversely affected by motion during the seizure.

Other studies have demonstrated good results for generalised tonic-clonic seizure (GTCS) detection and variable results for focal motor seizures. Detection algorithms are often developed on a single sensor modality, but there is also a drive for multimodal seizure detection with the expectation that it should be more accurate and potentially cover a wider range of seizures.

In this study the major aims were to build a multimodal seizure classification pipeline for motor seizures, both generalised tonic-clonic and focal, and to determine the modality and feature importance. The feature set used in this study is quite typical and based on features described in the non-EEG seizure detection literature. The only major divergence is the use of an approximation of the Euler angles (pitch and roll) rather than using low-pass filtered acceleration.

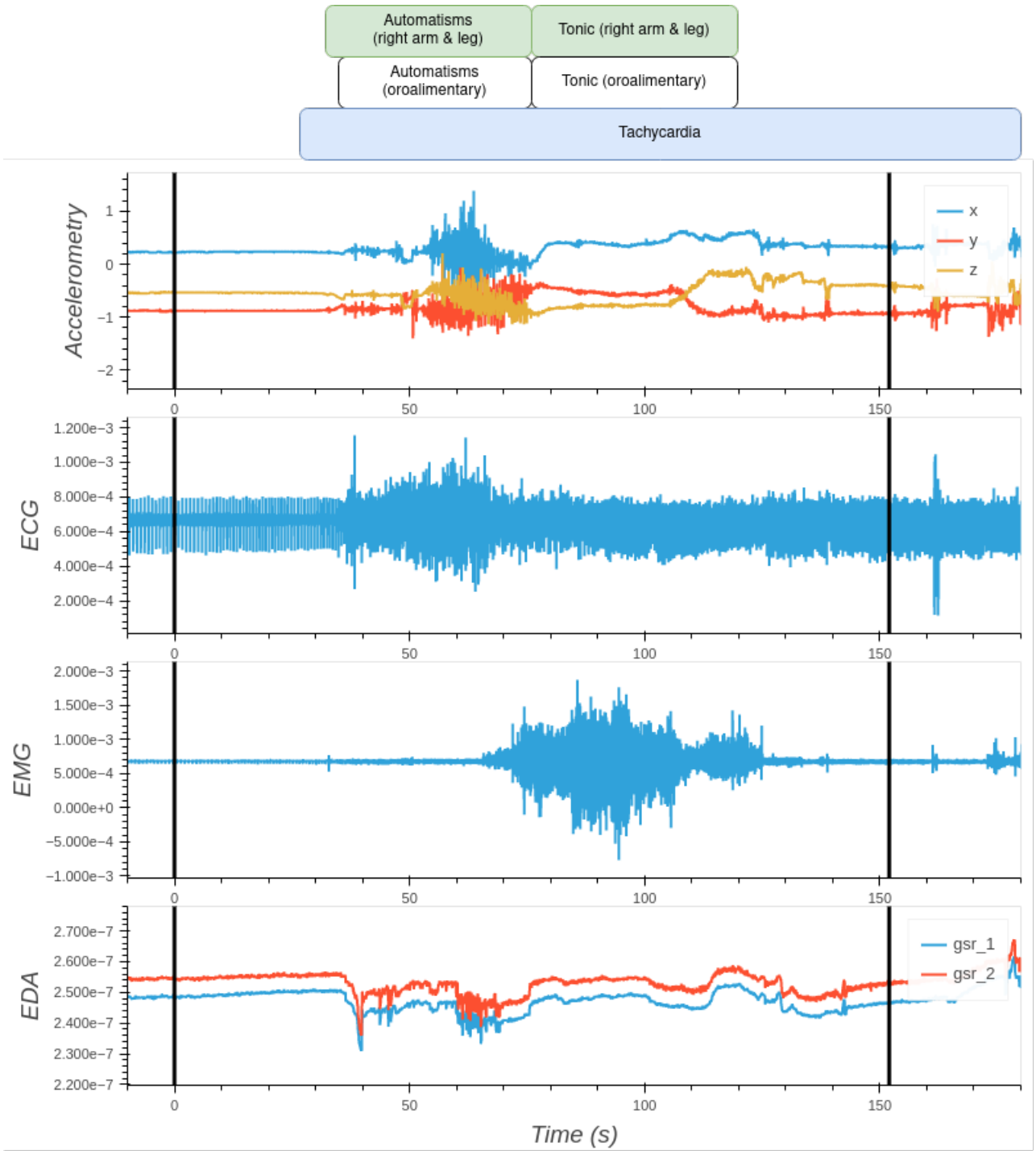


Fig. 3.1 Raw physiological signals during an example seizure
The boxes at the top show the duration of components of the seizure. Green boxes are motoric components that could conceivably be captured by a wearable device. White boxes are motoric components that are unlikely to be captured. The blue coloured box represents an autonomic change. Below are four raw signals from the IMEC device: accelerometry, ECG, EMG, and EDA. The black vertical bars represent the start and end of the seizure.

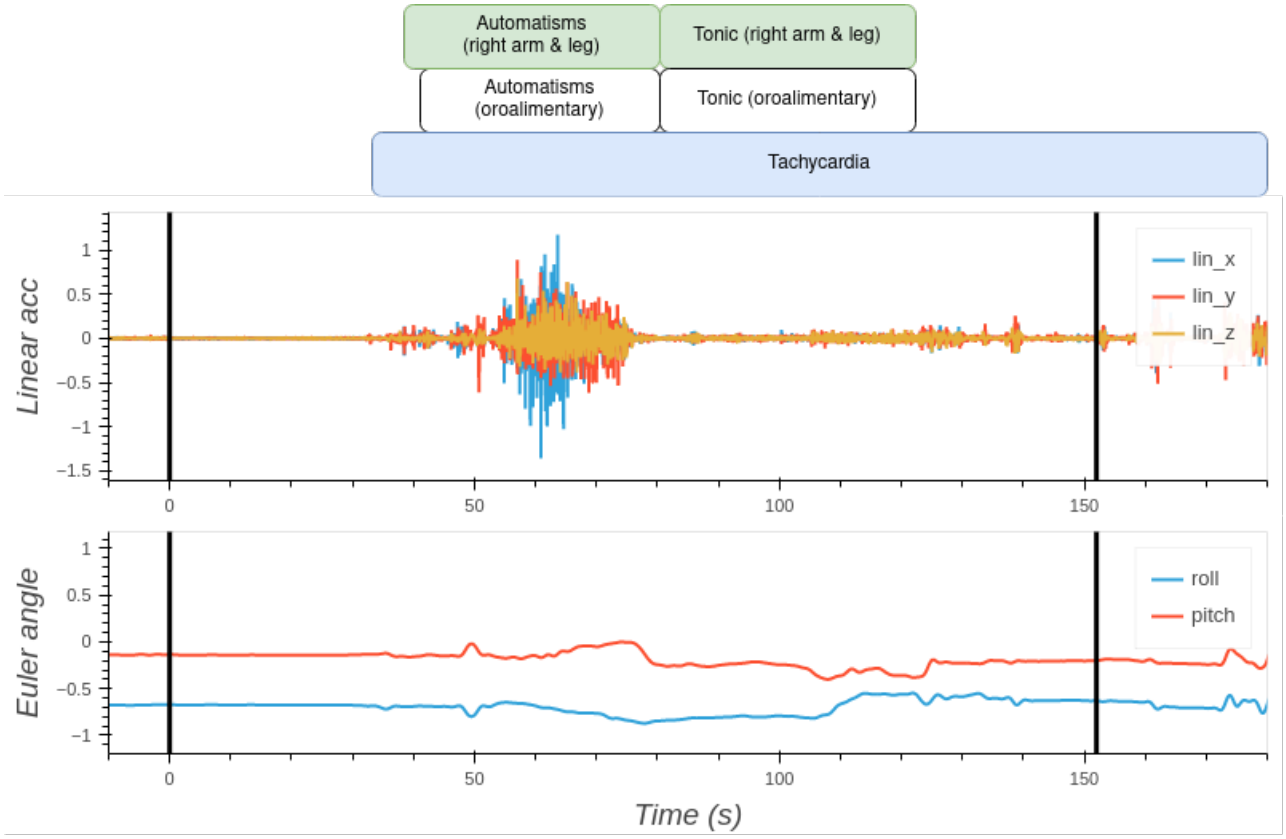


Fig. 3.2 Preprocessed accelerometry during an example seizure
The boxes at the top show the duration of components of the seizure. Green boxes are motoric components that could conceivably be captured by a wearable device. White boxes are motoric components that are unlikely to be captured. The blue coloured box represents an autonomic change. Below are the linear acceleration and roll & pitch Euler angles generated by accelerometery preprocessing.

3.2 Methods

3.2.1 Dataset

The dataset used here is derived from RADAR-EPI, a part of the RADAR-CNS research programme.[155] As was mentioned in the previous chapter, the study was conducted between July 2017 and February 2020 at the epilepsy monitoring units of two hospitals, King's College Hospital (KCH) in London and University Medical Center in Freiburg. A subset of participants in the KCH cohort used a prototype wearable device developed by IMEC, and hereafter referred to as the *IMEC device*. The device records ACC, EMG, EDA, and ECG. Participants who had a motor seizure while the device was recording are included in this study. In total 11 participants with a combined 30 motor seizures are included. Eight of these are GTCS or FBTCS. Up to an hour either side of the seizure event is included, dependent on the data being available. In total the dataset covers 50 hours of recording. Accelerometry, EMG, and EDA are included. ECG is excluded because of a high level of missingness covering the available seizures.

3.2.2 Features

Preprocessing

Accelerometry was split into two components. The gravitational component was extracted by applying a 0.5Hz low pass 5th order Butterworth filter. The linear was extracted by a high-pass filter of the same type. To reduce the dimensionality of the feature set, an approximation of pitch and roll were derived from the gravitational accelerometry signals using the equations given in the previous chapter.

A 70-500Hz bandpass filter was applied to the EMG. While there may be some useful information below 70Hz, extensive ECG artefacts and power line interference on the signal compromised the lower frequencies. In general, there is still useful information contained above 70Hz[156]

The EDA signal has a major regular artefact caused by DC polarity switching (Figure A.1) which inhibited the use of the EDA, especially with respect to the tonic level. The artefact is introduced either once per hour or once per minute depending on the recording. While several techniques were applied to reduce it, none were satisfactory and so a period of 5 minutes or 5 seconds was excised around the artefact. A band-pass filter from 0.1-2Hz was applied to produce a phasic EDA signal and a 0.02Hz low pass filter was applied to produce the tonic signal.

Feature definitions

The features chosen were on the basis of commonly reported features in the literature. Table 3.1 gives a list of the features extracted for each signal. In total 59 features were extracted in windows along the signal. A window step size of 2 seconds was used. The accelerometry (linear and Euler) features were extracted in 10 second windows. The EMG features were extracted in 2 second windows. Phasic EDA features were extracted in 30 second windows. Tonic EDA features were extracted in 300s windows.

3.2.3 Classification

Labelling

Patients with motor seizures recorded by the IMEC device were. A binary classification is assigned. The windows starting during the clinician-labelled motor seizures are assigned the positive class. All other windows are assigned to the negative class. A 5-minute post-ictal period is excluded because often the patient will interact with medical staff during this period. Autonomic changes in this period may still be included because of the large EDA window size.

Model

Classification is performed by a random forest model implemented in the scikit-learn library. [142] Default parameters were used except for $N_{estimators} = 1000$, max features per tree set to 10, and max depth set to 12. The model is trained in a leave-one-participant-out cross-validation to produce a group-wide model. To determine the viability and performance of an individual model, a leave-one-seizure-out model was trained on the data of P2. In addition to point classifications, a 5-length moving average is applied to the predicted probability along the participant's recording.

Evaluation

Evaluation metrics were given in the methods chapter. However, there are a few slight differences that are often used in seizure detection work. Sensitivity is the true positive rate, how many of the positive class are correctly predicted. Within a typical machine learning classification task, each individual data point would be considered independent. In this case, a single seizure event is split in to many individual data points. Considering a participant with two recorded seizures, if a model managed to correctly predict 40% of the individual data points, it may either be the case that an alarm could be raised for both seizures, if the

Signal	**Feature**
Accelerometry	Mean
	SD
	Skewness
	Kurtosis
	Max
	Hjorth mobility[133]
	Hjorth complexity[133]
	Hurst component[157]
	Zero-crossing rate
	Zero-crossing rate of $f''(x)$
Euler angle	Mean
	SD
	Skewness
	Kurtosis
	Max
	Min
	Hjorth mobility
	Hjorth complexity
EMG	Mean absolute value
	SD
	Zero-crossing rate
	Zero-crossing rate of $f''(x)$
	Line length
Phasic EDA	Mean absolute value
	SD
Tonic EDA	Mean
	SD
	Min
	Max
	Mean of $f'(x)'$
	SD of $f'(x)$

Table 3.1 Features used in seizure detection model.

points are spread across both events, or else only for one, if they belong predominantly or exclusively to one event. The actual metric of success that we are interested in is the true alarm rate, which will be referred to as sensitivity in this chapter, which is to say how often an alarm for a seizure event could be raised from a model's prediction.

Similarly, the standard definition for specificity, or the false positive rate, is not quite adequate. If five data points were to be incorrectly positively labelled, an alarm could be raised only once if they were all clustered in quick succession, or five times if they were spread out over the recording. Instead a false alarm rate (FAR) over a certain period of time is used. The exact circumstance under which an alarm is said to be raised or not can differ between studies. Here, a positive label within 2 minutes of another positive label is merged into a single alarm event.

PPV is an often used evaluation metric, but it is problematic because it depends entirely on how often a person has a seizure. It makes more sense to evaluate whether a model would be suitable for a participant on the basis of the sensitivity for their particular seizure type and whether it has a false alarm rate that they are able to accept.

Feature importance

Feature importance is ranked by the Gini-impurity-based importance of the feature in the random forest model.[158] The mean of the importance across all cross-validation folds is taken.

3.3 Results

3.3.1 Classification performance

The performance of the models across all participants for GTCS, other motor, and all seizures is given in Table 3.2. A visualisation of the seizure events, point classification, and the 5-point averaged probability is provided in Figure 3.3. The model correctly classifies all GTCS but only 12/22 (54.5%) focal motor seizures. A false alarm rate of 0.3/hour is calculated based on the number of merged predictions (Prediction are merged if they are within 2 minutes of each other). The distribution of false alarm events is not equal between participants. P8 has the highest rate of false alarms and a much higher baseline probability throughout much of the recording than other participants.

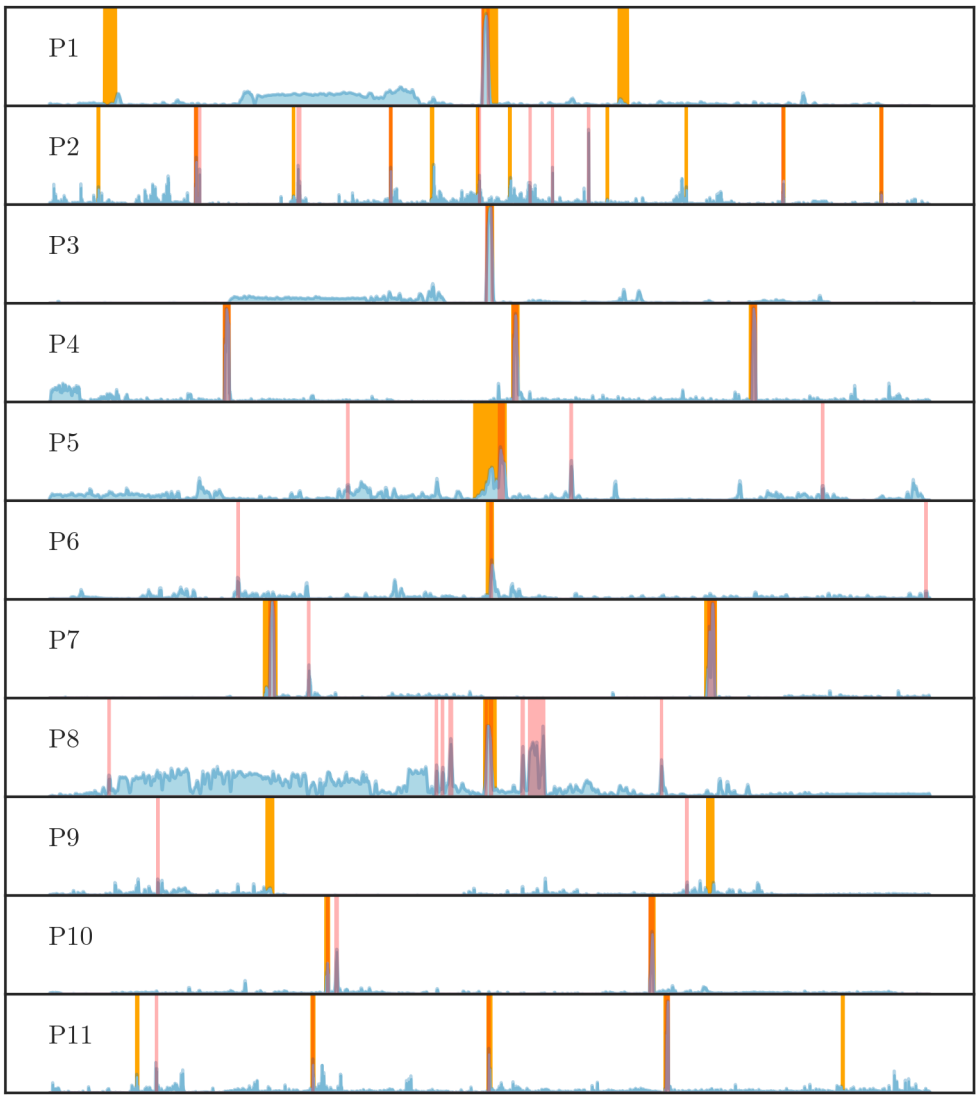


Fig. 3.3 Random forest seizure classification across all participants
A visualisation of clinician labelled v-EEG seizures (orange), per-point random forest classifications (red), and the 5-point moving average of the random forest prediction probability (blue). Each row corresponds to the recording of a single participant. The time period is not equal between participants and so the x-scale is different. The first and final seizures marked for P1 are not motor seizures and are so not included in the classification, but are close in time to the motor seizure.

Seizure type	Sensitivity	FAR (per hour)
GTCS and FBTCS	8/8 (100%)	0.3/h
Focal	12/22 (54.5%)	0.3/h
All	20/32 (62.5%)	0.3/h

FAR: False alarm rate. GTCS: generalised tonic-clonic seizure. FBTCS: focal to bilateral tonic-clonic seizure

Table 3.2 Group-wide seizure detection model results

3.3.2 Individual model performance

A higher sensitivity (9/11 vs 5/11) is achieved by the individually trained model for P2, compared to the group-wide model that did not include P2 data. The false alarm rate is higher if point predictions are considered (12 false alarms vs 5), but lower if the 5-point moving average of probability is used (2 false alarms vs 4).

3.3.3 Feature importance

The top 15 features according to the Gini-impurity-based importance measure averaged across cross-validation models are all accelerometry (Figure 3.5). Features belonging to both the linear accelerometry and the Euler angles (roll/pitch) are ranked highly. EMG features do appear, and are interspersed throughout the middle of the ranking. EDA features are resolutely at the bottom, with only two tonic features being assigned any importance.

3.4 Discussion

3.4.1 Evaluation of model performance

Motor seizure detection is viable in tonic-clonic seizures the performance in other types of motor seizure vary. Some could be particular to the participant, and so are not detected by a model trained in a leave-one-participant-out fashion where similar examples are not available. The dataset has multiple participants with (G)TCS, so not only are they one of the most extreme seizure types, the model also has multiple similar instances from other participants to train on. If the model was only trained on participants with GTCS, the FAR would probably decrease because GTCS tend to be longer in duration and with a great amount of

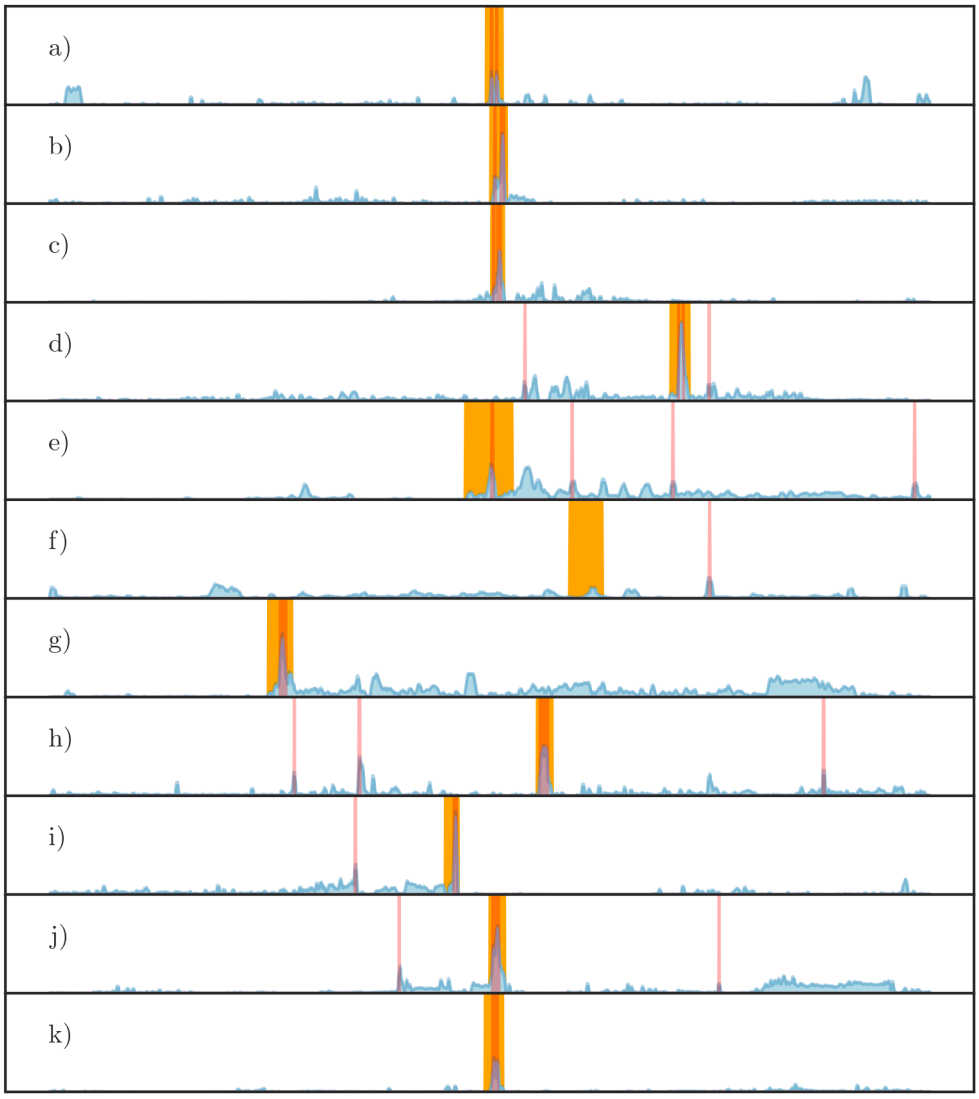


Fig. 3.4 Seizure classification in an individual model trained on P2
A visualisation of clinician labelled vEEG seizures (orange), per-point random forest classifications (red), and the 5-point moving average of the random forest prediction probability (blue). Each row corresponds to the recording around a single seizure belonging to participant P2. Up to one hour of data is included either side of the seizure, based on its availability.

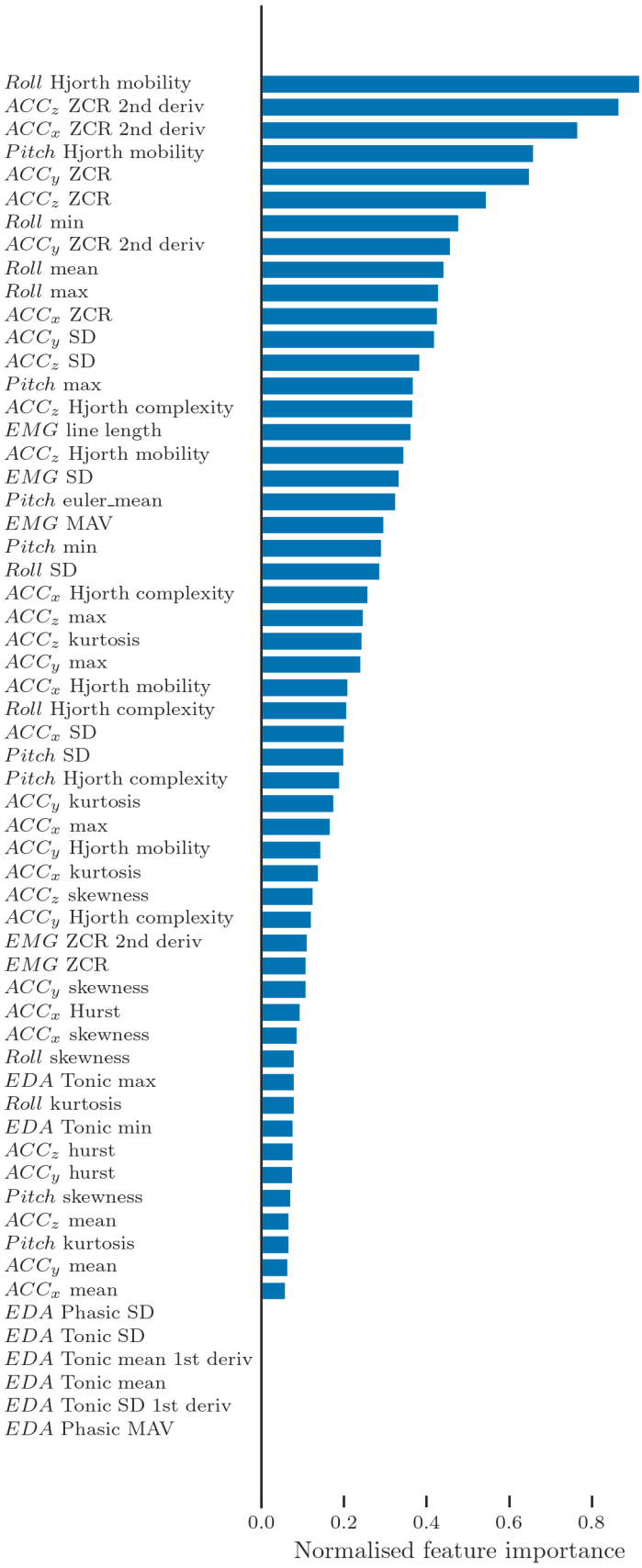


Fig. 3.5 Feature importance in group-wide models

movement. However, the GTCS classification performance is comparable already to GTCS-only studies in the literature,[159] albeit in a fairly small sample.

As well as participant-level differences in sensitivity, the false alarm rate differs between participants. P2 in particular has a high baseline prediction probability across most the recording. This is another area in which individual or personalised models may benefit, even if the seizure itself is detected. A high false alarm rate is likely to be intolerable, or to reduce the attention that is paid to an actual event because of a high exposure to false positives.

Training on a single participant in a leave-one-seizure-out manner produced a model with greater sensitivity. The vast majority of patients monitored in this study did not have enough seizure events while on the ward to train individual models despite staying as an inpatient for many days.

Important features tend to be based on the accelerometer or EMG, which is unsurprising given that all the seizures included contained a motor component and the other signals had data quality issues. The zero-crossing rate is highly represented at the top of the feature importance ranking, likely because it is a good signifier for clonic, convulsive, or otherwise highly repetitive high-frequency movement. Someone unexpectedly, the Hjorth mobility for roll and pitch were both highly ranked (1st and 4th). How the approximated angle of rotation changes over time might be a useful measure of slower acting epileptic movements, such as tonic contractions. It is not clear if they are necessarily better than simply using low-pass filtered X, Y, and Z acceleration directly, but roll and pitch are informative signals and by using them the feature set for low pass acceleration is decreased by 33%.

3.4.2 Comparison to other work

The performance of models reported in the literature are very varied. Partly they depend on seizure type, the sensors used, the study location, and the time period recorded (e.g. nocturnal only[160]). Generalised tonic-clonic seizures are often detected with high accuracy, up to 100% TPR with a low false alarm rate (FAR), both in studies based in hospitals and more recently in field studies.[161,162] Performance in ambulatory participants tends to be lower.[162]

High accuracy has been demonstrated in GTCS. An EMG-based detection algorithm achieved a 93.8% sensitivity (30/32 seizures) with low (average 9s) latency in a hospital-based study.[159] An earlier accelerometry-based study reported an 89.7% sensitivity (35/39 seizures) and low false alarm rate (0.2 per day). It was also noted that the majority of false alarms were restricted to only a few patients. A combined accelerometry and EDA based classifier also achieved high GTCS classification accuracy, with 94.55% sensitivity and 0.2 FAR/day.[163] Studies that try to classify other types of motor seizures have lower performance.

Andel et al. reported an overall sensitivity of 71% and FAR of 17.7/day in a study of motor seizures including tonic-clonic, tonic, and hypermotor seizures.[164]

The approach to modelling taken normally falls into one of two camps, either a typical machine learning classifier, or a threshold or defined algorithm. Typical machine learning classifiers, such as SVMs,[165–167] random forests,[168,169] LDA[165] are routinely used. Some papers explicitly define an algorithms,[170] or threshold a certain feature.[159,171,172] A small number of attempts have been made using deep learning or neural networks,[162] although often the datasets are too small to reasonably train a deep network.

3.4.3 Data quality

The accelerometer and EMG were both typically good. The accelerometer is in general robust, whereas even with better quality, the electrode-based signals would be sensitive to motion artefacts and detachment during seizures.

The ECG had various issues that precluded its use - EMG noise, missingness due to electrode detachment, and periods of low amplitude and high white noise. It was not used in this study because many participants did not have usable ECG around seizures. The future inclusion of heart rate metrics would likely lead to improvements in discrimination because it should be able to recognise seizure components or symptoms that are invisible to a motion sensor, such as a tachycardia, or provide the means to detect non-motor seizures.

EDA always has large artefact cause by DC polarity switching. Other devices do not have this issue, and like the ECG it may be able to pick up on autonomic symptoms that a motion sensor can not. It would therefore also be useful to include in future seizure detection algorithms, but was unfortunately not viable in this particular dataset.

3.4.4 Limitations

A limitation, common to many seizure detection studies, is that the seizures were recorded in a clinical setting, where performance likely lower in free living conditions. High performance has been reported for tonic-clonic seizure detection in field studies,[161] but it may be expected that accuracy decrease and false-alarm rate increase when a participant can move and act freely, rather than being restricted to a bed or ward.

Aspects of the pipeline could be improved. The pre-processing and feature set was limited by signal quality and availability, but similar recordings with better data quality could benefit from more robust EDA pre-processing, the inclusion of ECG, and more specific features for the EDA. Performance of a random forest model can be sensitive to hyperparameter

tuning.[173,174] Given the fairly small dataset size and the goals of the study, it was decided that fairly default hyperparameters would be set and not fine-tuned.

Several methodological limitations of the standard windowed-feature machine learning approach became apparent throughout this work. Firstly, classifying the segmented signal loses any time-dependence that is not explicitly captured in the features because the model has no concept of temporal locality. Secondly, seizures belonging to the same participant often looked similar in the physiological signal recordings. However, if the symptom was not of a common type, such as a GTCS, it is unlikely that a similar seizure from another participant will form part of training set of the model. On the other hand, there is rarely enough data from a single participant to train an individual model, with only one or two recorded events for many participants. Personalisation strategies in the literature typically depend on, sometimes manually, tuned thresholds in set algorithmic classifiers.[175] This led to considering seizure detection from a few-shot learning perspective and the use of meta-learning methods in the following chapter. Neural networks are also apt at modelling time series and so could potentially capture time-dependent information as well, addressing the first issue.

3.5 Conclusion

In this chapter I train and evaluate seizure detection models using a multi-modal machine learning pipeline. Accelerometry, and to a lesser extent EMG, features were important, but with the caveat that only motor seizures were considered and the quality of the electrode-based signals was low. Approximation of the Euler angle from the accelerometer is a potential alternative to using the 3-dimensional gravitational component of gravity that is commonly used in other seizure detection algorithms, but requires further validation and direct comparison. The processing pipeline was made into pymhealth, a reusable LLVM compiled Python library.

Epileptic seizures are an extreme example of heterogeneity and a model trained on a certain types of seizure is unlikely to generalise well to unseen variants. Even where the difference is less extreme, inter-individual variation is a common problem in machine learning approaches to medical machine learning tasks. While collecting enough data from a single person is infeasible in many applications, often a small amount of labelled data could be opportunistically collected. In the following chapter I consider a few-shot learning approach to a similar but simpler physiological classifation problem.

Chapter 4

A Meta-Learning Approach to Model Personalisation in Stress Detection

4.1 Preamble

The seizure classification problem illustrated the need for model personalisation. However, despite comparing favourably in size to other studies, the dataset size available in the epilepsy study is fairly small with limited repeat seizures. The idea of the work in this chapter was to evaluate a meta-learning approach in a simpler physiological classification task.

Stress is a psychological and physiological response to a change in external conditions and is critical to help the body meet external or internal challenges. However, prolongued or particularly intense stress can become maladaptive and implicated in various health conditions.[176] Offline stress classification using wearable data could be a useful intermediary in understanding the relationship between health outcomes and both acute stress events[177] and chronic activation of stress pathways.[176]

Stress classification has certain similarities to seizure detection. The autonomic response to acute stress shows similarities to the autonomic manifestations of some types of seizure, both typically activate the sympathetic nervous system. Many of the derived features used as variables in machine learning models in the literature are also similar. Moreover, stress is frequently reported precipitating seizures. A holistic mobile health approach to epilepsy monitoring may therefore incorporate acute stress detection.

Detection of stress using physiological signals has a long history. Early efforts included using monitored stress in drivers to inform road planning[178] and monitoring with the goal of adapting workload in various professions.[79,179–181] More recently, with the development of sensor and wearable technology there has been an increased focus on worn devices, longer-

term monitoring in everyday life, and its use in healthcare. Particularly with the release of public datasets like DriveDB[79] and WESAD,[78] there has been a large output of studies. There are a couple of useful reviews that cover the published stress studies[182] and the machine learning approaches used,[183] while the approaches of the most pertinent studies will be briefly described here.

A major problem in stress detection is the inter-individual variation in the measured stress response. Several techniques to personalise models and their apparent performance increase over general models had been reported. A common approach has been to train a separate model for each participant.[184,185] Saeed and Trajonvski used multi-task learning, using a shared neural network connected to a subject-specific classification layer, on DriveDB and found better performance than a single-task neural network.[186] Another approach used clustering to group participants are train models for each group.[187] While each of these studies reported an increased performance in their personalised models, there are important limitations. Primarily, each trained their models on a random split of a participant's data. Physiological time series exhibit autocorrelation and so the improved performance may be partially explained by correlations between the train and test data, particularly where there are overlapping feature windows, rather than an increased ability to differentiate stress. Secondly, several approaches required training a personlised model in combination with the full dataset which could make adding new people computationally expensive.

Using public stress classification datasets, in this chapter I test whether providing a small amount of 'contextual' data from a baseline recording can improve the classification performance for that participant in the subsequent recording. Importantly the person-specific training samples are only taken from the beginning of the recording, reducing the confounding effect of correlation between nearby points. Additionally, 'personalising' a neural process only requires a forward pass of the contextual data through an encoder network, reducing the computational cost of adaption to new people. The aim of this study was to assess the viability of a meta-learning approach in a physiological dataset and to develop a per-participant stress classification algorithm.

The chapter is included as a pre-print of a paper that is currently under consideration.
DOI 10.48550/arXiv.2002.04176

Personalized acute stress classification from physiological signals with neural processes

Callum L Stewart[1]
callum.stewart@kcl.ac.uk

Amos Folarin[1,2]
amos.folarin@kcl.ac.uk

Richard Dobson[1,2]
richard.j.dobson@kcl.ac.uk

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, U.K.
[2] Institute of Health Informatics, University College London, London, U.K.

February 12, 2020

Abstract

Objective A person's affective state has known relationships to physiological processes which can be measured by wearable sensors. However, while there are general trends those relationships can be person-specific. This work proposes using neural processes as a way to address individual differences.

Methods Stress classifiers built from classic machine learning models and from neural processes are compared on two datasets using leave-one-participant-out cross-validation. The neural processes models are contextualized on data from a brief period of a particular person's recording.

Results The neural processes models outperformed the standard machine learning models, and had the best performance when using periods of stress and baseline as context. Contextual points chosen from other participants led to lower performance.

Conclusion Neural processes can learn to adapt to person-specific physiological sensor data. There are a wide range of affective and medical applications for which this model could prove useful.

1 Introduction

Wearable devices are increasingly used in affective computing because they provide continuous information on a person without requiring their attention;

furthermore, some disorders or affective states have known relationships to measurable physiological processes [1, 2] and are therefore candidates for remote monitoring. However, symptoms of disease and manifestations of affect can differ from one person to another, hindering the generalizability of models within mobile health and affective computing.

1.1 Stress background

Stress is a natural collection of responses to a change in homeostasis or a perceived threat. Stressful stimuli can elicit a variety of different behavioral, emotional, and physiological responses. The physiological response is predominantly mediated by the hypothalamic-pituitary-adrenal (HPA) axis and the autonomic nervous system (ANS) [3], which in turn affect a range of physiological functions. In particular, the correlation between stress and heart rate and galvanic skin response (GSR) has long been known [4]. Recent developments of wearable physiological sensors provide the ability for continuous, long-term, passive, and remote measurement. Their use, therefore, allows for an objective measure of systems mediated by the ANS and HPA in response to acute stress.

Accurate detection of stress has wide-ranging application. It could be used in intelligent feedback systems, altering the system's behavior in response to stress [5]; as part of a system monitoring the progression of diseases or disorders with a known relationship with stress, such as depression [6]; detecting and managing maladaptive stress [7]; or measuring response to medication or therapy.

1.2 Machine learning personalization strategies

Model personalization acknowledges that a single medical problem can have a diverse range of outcomes and symptoms between individuals, and attempt to improve model performance by making it specific to an individual. This has been approached in a number of ways; Non-exhaustively they include training completely separate models for each individual, selecting different features, setting personalized cut-off thresholds, additional training or hyperparameter selection of a general model using an individual's data, and clustering individuals and creating a model for each cluster.

There are problems with some of the traditional personalization methods. Most generally, a lot of relevant data can be ignored if only a subset of a cohort is used to build a model, and the collection of adequate person-specific data is often time-consuming and expensive. Few-shot and meta-learning techniques developed in adjacent fields, such as image classification, offer potential frameworks to approach the problem of personalizing models.

1.3 Meta-learning and related approaches in biomedical datasets

Various few-shot and meta-learning techniques have been developed recently. They are typically first used in generic open access few-shot datasets but have

found some application in low-data domains like medical imaging.

Non-parametric or metric based networks, such as siamese networks and matching networks [8], learn a distance metric or comparable embeddings between input vectors. This method of few-shot deep learning appears to have had the greatest uptake in biomedical problems, with applications in seizure detection, histopathology [9], drug discovery [10], and fall detection [11], among others.

Optimization-based meta-learning, exemplified by MAML [12], aim to learn a set of initial parameters which can be quickly adapted to a new dataset through few additional gradient steps. They have found some promising use in low-data medical image classification tasks [13, 14], but are typically not used for personalizing a model to an individual in a longitudinal dataset.

Another technique, broadly categorized as parameterizing or black-box meta-learning, consists of a classification network and an encoder network which is used to parameterize the classifier. An example is neural processes [15]. Neural processes are used as personalizable models in this study. They can be thought of as a distribution of functions, parameterised by a few 'context' x-y pairs. If we consider there to be an underlying biological trend to affective states or disorders, which manifest differently depending on the individual and their context, then a neural process forms a distribution of functions over the general trend which can be parameterized for a person using a small number of x-y pairs from that individual. Once trained on a meta-training set, individualization is provided only at the cost of a forward pass through the encoder, the original training data is no longer required.

1.4 Study datasets and objective

This work uses meta-learning techniques developed for few-shot learning to individualize models used for classifying periods of stress in participants from two publicly available datasets: Stress Recognition in Automobile Drivers (drivedb) [16] available from physiobank [17] and Wearable Stress and Affect Detection (WESAD) [18]. Both datasets contain continuous electrocardiogram (ECG) and galvanic skin response (GSR) recordings in healthy participants during a series of tasks designed to elicit an affective response. WESAD is a public dataset for affect and stress detection using motion and physiological recordings, including ECG and GSR. In addition to a task inducing stress, it includes an amusement task which is negatively labeled for our binary stress classification models. Drivedb is a dataset with multiple sensor recordings, including ECG and GSR, taken while a healthy participant drives on a predefined route containing sections of in-city driving assumed to be stressful, and sections of highway driving assumed to be relatively less stressful.

We investigate the applicability of neural processes (NPs), as a representative of meta-learning techniques, to personalization in a biomedical problem. Firstly, the performance of general models built with k-Nearest Neighbors (k-NN), support vector machine (SVM) with a radial basis function kernel, and an L1-regularized logistic regression (Lasso), are compared against a neural process individualized with either baseline-only or randomly chosen context

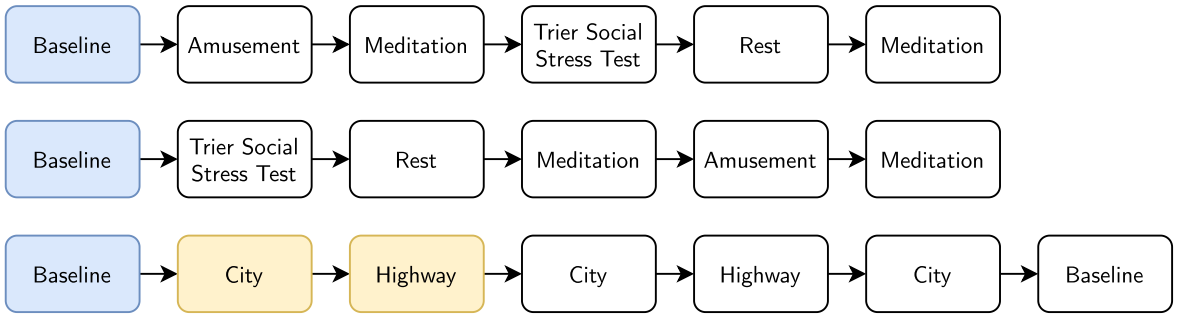


Figure 1: Study protocols for datasets. The two protocol variations in WESAD (top, middle) and the route driven in drivedb (bottom).

pairs. Secondly, a baseline-only personalization is compared against a neural process which uses the first highway and city driving segments in addition to the baseline in the drivedb dataset, which includes repeated sections intended to induce stress (city) and relatively reduced stress (highway). The sections used as context points are excluded from the test performance.

2 Methods

The WESAD dataset contains 15 participants (age $= 27.5 \pm 2.4$) with an average recording duration of 96 minutes. Two affective responses, stress and amusement, are evoked in two tasks. There is a baseline period, rest period and two meditation tasks 1. Two devices are worn by participants, but only ECG and GSR signals from the chest-worn RespiBAN are used here. Drivedb contains recordings of 17 participants lasting between 65 and 93 minutes, depending on the road conditions during the test. Of these, only 13 are used (4-16) because of missing data or unclear label markers. Again, only the ECG and GSR signals are used, collected from a custom wearable system.

2.1 Processing and feature extraction

Manually defined features are used to facilitate comparison between general machine learning models and the neural process models, and to evaluate whether the neural process is able to use data representative of an individual to improve performance rather than the ability of a neural network to learn representations from raw or preprocessed signals.

The Hamilton-Tompkins algorithm is used to detect the R peaks in the ECG signal [19]. Tonic and phasic GSR are filtered from the raw GSR signal using a 0.2Hz lowpass and a 0.5-2Hz bandpass filter respectively; both are 5th order Butterworth filters. Participants in the drivedb dataset can contain GSR recorded at either the hand, foot, or both. Only one is used, and the hand GSR

Signal	Feature	Equation / Reference
ECG	HR range	$max(HR) - min(HR)$
ECG	HR mean	$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
ECG	HRV SDNN	[20]
ECG	HRV RMSSD	[20]
ECG	HRV CSI	[21]
ECG	HRV sample entropy	[22]
ECG	RQA determinism	[23]
ECG	RQA length entropy	[23]
ECG	HRV LF absolute power	[20]
ECG	HRV LF relative power	[20]
ECG	HRV LF peak frequency	[20]
ECG	HRV HF absolute power	[20]
ECG	HRV HF relative power	[20]
ECG	HRV HF peak frequency	[20]
ECG	HRV HF/LF ratio	[20]
GSR tonic	Mean	$\frac{1}{N}\sum_{i=1}^{N} x_i$
GSR tonic	SD	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
GSR tonic	1st deriv. mean	$\frac{1}{N}\sum_{i=1}^{N} x'_i$
GSR tonic	1st deriv. SD	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x'_i - \bar{x'})^2}$
GSR phasic	SD	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
GSR phasic	mean absolute value	$\lvert\bar{x}\rvert = \frac{1}{N}\sum_{i=1}^{N} \lvert x_i\rvert$

Table 1: Feature definitions

is preferred if it is available.

Typical heart rate, heart rate variability (HRV), phasic GSR, and tonic GSR features are extracted in 40s windows with 20s overlap from each participant 2.1. The class label for the window is determined by the largest proportion of time spent in either the stressful or relaxing task. Features are min-max scaled between -1 and 1 for all general models, but are not scaled for the neural processes.

2.2 General model

Both general and personalized models are trained and tested using leave-one-participant-out cross-validation. Traditional machine learning models were built using scikit-learn [24], using default hyperparameters. Three general models are used: A logistic regression with l1 penalization and $C = 1.0$, a radial basis function kernel SVM with gamma = 0.0526 ($1/N_{features}$), and a 20-neighbor k-NN classifier.

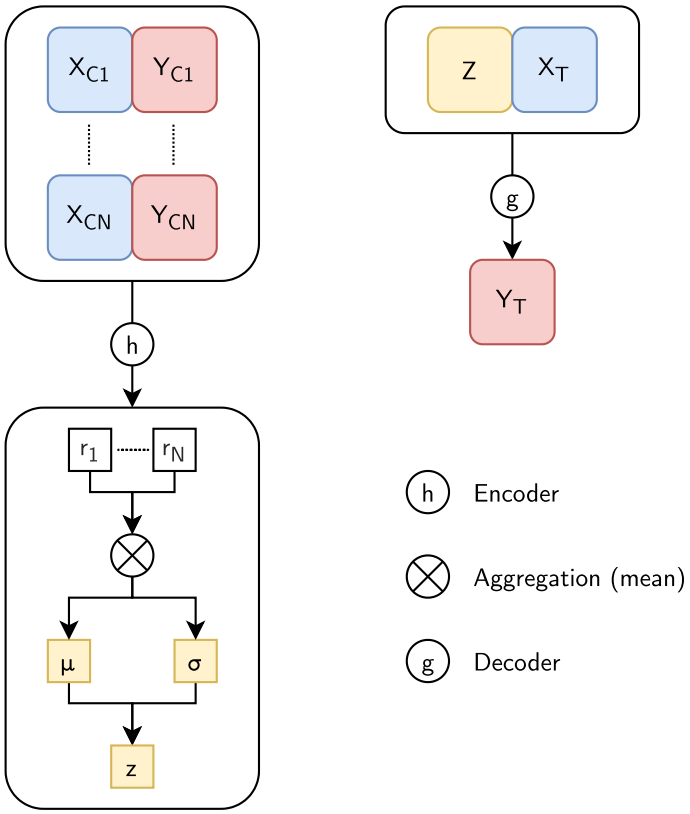


Figure 2: A general view of a neural process model. The encoder (h) takes X-y pairs and transforms them into a latent distribution. A sample from that distribution is concatenated with unlabeled data and passed to the decoder (g) which predicts class labels for the unlabeled data.

2.3 Personalized neural process models

A neural process is a latent variable neural network which aims to adapt to specific context data at test-time. It is formed of three components: an encoder $h(r_i|x_{ci}, y_{ci})$ which transforms a set of contextual input data pairs (x_c, y_c) into a representation r, an aggregator which aggregates multiple representations from the encoder into a single vector which is used to parameterize a latent distribution z, and a decoder $g(y_t|x_t, z)$ which samples z and transforms unlabelled data x_t into a predicted value y_t 2. The model is built using the pytorch library [25].

The specific architecture used here consists of an encoder of 3 dense hidden layers each with 30 nodes, a mean aggregator and a latent variable with 15 nodes, and a decoder with 3 dense 30 node hidden layers and a single output node. The decoder has a dropout rate of 0.2. Each model is trained on the data of all but one participant.

During training, the data of each training participant is looped through.

Between 5 and 10 points are randomly chosen as the context points. All of the data belonging to the current training participant is used as target points. Both context (xy_c) and target (xy_t) x-y pairs are passed through the encoder to create latent distributions Z_c and Z_t respectively. The encoded context points (Z_c) are concatenated with the unlabeled target points (x_t) and passed through the decoder to predict the target label ($\hat{y}_t$). As well as minimizing the binary cross-entropy between the predicted target points ($\hat{y}_t$) and the true values (y_t), the Kullback Leibler divergence between the distributions of the encoded context points (Z_c) and the encoded target points (Z_t) is minimized (3).

At test time the context points are unique from the target points and selected according to the personalization strategies mentioned in the following paragraph. In each case 6 x-y data pairs are used as the context points. Any period or task from which the context points are chosen are not used as target points. Stress predictions for the remaining data for the participant are calculated.

2.4 Personalization strategies

Three methods for selecting context points during testing are chosen. Firstly, each model is personalized using context points selected only from the baseline segment. Secondly, context points are randomly selected from the entire recording with a uniform distribution. Thirdly, two points from each of the baseline recording and the first occurrence of the city (stress) and highway (non-stress) driving segments are used as context points. Data from the recording following the sections chosen for context points; two city driving sections, a highway section, and a relaxation section; are subsequently predicted using the personalized model. Because the WESAD dataset only has a single stress assessment task, only the drivedb dataset can be used in the third personalization strategy.

2.5 Performance metrics

Three performance metrics are reported here: the area under the curve (AUC) of the receiver operating characteristic (ROC), the average precision, and the log-loss. Because of the class imbalance in the WESAD dataset and the differences in class proportion between datasets, the average precision may be more informative than the AUC.

3 Results

Both the baseline-only and randomly chosen context NPs perform better than all of the general models, with the randomly chosen context performing best 3. Randomly chosen points, while not included in the test scores themselves, are likely to be strongly correlated with points from the surrounding time and from the same task. The third personalization strategy, in which the context points are selected from the baseline and first occurrence of each task, is used to address this problem in models for drivedb participants. Similarly to the

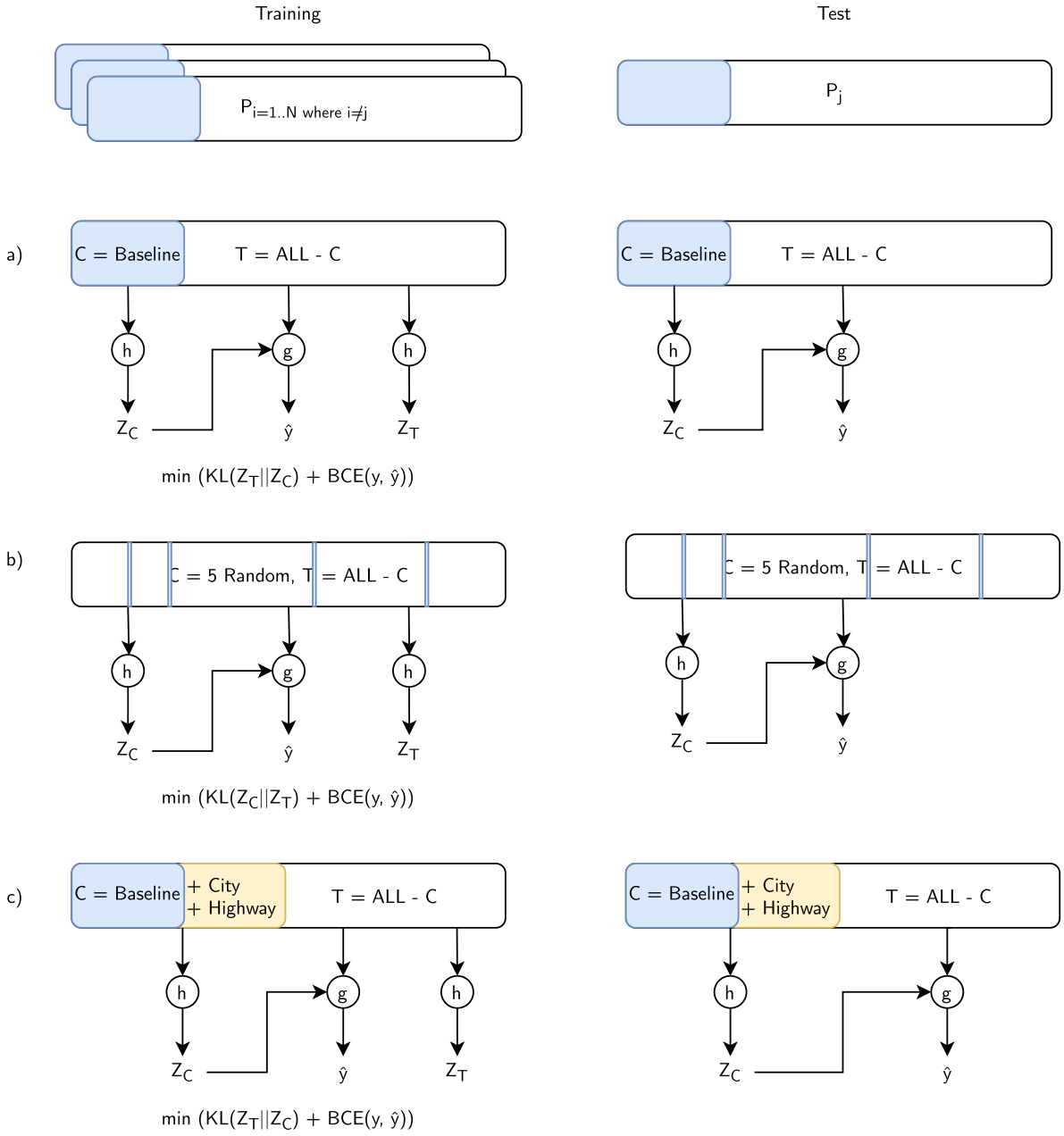


Figure 3: Context strategies for training and testing in each participant's (j) neural process model. h: Encoder, g: Decoder, BCE: Binary cross entropy, KL: Kullback–Leibler divergence. In each case, the training loss is the sum of the Kullback-Leibler divergence between the two distributions formed by target and context points passing through the encoder and the log-loss between y and y-pred. a) Context points are taken from the baseline recording, target points are taken from the remaining data. b) Context points are randomly selected using a uniform distribution. c) Drivedb only - Context points are taken from the baseline, first city, and first highway sections. An equal number of points

Model	AUC	Average precision	Log loss
Lasso	0.954	0.881	0.222
SVC (RBF kernel)	0.943	0.882	0.234
K-Nearest Neighbors	0.870	0.740	0.563
NP (baseline)	0.970	0.924	0.182
NP (random choice)	0.984	0.957	0.133
NP (other participant)	0.880	0.780	0.470

Table 2: WESAD dataset results

Model	AUC	Average precision	Log loss
Lasso	0.695	0.736	0.669
SVC (RBF kernel)	0.704	0.733	0.645
K-Nearest Neighbors	0.680	0.721	3.09
NP (baseline)	0.776	0.797	0.570
NP (tasks)	0.787	0.804	0.553
NP (other participant)	0.722	0.757	0.663

Table 3: Drivedb dataset results

previous results, the personalized NP models perform best and the models which include stress-task context perform better than those which use baseline-only data 3, although the improvement in comparison to the baseline-only models is much slighter.

Using the neural process models with context and target points selected from different participant results in greatly reduced performance (WESAD: 0.957 average precision, same-participant vs 0.780 other-participant, drivedb: 0.804 vs 0.757), indicating that the neural processes do rely on and gain benefit from individual-specific data points 4. The increase in performance between the general and personalized models appears partly due to a decrease in variance of performance between each participant's model 5. Each of the general models have a subset of low-performing participants, although they do not completely overlap. The distribution of performance over participants for the general models contains a large number of participants with very high performance, 0.9+ accuracy, and a tail of drastically lower performing participants. The personalized NP models performance is mostly improved through increased performance on those lower performing participants. Within the WESAD cohort, the average accuracy between the NP and lasso models was increased by 0.0995 for those participants whose accuracy was less than 0.9 in the general lasso model, compared to a decrease of 0.002 for those above 0.9 in the lasso model. The drivedb dataset shows a broadly similar pattern; participants with an accuracy score in the general lasso model lower than the average across the dataset (0.66) have an average accuracy increase of 0.117, compared to an increase of 0.035 in those who had lasso model accuracies above the average.

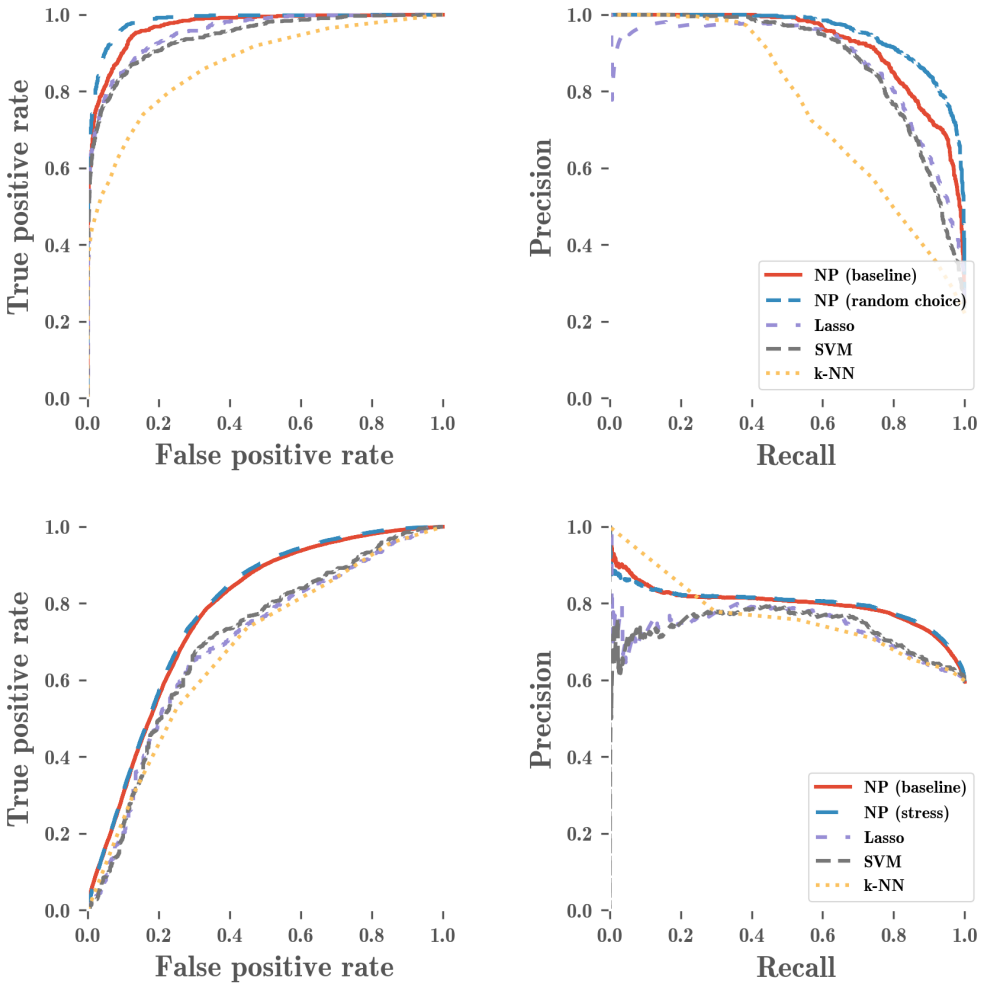


Figure 4: Receiver operating characteristic (left) and precision-recall (right) plots for WESAD (top) and drivedb (bottom) models.

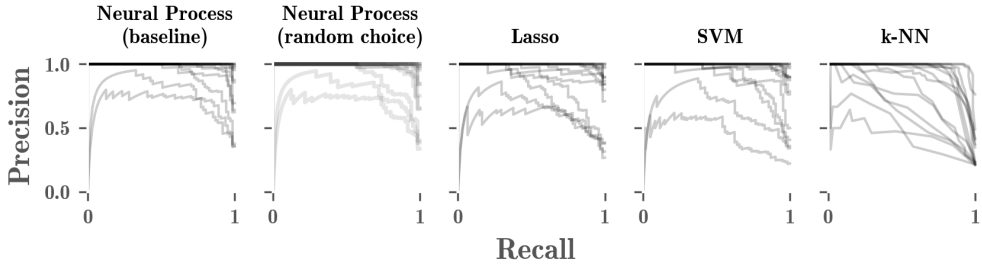


Figure 5: Precision-recall plots for each model in which models belonging to participants from the WESAD dataset are drawn as individual lines.

4 Discussion

Overall the personalized models performed better than the general models. Encouragingly, the improvement between personalized and general models is most marked in those participants with lower performance in the general models. The improved performance of the lower performing subgroup in the NPs suggests that by using a small amount of person-specific data, a base model can be successfully used in a greater range of people who are dissimilar to the training cohort or who are somehow divergent in comparison to the majority. The demographics for the participants in the two studies was quite homogenous; given a broader population, the usefulness of a more flexible or personalizable model may be greater.

There are two remaining participants with low-performing personalized models, visibile in 5 and both belonging to the WESAD cohort. One has low performance across all models, where no model can differentiate the amusement and stress task well. The second is particular to the NP model, and appears to be due to an atypical baseline recording, which includes a large tonic GSR amplitude, highlighting the importance of contextual data that is typical of the class it is representing. If the first few minutes of the recording are discarded, and a portion of the first rest period selected as the context, the performance of the model is similar to other participants.

Model personalization methods previously used in stress detection studies have been typically achieved through personal feature normalization [26], training a model on one participant's data[27, 28, 29, 30], or training models for groups of similar participants [27, 31]. Neural processes have a number of theoretical advantages over these methods; they do not assume that personal differences in features can be reduced to a linear proportion of a baseline measurement, they can make use of the entire dataset of participants, only a small number of labelled data points are required to personalize a model, and the computational cost of personalization is only a single pass through the encoder network.

The importance of correct use of cross-validation or training splits is demonstrated in the literature. High performance can be achieved when an individual model is trained using random cross-validation [31, 30] because temporally close data points will be highly correlated. This is also seen in the neural process models, in which randomly sampled context data points outperform context from a single task. For the purpose of building personalized models, it is therefore necessary to have a dataset with multiple assessments per participant, or to personalize based on unlabeled or negative case data.

In general, using meta learning techniques additional medical datasets with similar tasks and signals could be incorporated. Aggregation of similar small datasets could lead to improved performance for each individual task. To combine multiple tasks along with personalization through meta-learning, it may be necessary to pose the meta-learning procedure in multiple levels or hierarchically [32], where more prior knowledge is shared between individuals in the same task than between the different tasks.

In this study features are manually defined and extracted from the raw signals

because the objective was to discover whether personalization through neural processes is possible and useful, rather than looking at whether a neural network can learn a better feature representation. However, meta-learning can allow more sophisticated deep learning techniques and feature extraction where they would otherwise be intractable because of small dataset sizes. Addition of a neural network to learn features is therefore a prominent area to potentially improve performance.

That only baseline data points used as context can improve performance suggests that a representation built on unlabeled or weakly labeled data may be viable. Particularly in long-duration recordings, much of the data in biomedical datasets can be unlabeled or with a very large imbalance between positive and negative cases. Going forward, it would therefore be useful to be able to personalize a model based on that unlabeled data. Where the representations from the context data points are currently mean aggregated, it may make more sense to have an aggregation that recognizes the time-dependent nature of the data. Additionally, in the future it would be useful to compare the performance of different personalization techniques, both optimization-based deep learning and classical machine learning, against the neural processes demonstrated here.

5 Conclusion

Neural processes, presented as a method to generate personalized models, outperform general classic machine learning algorithms in stress detection tasks across two datasets and appear to use small amounts of person-specific context data to improve performance. Using only baseline data as context is useful, but the inclusion of data with the positive-label class further improves performance. The datasets used here concern affect and stress classification, but there are applications beyond: many problems in medicine and biology have large inter-individual differences or heterogeneity in classification which could be addressed using neural processes or similar methods. There is also a large space for improvement in various aspects of the modeling procedure.

Acknowledgments

This paper represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

References

[1] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia. Acute mental stress assessment via short term HRV analysis in healthy

adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, 18:370–377, April 2015.

[2] Joachim Taelman, S. Vandeput, A. Spaepen, and S. Van Huffel. Influence of Mental Stress on Heart Rate and Heart Rate Variability. In R. Magjarevic, J. H. Nagel, Jos Vander Sloten, Pascal Verdonck, Marc Nyssen, and Jens Haueisen, editors, *4th European Conference of the International Federation for Medical and Biological Engineering*, volume 22, pages 1366–1369. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[3] Yvonne M. Ulrich-Lai and James P. Herman. Neural regulation of endocrine and autonomic stress responses. *Nature Reviews Neuroscience*, 10(6):397–409, June 2009.

[4] Richard S. Lazarus, Joseph C. Speisman, and Arnold M. Mordkoff. The Relationship Between Autonomic Indicators of Psychological Stress: Heart Rate and Skin Conductance:. *Psychosomatic Medicine*, 25(1):19–30, January 1963.

[5] Olga C. Santos, Raul Uria-Rivas, M. C. Rodriguez-Sanchez, and Jesus G. Boticario. An Open Sensing and Acting Platform for Context-Aware Affective Support in Ambient Intelligent Educational Settings. *IEEE Sensors Journal*, 16(10):3865–3874, May 2016.

[6] Neil Schneiderman, Gail Ironson, and Scott D. Siegel. Stress and Health: Psychological, Behavioral, and Biological Determinants. *Annual Review of Clinical Psychology*, 1(1):607–628, April 2005.

[7] J. Bakker, M. Pechenizkiy, and N. Sidorova. What's your current stress level? detection of stress patterns from gsr sensor data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 573–580, Dec 2011.

[8] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *arXiv:1606.04080 [cs, stat]*, December 2017. arXiv: 1606.04080.

[9] Alfonso Medela, Artzai Picon, Cristina L. Saratxaga, Oihana Belar, Virginia Cabezon, Riccardo Cicchi, Roberto Bilbao, and Ben Glover. Few Shot Learning in Histopathological Images:Reducing the Need of Labeled Data on Biological Datasets. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1860–1864, Venice, Italy, April 2019. IEEE.

[10] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science*, 3(4):283–293, April 2017.

[11] Diego Droghini, Fabio Vesperini, Emanuele Principi, Stefano Squartini, and Francesco Piazza. Few-Shot Siamese Neural Networks Employing Audio Features for Human-Fall Detection. In *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence - PRAI 2018*, pages 63–69, Union, NJ, USA, 2018. ACM Press.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, July 2017. arXiv: 1703.03400.

[13] Gabriel Maicas, Andrew P. Bradley, Jacinto C. Nascimento, Ian Reid, and Gustavo Carneiro. Training Medical Image Analysis Systems like Radiologists. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11070, pages 546–554. Springer International Publishing, Cham, 2018.

[14] Xiang Jiang, Liqiang Ding, Mohammad Havaei, Andrew Jesson, and Stan Matwin. Task Adaptive Metric Space for Medium-Shot Medical Image Classification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2019*, volume 11764, pages 147–155. Springer International Publishing, Cham, 2019.

[15] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural Processes. *arXiv:1807.01622 [cs, stat]*, July 2018. arXiv: 1807.01622.

[16] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, June 2005.

[17] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23), June 2000.

[18] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 2018 on International Conference on Multimodal Interaction - ICMI '18*, pages 400–408, Boulder, CO, USA, 2018. ACM Press.

[19] Patrick S. Hamilton and Willis J. Tompkins. Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database. *IEEE Transactions on Biomedical Engineering*, BME-33(12):1157–1165, December 1986.

[20] Fred Shaffer and J. P. Ginsberg. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, 5:258, September 2017.

[21] Jesper Jeppesen, Sandor Beniczky, Peter Johansen, Per Sidenius, and Anders Fuglsang-Frederiksen. Using Lorenz plot and Cardiac Sympathetic Index of heart rate variability for detecting seizures for patients with epilepsy. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4563–4566, Chicago, IL, August 2014. IEEE.

[22] Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, June 2000.

[23] N Marwan, M Carmenromano, M Thiel, and J Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, January 2007.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *arXiv:1201.0490 [cs]*, June 2018. arXiv: 1201.0490.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas KÃ¶pf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv:1912.01703 [cs, stat]*, December 2019. arXiv: 1912.01703.

[26] Jonathan Aigrain, Severine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani. Person-specific behavioural features for automatic stress detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, Ljubljana, May 2015. IEEE.

[27] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1053–1060, July 2016.

[28] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. Call Center Stress Recognition with Person-Specific Models. In Sidney D'Mello, Arthur

Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, volume 6974, pages 125–134. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[29] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. StressSense: detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, page 351, Pittsburgh, Pennsylvania, 2012. ACM Press.

[30] Kizito Nkurikiyeyezu, Kana Shoji, Anna Yokokubo, and Guillaume Lopez. Thermal Comfort and Stress Recognition in Office Environment:. In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 256–263, Prague, Czech Republic, 2019. SCITEPRESS - Science and Technology Publications.

[31] Saskia Koldijk, Mark A. Neerincx, and Wessel Kraaij. Detecting Work Stress in Offices by Combining Unobtrusive Sensors. *IEEE Transactions on Affective Computing*, 9(2):227–239, April 2018.

[32] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically Structured Meta-learning. *arXiv:1905.05301 [cs, stat]*, November 2019. arXiv: 1905.05301.

4.8 Summary

One of the aims of the thesis was to consider how baseline or small quantities of person-specific data may improve performance of algorithms built for problems that exhibit inter-individual variation. While conceptualised in the epilepsy study, where there is a larger variety of inter-individual seizure manifestions than in stress, this study provided some evidence for the use of a meta-learning based approach to model personalisation. Conditioning a neural process on baseline data in WESAD or baseline and the first stress event in DriveDB did improve the performance compared to general models or conditioning on other participant's data. Importantly, the model was conditioned on contextual points that were separated in time from the test points, which has not been common in the literature.

The use of a meta-learning neural process to address inter-individual variation in a biosignal classification task was novel, but there is a large field of meta-learning and few-shot learning methods. Gradient-based approaches, such as MAML,[188] have been more widely applied,[189,190] and in the intervening period advances have been made within the neural process family[191] over the conditional neural process used in this study. In the papers that introduce them, NPs have competitive performance with other meta-learning methods, but a robust comparison of methods among digital health tasks would be useful. The preference for NPs here were due to a couple of factors. Firstly, the probabilstic prediction gives an error margin which is an attractive property in medical tasks. Secondly, the task-adaption only requires a forward-pass through an encoder, rather than additional gradient descent steps and the storage of per-person updated weights, which may be advantageous in an environment with constrained resources.

The following chapters change focus to a crowdsourced COVID-19 study, firstly the setup and software development required and secondly looking for signs of and risk factors for long COVID in mobile health data. While addressing separate areas of the mobile health study pipeline, one continuing theme is the use of baseline data and is part of the reason I felt it was important to include the retrieval of historic wearable data. While using a different analytical approach, the

Chapter 5

Mass Science: Software Development and Participant Engagement

5.1 Introduction

The sudden onset of the COVID-19 pandemic in late 2019 stimulated a flurry of research across academia. It necessitated a rapid reaction to understand the disease and develop treatments and management plans from the beginning of the pandemic. Prior to the pandemic, I had considered attempting to run a citizen science project looking at depression and had some prototype development of some of the app components necessary to support one. As the important of COVID-19 became clear, we redeveloped that work, along with parts of RADAR-base, to try and quickly set up a crowd-sourced COVID-19 mHealth study, which was ultimately named Covid Collab.

Mobile health received a reinvigoration of an already upward trajectory due to the pandemic, both as a tool for the delivery of treatment and management of conditions in general as well as a method for original research and investigation into COVID-19. The clear unique advantage of /acmhealth in this context is its remote nature; at a time when people were encouraged to socially distance, mobile health (mHealth) was an avenue for the delivery of care and research which did not require a person to put themselves at risk of infection. There are, in addition, multiple other advantages. Those include the ability to engage participants who may not be identified through clinical or other traditional recruitment practices; access historic or baseline data from existing repositories, such as fitness tracking data, from before enrolment or the start of the pandemic; and a continuous and ubiquitous source of data was especially useful in a disease for which we did not have prior knowledge of progression or

duration. Moreover, the high level of public interest in COVID-19 likely led to increased interest in research and particularly citizen science initiatives that focused on COVID-19.

The primary aim of this chapter is to describe the rapid development of an application in response to the COVID-19 pandemic, which included the ability for remote enrolment and interoperability with RADAR-base components. Additionally, I investigate participant engagement and adherence in the Covid Collab study.

5.1.1 Mass Science Application

In light of the above circumstances, we wished to run an mHealth study on COVID-19 that was open for enrolment to the general public and could be participated in remotely and without direct involvement from researchers. *Covid Collab*,[192] the resulting study, is the subject of subsequent chapters, but it required software development work and provided a view of participant engagement behaviour in a remote study which are the focuses of this chapter.

The RADAR-base platform,[8] used for the data collection in previous mHealth studies including the Epilepsy dataset in this thesis, included many useful components. However, because it was developed with in-person and clinical enrolment in mind, there were several shortcomings which made it unsuitable for a remotely enrolled crowd-sourced study. To address those issues in RADAR-base, the *Mass Science* application and accompanying backend infrastructure were developed. The Mass Science application is a cross-platform mobile app developed using the Flutter framework.[193] It takes many of the features of the RADAR-base active RMT app,[8] such as the general structure and display of active tasks, as well as making use of the questionnaire protocols created for RADAR-base. While designed primarily with the aim of launching a COVID-19 monitoring study, it was made to be straightforwardly used in other studies and to interoperate to some degree with components of the RADAR-base platform. In addition to the Covid Collab study, Mass Science was later used in the Convalescence long COVID study.[194]

5.1.2 Participant engagement

Attrition of users is a problem common in the mobile sphere in general,[195] with fitness and mobile health apps commonly abandoned within the first three months of use.[196,197] Understanding and encouraging adherence and engagement in mHealth studies is an area of increasing study because low adherence or abandonment can cause data quality issues that undermine further analysis: from biasing the usable data towards diligent participants who may not be representative of the wider population; to incomplete data requiring imputation or pruning; and reducing study power. When participant engagement differs between groups,

it is hard to determine whether differences to the outcome variable between groups are true effects or whether it is due to differences in retention.

Engagement and retention in mHealth studies

Various factors are important to consider when aiming to engage and retain a representative population in an mHealth research study. The low barrier for participants to enrol in mHealth studies can enable large, wide-ranging studies with representative populations, but that low barrier to entry can translate to a weaker bond to the study and a lower barrier to abandonment.

Socio-demographic characteristics often differ between recruited participants and the general population. Age,[196] gender,[198] ethnicity,[199] economic status, and education[200] have all been put forward as factors that may affect a person's motivation to enrol or continue to engage with mHealth studies and there are often discrepancies between socio-demographic characteristics in the study population and the wider population of either people who have the condition under study or the general population. Some differences between groups may not be due to different underlying interest in mHealth itself, but instead through recruitment and study design. As an example, several studies suggest men are more likely than women to report interest in mHealth[201,202] and gender imbalance is often present in mHealth studies. However, the direction of the imbalance is inconsistent, which may suggest recruitment strategies or interest in the particular medical issue are more important than an underlying gender-based difference caused by an interest in mHealth in general.

Disease or health outcomes, whether the primary outcome of the study or comorbidities, could directly affect a participant's willingness to adhere to the study. A person with a direct interest in the condition under study may have greater motivation to participate. For example, a person undergoing a depressive episode may be less adherent.[203] Depression is a common comorbidity in many chronic diseases[204] and could bias or reduce the quantity of data in a wide range of studies. Whilst it has been suggested that missingness in mHealth data can be informative[203] it is still a source of uncertainty, particularly within studies that aim to understand or investigate a medical issue rather than in a clinical or intervention-based mHealth paradigm where detection itself is paramount. The impact of mental health on attrition is something that we can investigate here because of the regular mental health questionnaires collected throughout the Covid Collab study.

Particularly where there is little direct interaction between researchers and study participants, self-motivation to adhere to a study's protocol is an incredibly important factor for continued participation and the production of useful data. However, motivation may well

have interactions with attributes of the participant which are impactful to how the study should be designed or how results should be interpreted.

Citizen science studies

Mobile health technologies can be used in a variety of research with differing enrolment strategies and degree of direct contact with participants, from a supportive or marginal role where enrolment and retention may be managed as in a traditional clinical or research trial, to a hands-off *citizen science* type study in which prospective participants are directed to download an application which takes them through an automated enrolment procedure. The Mass Science app was designed for the later model, and we would expect participant engagement to previous citizen science mHealth studies, with the caveat that interest in participation could wax and wane with attention on the COVID-19 pandemic as opposed to the chronic conditions previously studied which may, tentatively, have a more stable level of baseline interest. In this section we will briefly consider the existing knowledge of engagement behaviours in a number of completed studies. Some of these created their study app using the ResearchKit framework,[205] an iOS specific framework developed by Apple to create app-based surveys, consent flows, and active tasks for research studies.

Several citizen science mHealth studies published the socio-demographic breakdown of their participants or even detailed engagement patterns prior to the pandemic. Several more similar studies, targeted specifically at monitoring COVID-19, were launched near the beginning of the pandemic, alongside Covid Collab. Although engagement and attrition are not exactly the same between all of these citizen science studies, there are some shared patterns.

The most noticeable is a severe drop in participation at the very beginning of each study, followed by continued attrition throughout the study. In the MyHeart Counts study mean engagement duration was 4.1 days[206] and in the Cloudy with a chance of pain study 33% of the participants disengaged from the study immediately after enrolment (total N=13207, 2623 without baseline questionnaires and a further 1733 with the baseline questionnaire but no further engagement). The asthma mobile health study recruited 7593 participants, with 30.5% (n=2317) completing at least 5 daily or weekly surveys and 2.3% (n=175) with a 6-month milestone survey complete.[207]

As in the general mHealth case, there are often reported differences in engagement and retention in citizen science project along socio-demographic lines. The distribution of age at enrolment is often either proportionately younger[115,207,208] or thin-tailed such that the youngest and oldest age groups are less represented.[209] However, where it is noted, older participants seem to have a higher retention rate.[207]

Gender is often skewed; some studies have a predominantly female cohort[14,210,211] while others are predominantly male.[115,207,212] This may be partly, but not entirely, explained by the condition under study effecting one gender to a greater extent; for example men accounting for 82% of the MyHeart Counts study on cardiovascular health. The studies based on COVID-19 research had a greater proportion of female participants[210,211,213]

The severity or presence of the disease under study is also occasionally mentioned as a factor in engagement. People with worse asthma control were more likely to enrol and those reporting more frequent symptoms were more likely to continue in the asthma mobile health study.[207]

The Cloudy with a chance of pain study reported an in-depth analysis of engagement.[209] In it, they devised a model for engagement based on a 3-state hidden Markov model. The three states corresponded to high engagement, low engagement, and disengaged. This model was used to cluster the study's participants into four clusters, *high engagement*, *moderate engagement*, *low engagement*, and *tourists*, where tourists are those who disengaged from the study almost immediately.

Increasing participation

Given the proportion of people who disengage from studies after a short period there is, understandably, a large focus on how studies and their supporting software can be designed in a way which retains General recommendations to increase participation follow a few major themes.[214]

App design Some focus on app design; aesthetic features include a clean and consistent user interface (UI), colour scheme, and a well functioning bug-free experience.[215] Aspects of app design can be more or less attractive or accessible to certain groups. For instance, legibility or complexity concerns in the older adult population.[196]

Study burden The instruments and tasks required to be completed require time and place a certain level of burden on the participant. Within online studies, the relevance and length of questionnaires is a factor in attrition,[216] as is the length of the baseline assessments.[217]

Value to participants Providing value to the participant through features such as feedback and access to their results through graphs, history, or other views, gamification of app elements,[218] and monetary compensation[219,220] can increase engagement in the study. Direct access to medical professionals is available in some mobile health apps and likely provides value to the participant, but more oriented to intervention style apps

with manageable cohort sizes. To some extent the value that participants perceive in the study may be balanced against the perceived burden of taking part.

Communication Communication can cover several areas. It is important that participants fully comprehend what is required by the study and how to complete any assigned tasks so that they are able to make an informed decision to take part and so that the data they generate is useful. Communication is naturally limited in online or citizen science style studies, but comprehension can be partially achieved through well designed instructions. Notifications and reminders are an important method for driving engagement[221] but the desired frequency and under what circumstances they should be delivered is an open topic, for instance large numbers of notifications can elicit a negative affective response[222]

Recruitment practices It is important to try and create a representative study population so that any results are not tainted by bias and are more likely to be generalisable. Targeted recruitment of specific demographics can help ameliorate underrepresented groups in the study population, but requires either prior knowledge of enrolment and attrition patterns across different groups, which may differ between studies, or rapidly alter recruitment targets based on incoming study data.

The focus of this chapter is to describe the development process and final state of the mobile application and associated backend infrastructure made in response to the COVID-19 pandemic. Secondarily, it is to provide insights into participant engagement and retention in Covid Collab, a remotely enrolled crowd-sourced study for which the mobile app was developed. The methods and development section is split fairly evenly between the two aims, while the results are primarily focused around participant engagement.

5.2 Methods and Development

The first part of this methods section will deal with the structure and development of the Mass Science application and backend infrastructure. The second part will present the analysis of participant engagement in the *Covid Collab* study.

5.2.1 App Structure and Functionality

The application is built up of several components: a user interface (UI) which the participant interacts with; modules around active tasks and surveys, including the models for each task, scheduling, and notifications; passive data collection; and flows that allow the connection of third party accounts and services.

User Interface

The three major screens once a user is logged in are the Home, History, and Sources pages, shown in Figure 5.1. They provide the user interface to view available tasks, visualise previously submitted data, and connecting third party wearable devices respectively. In addition, there are screens for login, enrolment, onboarding, displaying an in-progress task, and leaving the study.

Home The app is primarily a data collection platform. Therefore, the home page consists of a widget at the top of the page to enter ad hoc COVID-19 related information and a widget underneath display active tasks that are available for the participant to complete (Figure 5.1a).

History dashboard The history dashboard displays a plot of previous self-rated happiness and energy responses from the symptoms task (Figure 5.1b).

Sources The sources page allows the connection of and toggle the collection of passive data. Currently, Fitbit and Garmin accounts can be connected. Previously it was possible to enable and disable the collection of location data on this page. The rationale was to allow precise control by the participant over exactly what and when data is shared.

Enrolment The enrolment and onboarding screens aim provide information so that a person can make an informed decision on whether to take part, explain the study and how to navigate the app, and to ensure ethical obligations (such as providing a participant information sheet) are carried out.

Login A screen to allow a person to go into the enrolment process or log in to an existing account.

Questionnaires and Active Tasks

The Mass Science app is designed to facilitate the collection of passive and active mobile health data. By its nature, the passive data is collected opportunistically and with little or no

input required from the participant. The vast majority of the functionality and experience of the app that a participant will interact with are the active tasks that they are requested to complete. The following section will describe the structure of an active task within the app. A more detailed description of Covid Collab tasks is given in the following chapter, and a full list of tables of implemented tasks is available in Appendix B.

Within the app code base, an active task consist of a *Procedure* and an attached *Schedule*. A procedure is a collection of *Tasks* which may be conditionally dependent on the results of other tasks within the same procedure, previous responses of the same survey, or values from tasks in other procedures. Widgets are created for each task type and displayed in a sequence to the participant within the particular active task procedure. An example of an individual *Task* would be a single item of a questionnaire, while the questionnaire in its entirety is the *Protocol*. The following generic task types are implemented:

Dropdown Allows the selection of a single item from a predefined list of options in a dropdown box.

Radio / Tickbox Allows the selection of a single item (Radio) or multiple items (Tickbox) from a list of displayed options.

Datepicker Allows the entry of date, time, or both through either text entry boxes or a calendar and clock widget.

Listbuilder A Listbuilder task allows the construction of multiple values of another type. For example, the COVID-19 symptom entry task is a listbuilder task in which each item is a symptom name attached to a radio task of severity on a 4-item Likert scale. Symptom severity scores can be added and removed from the response by the participant.

Slider A continuous or segmented numeric scale. Optionally it can be labelled rather than directly displaying a numeric values.

Text A text entry field that allows the participant to enter arbitrary text under certain constraints (e.g. length, numeric or date comparisons, or pattern matching). It can also be combined with other types, such as Dropdown or Tickbox, to allow for custom entry if preexisting options are not sufficient.

Webview (Cognitron) A webview displays an external online web Uniform Resource Locator (URL). It can inject JavaScript to intercept data of interest. It is currently solely used to display Cognitron[223] tests for Convalescence participants.

A procedure can be attached to one or more *Schedule*. As the name suggests, a schedule is responsible for determining when a procedure is available to complete by a participant. In addition, schedules can be conditional on the presence of or values contained in other active tasks. For example, in the Covid Collab study some questionnaires are only scheduled once a positive diagnosis has been reported in the COVID diagnosis protocol.

Regular For procedures that are to be completed on a fixed frequency such as every two weeks.

Cooldown For procedures that become available after a fixed cooldown period starting from the point that the last response was submitted.

Oneshot A one-time schedule

AlwaysDue For procedures that should always be available to complete.

As mentioned, whether to display a task or schedule a procedure can be conditional on other values. A condition takes an identifier of the *task id* (and optionally the *procedure id* that the task belongs to, it is otherwise assumed to come from the same procedure) and a comparison to run on that value. The comparison operators currently available are:

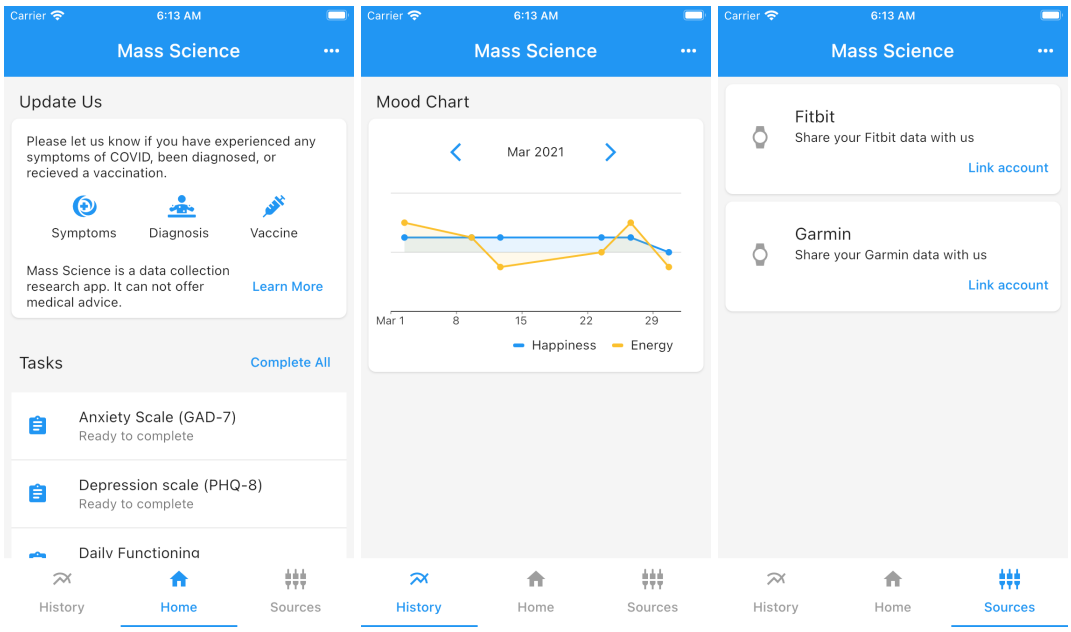
empty, notempty Whether the task value was completed or not.

gt, lt, eq, gte, lte Whether the numeric or date task value is greater than, less than, or equal to a comparison value.

contains, containsAll, containsAny Whether a list contains a (or multiple) comparison value.

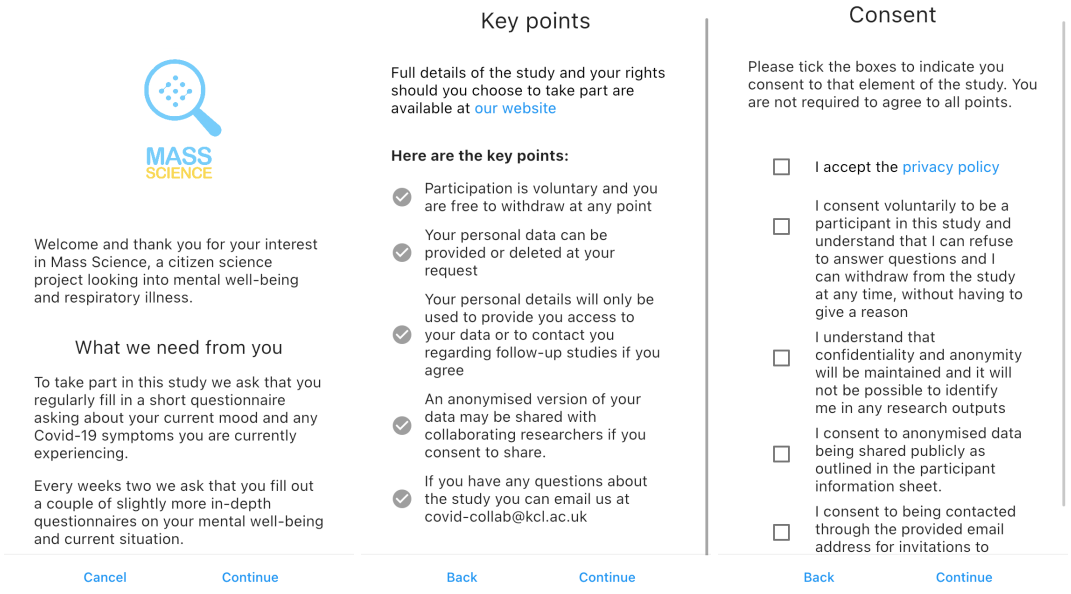
Passive Data Collection

There is a small amount of passive data collection directly from the app. Baseline information on the phone and operating system are taken. Originally, high-frequency raw location data was collected. The collection of location data was discontinued after several months due to a change in Google Play Store policy. Most passive data is collected from third-party wearable devices through the device manufacturer's web application programming interface (API)s. Participants with Fitbit[224] or Garmin[225] devices can connect their account through the Sources page of the Mass Science app.



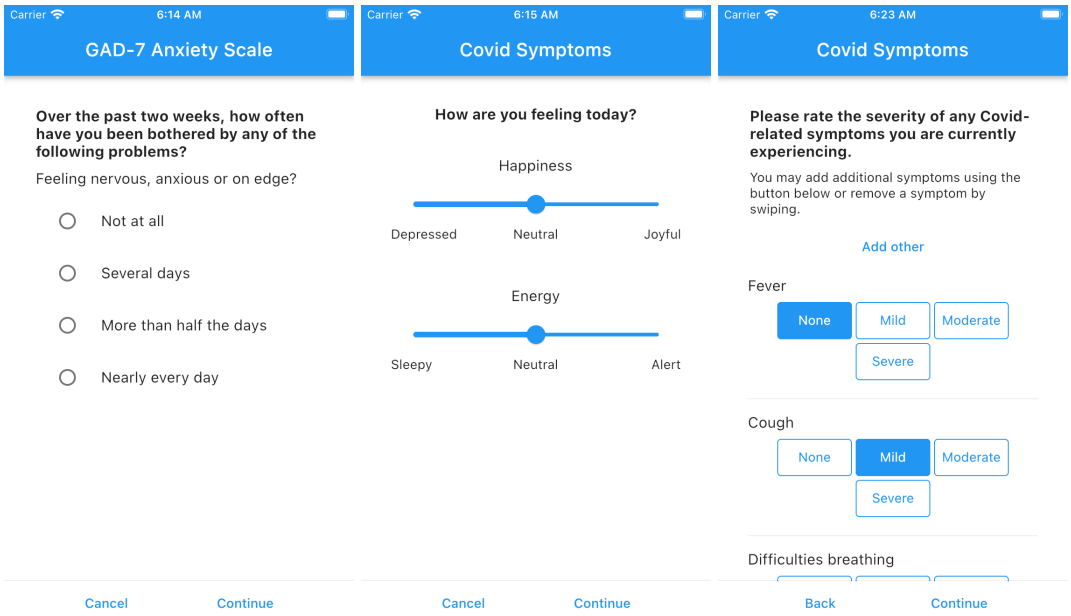
(a) Home page (b) History page (c) Sources page

Fig. 5.1 The three main pages of the Mass Science application



(a) Enrolment information (b) Enrolment key points (c) Enrolment consent page

Fig. 5.2 Part of the Covid Collab enrolment process in the Mass Science application



(a) GAD-7 questionnaire (b) Arousal-valence scale (c) Symptom severity rating

Fig. 5.3 Questionnaire examples in the Mass Science application

5.2.2 Backend Structure

The background infrastructure supporting the app is largely based around leveraging Google Cloud Platform components[226] alongside the RADAR-base RADAR-REST-Connector.[227]

Google Cloud Platform and Firebase

Firebase is a framework within the Google Cloud Platform focused on enabling web and app development. Several of the components are used within Mass Science. The most important distinction between the requirements for Mass Science and what was implemented in RADAR-base at the time was the ability to remotely enrol and authenticate participants. Firebase Authentication is an authentication backend which provided a few useful features: the ability for a user to sign up and sign in with an email and password, in-built account recovery options, and easy authorisation of users to specific parts of other Firebase services.

Receipt and storage of participant data is handled through Cloud Firestore, a NoSQL document cloud database. Access is restricted so that participants can only read and write their own data. Because it is a standard database that can be accessed from within the app, despite being persisted online, providing data and visualisations directly back to the participant is made fairly straightforward.

Cloud Functions is a framework for running backend code written in one of several languages in response to an event, such as an HTTP request or a change to the Firestore database. It is used for most of the remaining backend tasks: scheduling and sending notifications, supporting participants leaving the study, linking and authenticating third party sources, and data processing.

Notifications

Notifications to complete active tasks are sent remotely rather than from within the app itself. A server side implementation of the schedules found in the app are used to schedule reminder notifications to be sent to the participant. To schedule a notification a document is created within Firebase with details of the time to send, message, and associated active tasks. When the notification is within one month of the time it is to be sent, a job is created on Cloud Scheduler. The job ultimately runs at the time set in the document and the job triggers the *send notification* cloud function with the notification document ID as an argument. The notification document is then updated with the result of the notification.

RADAR-base components

An in-depth description of the RADAR-base platform, including the REST connector, is available in the RADAR-base paper.[8] Briefly, the REST connector takes the Garmin and Fitbit accounts linked through the app and cloud function and uses the companies respective REST API to request the participant's data.

5.2.3 Enrolment and Attrition

Recruitment Strategy

Recruitment began in June 2020. The study was first publicised through newsletters, news articles, and the university mailing list. Between August 2020 and May 2021 the study was linked to from within a section of the Fitbit app[228] for users based within the United Kingdom (UK).

Length of engagement

The total duration spent in the study for a participant is defined as the time between enrolment and the date of completion of the last active task. There are several active tasks available to complete with schedules of differing frequencies, the details of which are given in the next chapter. Engagement may also differ between different active tasks because of

the frequency of notifications, the length of the questionnaire, or the participant's impression of the importance of the task. However, when considering attrition, all the active tasks are included. Passive data belonging to a participant may still be available through a previously connected third-party wearable device manufacturer's API after the participant has stopped otherwise engaging with the study. While that can still provide useful information, it is not included here for determining attrition. Up to one year of data is included for each participant from the point that they enrol.

Factors predicting attrition

To understand how baseline characteristics of a participant might affect the rate of attrition we use survival analysis techniques, which are commonly used where time-to-event data with censoring is present.[229] Censoring refers to the case where the exact time-to-event is unknown because data becomes unavailable before that point (in the case of right censoring), for example, because the study ends. The event in our case is the point that the participant stops engaging with the study. A proportion of participants are still engaged with the study and therefore the event, the point at which they will drop out, is right censored. The analysis is taken up to September 2022. If a participant has submitted at least one active task in the month prior (August 2022) they are assumed to still be engaged.

Survival probability, here the probability that a participant is still engaged in the study, can be calculated at time t with both censored and uncensored data using the Kaplan-Meier (KM) method.[230] To calculate the probability of survival $S(t_i)$ at time t_i the probability of survival from the previous time point is multiplied by the proportion of participants who survived the current time step. The proportion who did not survive is the number who had an event e_i at time t_i divided by the number of surviving participants n_i directly prior to t_i. The full equation is given in Eq. 5.1. KM plots are generated separately for sex and age categories.

$$S(t_i) = S(t_{i-1}) \left(1 - \frac{e_i}{n_i}\right) \tag{5.1}$$

We can visualise the probability of survival in different groups and estimate the hazard ratio of two groups straightforwardly, but to compare multiple groups or variables we need a more comprehensive model. A Cox proportional hazards model is a parametric regression model.[231] It is similar to a linear regression but where the outcome variable is the hazard function at time t. The outcome is assumed to be equal to the baseline hazard function multiplied by the exponential of a function of the covariates (Eq. 5.2). The hazard function is the instantaneous rate of an event at a particular time.[232]

$$\lambda_i(t|x_i) = \lambda_0(t)exp(\beta^T x_i) \tag{5.2}$$

A Cox model is fit to sociodemographic, mental wellbeing related, and historic wearable data predictor variables. The sociodemographic variables are age, sex, employment status, the presence of one or more mental comorbidity, and the presence of one of more physical comorbidity, smoking status, and body mass index (BMI). Additionally, depression and anxiety are the two most commonly reported mental health related comorbidity in the sociodemographic survey and are included. Where it is available, historic Fitbit data is taken from the year 2019 to generate three predictors: historic sleep, activity, and heart rate. Historic activity is the average time spent per day in the 'very active' Fitbit activity category. Historic sleep is the average duration of each discrete sleep log. Historic heart rate is the mean of the daily resting heart rate.

The predictors are from different sources with different data availability. There is also a certain amount of covariance between some of the predictor variables. Therefore, several models are fit to different groups of predictors to maximise the number of observations and to show the effect of a group of predictors in isolation. All groups include age and sex. In addition, the groups contain the following:

1. Employment status
2. Historic sleep, heart rate, and activity
3. The presence of physical and mental comorbidities
4. Depression and anxiety
5. Smoking status
6. BMI

Temporal behaviour pattern clusters

The total duration of engagement is not the only factor that is important when understanding how different people interact with a study. There can be periods of high, low, and no engagement, and different aspects of the study may be engaged with differently. To try and capture some of temporal similarities between participants I aimed to cluster the sequences of engagement for each participant into groups. To cluster, it is necessary to know the distance between the individual objects. Clearly there is not a straightforward way to calculate distance between two temporal sequences that can take into account how it varies over time

and where similar patterns among different sequences may not occur at the same time point. One method, used here, is to fit a HMM to each sequence and use the likelihood of observing other sequences under that model as the distance.

A HMM is a generative probabilistic model and are often used in time series analysis. They provide a way of matching an observed sequence to a hidden state, where the underlying state may have some meaning. A model consists of N hidden states, each with a probability distribution b_i that emits elements of an observed sequence where i is the particular state. There is a distribution A for probability of transitioning between states and a distribution π for the probability of starting in a particular state. Originally the emission or observation sequence was typically discrete and so there was an additional parameter M referring to the length of the observation alphabet.[233] However, the emission probability distribution can be a distribution over a continuous variable as well. The three main questions for a HMM are: what is the most likely sequence of hidden states for an observed sequence of emissions given a model, what is the likelihood of an observed sequence given a model, and what are the most likely parameters for a model given an observed sequence.

The process of generating a distance matrix between each participant's sequence of engagement is as follows:

1. Create an engagement sequence for each participant. The sequence is the number of surveys submitted by the participant in a week for each week after enrolment for one year, creating a 52-length vector.

2. Fit a HMM to each participant's engagement observation sequence. Each model has 3 hidden states and a Poisson emission distribution. Models are fit using the hmmlearn Python library.[234]

3. For each HMM, calculate the log likelihood of every other participant's sequence under that model.

4. Use the above log likelihoods to create a distance matrix D of the absolute log likelihood. The distance matrix is made symmetrical by taking the minimum of the two absolute log likelihood values $P(X_i|\theta_j)$ and $P(X_j|\theta_i)$ where X_i is the observed sequence for the ith participant and θ_i are the parameters of the HMM for the ith participant (See Eq. 5.3).

$$D_{ij} = min(P(X_i|\theta_j), P(X_j|\theta_i)) \tag{5.3}$$

5. Generate an affinity matrix A from the distance matrix using Eq. 5.4

$$A = exp\left(\frac{-D}{\sigma(D)}\right) \tag{5.4}$$

In a general sense, it may be expected that a participant who disengages very shortly after enrolment would have a model in which there is a state with a low probability of emitting a value greater than 0 and with a high probability of staying in that state. A participant with high engagement would have a state with a high probability of emitting an observation greater than 0. Meanwhile, a participant who submits surveys infrequently but remains in the study may have a state which is likely to emit a value greater than 0, a state which is likely to emit 0, and a higher probability of moving between states than the previous two theoretical participants. An observation sequence from a participant would be expected to have a higher likelihood under another participant's model the more similar their engagement behaviour is.

Given that we have the distance (and affinity) matrix, there are a number of clustering algorithms available to us. It is likely that the clusters will have different variances and may potentially be overlapping. To illustrate why the variance between clusters may not be expected to be equal, consider two groups: a low and moderately engaged. A group of participants who have very low engagement, such as dropping out after the first day, will probably have models that have a very high likelihood across all similar participants. This is because after the initial emitting state, the model can transition to a state with a very high probability of emitting a 0 and then stay in that state. A moderately or highly engaged group, however, may require more state changes and have emission distributions spread over a greater range of values. These constraints on the way the data is expected to be clustered led to the decision to use a spectral clustering algorithm[235] implemented in the Python library scikit-learn.[142]

Spectral clustering requires specifying the number of expected clusters. Since we do not have a prior notion of how many behavioural engagement clusters there are in our dataset, we follow the eigengap heuristic method to determine the optimal number of clusters based on the eigenvalues of the affinity matrix.[236] Firstly, the graph Laplacian is calculated from the normalised affinity matrix using the *csgraph.laplacian* function in the scipy library.[237] Secondly, the optimal cluster number is estimated based on the largest distance between eigenvalues.

The resulting clusters are described qualitatively on the basis of the overall group periodicity, engagement duration, and consistency. Several basic statistics over sociodemographics and engagement are given for each cluster. Finally, a multinomial regression is fit with cluster group as the dependent variable and age, sex, and depression as a comorbidity as the

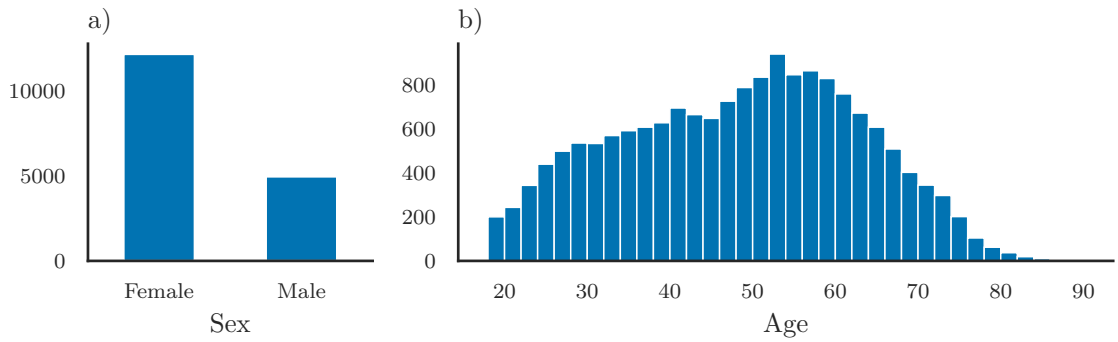


Fig. 5.4 Participant age and sex distribution

a) The number of female and male participants enrolled in the study. b) The age distribution of participants in the study.

independent variables. The intention is to see whether a common coexisting mental health concern is associated with any specific engagement pattern.

5.3 Results

5.3.1 Mass Science Application

The Mass Science application was launched on the Apple and Google app stores in May 2020. Since that date it has been downloaded over 20,000 times and has had over 17,500 enrolled participants. It has been iterated on several times. The most substantial update was the introduction of the ability to run multiple studies and the associated changes required for the Convalescence study.

5.3.2 Descriptive Analysis

Even at the point of enrolment there is a large imbalance across sociodemographic groups in the Covid Collab study. Female participants (N=12137) far outnumber male participant (N=4950). There is also an under-representation of the age groups at the extremes of those allowed to participate, 20- to 30-year-olds and those above 65, while there is an over-representation of 50- to 60-year-olds, which can be seen in Figure 5.4. Participants who completed the extended sociodemographic survey question on ethnicity, were also majority white British (90.2%, N=5335/5916). Additionally, the majority of participants with location information and the majority of app downloads (roughly 90% in both cases) were from the United Kingdom (UK).

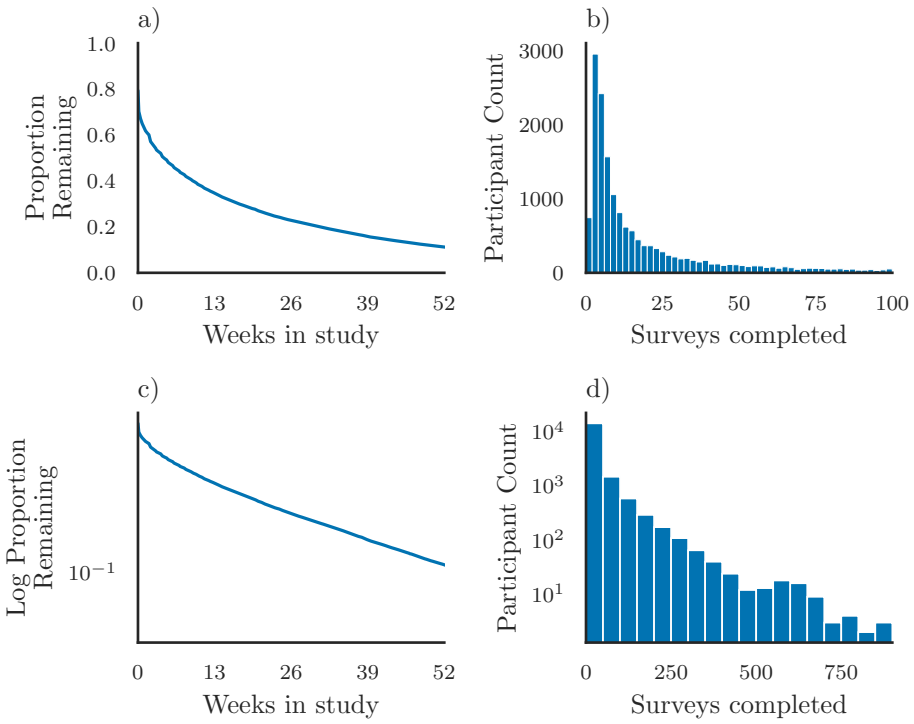


Fig. 5.5 Participant contribution and attrition

a) Proportion of participants remaining in the study at a certain point after enrolment. Contribution is defined as whether the participant has contributed an active task either on the week or at a later date. b) The number of active tasks completed per participant. The plot is truncated at 100 surveys completed because of the small proportion above this point, but the maximum number contributed by a participant is 889. c) Shows the proportion of participants remaining on a log scale. d) Shows the number of completed tasks per participant on a log scale.

Within the study population there is clearly unequal levels of contribution. Figure 5.5 shows a large drop in participation in the first few weeks of study, followed by a gradual levelling off. After the first couple of weeks during which there is increased dropout, the rate of attrition roughly follows a power-law distribution, demonstrated by the straight line of the log plot. There is a similar pattern in the number of surveys completed, with a small number of participant responsible for a large number of survey responses.

5.3.3 Survival analysis

To start to understand whether those unequal levels of engagement are associated with different sociodemographic categories,we can consider how survival (here continued engagement in the study) differs between different groups. Figure 5.6 shows how engagement differs

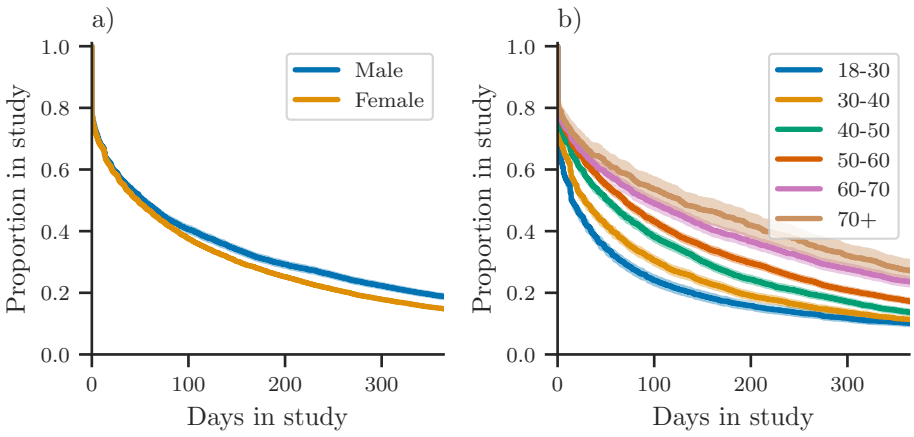


Fig. 5.6 Attrition by age and sex

Proportion of participants remaining in the study over time across (a) sex and (b) age categories. Shaded bands represent 95% confidence intervals.

between men and women and between different age groups. Ostensibly, male participants appear more likely to stay in the study despite forming a smaller proportion of the study. There are also a clear increasing level of participation the older the age group, such that at half a year around 50% of the 70+ age group are still engaged, compared to around 20% of the 18-30 age group. The proportion of remaining participants for each group does level off at around 15% by the end of the year for the younger groups, while the older groups are still more engaged but continuing to decrease.

5.3.4 Proportional hazards regression

The proportional hazards model allows us to see the affect of multiple groups and continuous variables on the hazard ratio. Figure 5.7 displays a visualisation of hazard ratios across the different groups of predictor variables run in each proportional hazard model. A table of the numerical results is available in the appendix Table A.1.

While on the basis of the first survival plots male sex seemed to be associated with reduced attrition, when taking age into account there is no longer a significant affect, and the spurious relationship is likely due to a higher average age for male participants. Age continues to be significantly related to the rate of attrition.

The following groups of predictors are all part of models that include age and sex, but not the other variables. Employed participants were not significantly different to retired, unemployed, or student participants.

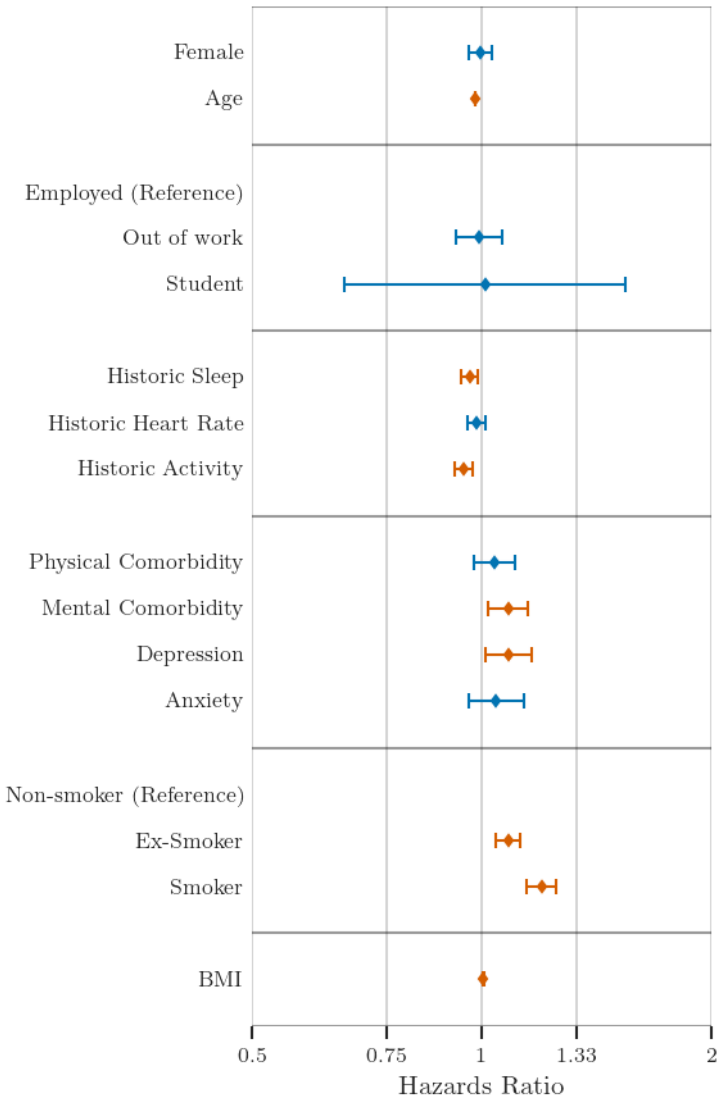


Fig. 5.7 Proportional hazards model hazard ratios for attrition in the Covid Collab study

An orange marker signifies a p-value under 0.05 while a blue marker signifies a non-significant result.

Of the historic wearable metrics, both an increased mean duration of sleep and an increased mean duration of high activity per day caused a lower hazard ratio, 0.964 and 0.947 respectively. The wearable metrics were normalised by taking the z-score across all participants and so the hazard ratio is in reference to a change in the cohort-level standard deviation.

The presence of at least one physical comorbidity was not significantly associated with attrition, but the presence of at least one mental comorbidity had an increased hazards ratio (1.08). The two most common mental comorbidities in the cohort were selected. A reported depression diagnosis had a similar hazards ratio (1.08) while anxiety was not significantly associated.

Smokers (HR=1.20) and ex-smoker (HR=1.08) were both significantly more likely to leave the study than their non-smoking counterparts. Finally, there was a small but significant increase in attrition as BMI increased (HR=1.004).

5.3.5 Engagement Clusters

The clustering of participants into engagement groups on the basis of the log likelihood of every other participant's engagement sequence under the parameters of a HMM fit to a particular participant was carried out on all participants with over one weeks worth of data. Participants who immediately disengage from the study clearly form a single particular pattern, and given the fairly large number of those participants (N=5034) there was a potential to affect the clustering of other groups without providing any useful information. The clustering algorithm was therefore run on the 11299 remaining participants.

The five largest differences between eigenvalues of the affinity matrix, and therefore the five most likely to be optimum numbers of cluster were 1, 7, 5, 3, 8. The spectral clustering algorithm was therefore run with 7 groups. Figure 5.8 displays a heatmap showing the engagement sequence for every participant ordered by their cluster. Table 5.1 provides descriptive statistics of number of participants, average number of surveys completed, average duration of study engagement, age, sex, and proportion of participants reporting depression for each cluster.

Reducing groups of engagement behaviour into neat descriptions or values is hard, but there do appear to be some qualitative and quantitative differences. Across clusters, and in line with the survival analysis, more engaged groups tended to be older. The least engaged clusters had the highest rates of depression. Roughly, clusters 1, 2, and 4 are highly engaged, while cluster 3, 5, 6, and 7 are moderately engaged. In the following list, clusters are rated on consistency, whether or not participants miss weeks while still in the study; contribution, the number of surveys submitted; and duration, the total number of days in the study.

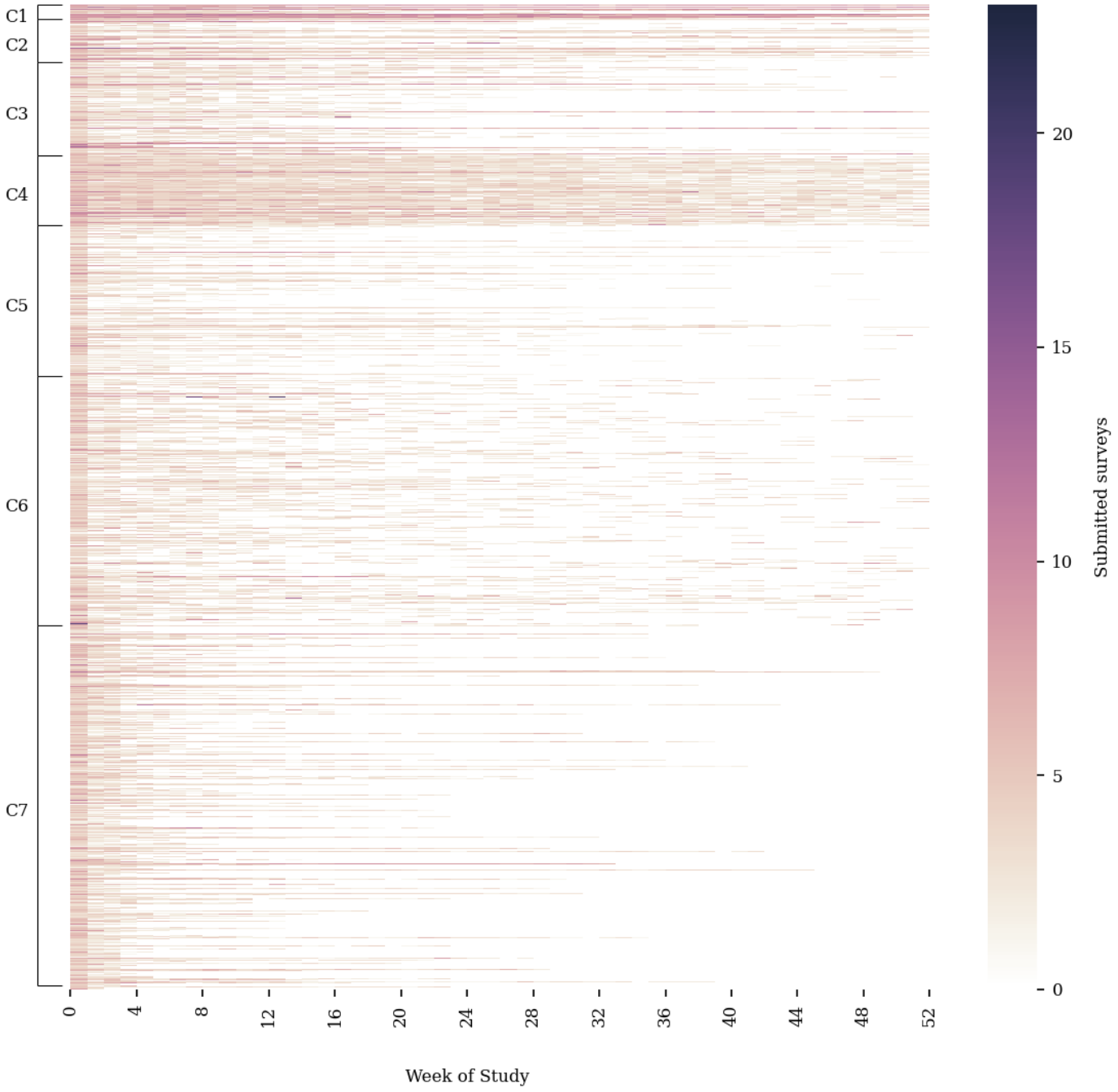


Fig. 5.8 Participant engagement clustered into 7 groups

Cluster 1 *High consistence, extreme contribution, long duration*
The smallest and most engaged group. The average amount of time spent in the study is only the second highest, but as a group they provided almost twice the number of surveys as the next closest group. The proportion of participants with depression (12.9%) is noticeably lower than any other group. Although still predominantly female, it is the group with the highest proportion of male participants (36.7%)

Cluster 2 *Moderate consistency, high contribution, moderate to long duration*
This group appears to be less consistent than the other two high engagement clusters. There is a large variance in the number of days in study and while the mean number of days is lower than clusters 1 or 4, the number of surveys submitted is equivalent to cluster 4 participants. The main distinguishing characteristic appears to be a highly engaged first third of the study followed by a moderate amount of engagement and dropout.

Cluster 3 *Moderate consistency, moderate contribution, moderate duration*
Similar to cluster 2 in that the number of surveys is fairly high in relation to the to amount of time spent in the study. Despite having an intermediate average number of days in study (d=171) compared to cluster 5 (d=143) and 6 (d=243), there are roughly twice as many surveys submitted. Contribution is high in the first few months, but attrition is substantial afterwards.

Cluster 4 *Extreme consistency, high contribution, long duration*
Members of this group appear to stay engaged, not drop out, and consistently submit surveys every week.

Cluster 5 *Low to moderate consistency, low to moderate contribution, moderate duration*
In general, participants in cluster 5 appear to start with a fairly consistent contribution, become inconsistent, and finally drop out after, on average, 4 or 5 months. They are similar to participants in cluster 3, but with lower rates of contribution.

Cluster 6 *Low consistency, low to moderate contribution, long duration*
This group is seems characterised by a pattern of irregular engagement. Participant are often retained in the study for a long period but are likely to skip one or more weeks at a time.

Cluster 7 *High consistency, low to moderate contribution, short duration*
Most participants in this group consistently provide data for a few months and then drop out. Unlike cluster 5 or 6 participants, once they stop they do not reengage.

Cluster	N	No. Surveys	Days in study	Age	Sex=Female	Depression
1	169	265.5±167.8	420.0±203.4	58.2±12.5	63.3%	12.9%
2	490	110.5±105.1	295.9±234.6	54.4±13.2	67.3%	23.3%
3	1077	62.1±74.6	171.3±147.4	51.2±13.4	73.3%	21.1%
4	793	153.5±44.4	516.3±118.6	56.4±12.2	66.5%	18.9%
5	1730	27.5±33.7	143.2±139.1	47.7±14.2	70.0%	24.6%
6	2867	38.6±30.5	243.6±152.6	50.4±13.5	74.2%	27.8%
7	4173	29.0±36.4	62.3±85.6	46.5±14.2	71.7%	28.5%

Table 5.1 Descriptive statistics for engagement clusters

In addition to the description above based on descriptive statistics and a visual inspection of the engagement sequences belonging to each cluster, a multinomial logistic regression of age, sex, and depression to cluster group was carried out. The results, given in Table 5.2, take cluster group 7 as the reference group. While the majority of the results reinforce what was demonstrated in the survival analysis, there is a significant difference in the proportion of female participants between clusters 6 and 7. This may suggest that groups or variables that are not significantly different when considering total engagement time may be different when considering the pattern of engagement over time.

5.4 Discussion

This chapter focused on both the design and software development of the Mass Science application and the patterns of adherence among participants of Covid Collab, a citizen science study run through the Mass Science app. While on the surface the two aims do not appear similar, design choices in an app and study can have consequences on adherence and many of the recommendations in the literature to increase engagement are based around or rely on app design and development. It was therefore useful to first present the Mass Science app.

5.4.1 App design

The design of Mass Science largely followed the design of the RADAR-base active RMT app, which itself went through several design cycles with different focus groups.[238] Additionally, feedback from a session with the Young People's Mental Health Advisory Group on the design of a citizen science app for monitoring depression before the pandemic was

Predictor	**Coeff**	**Std Err**	**P-value**
Cluster 6			
age	0.022	0.003	7.54e-14
female	0.206	0.088	0.020
depression	0.029	0.089	0.745
intercept	-1.09	0.169	0
Cluster 5			
age	0.0033	0.004	0.392
female	0.0395	0.117	0.735
depression	-0.195	0.123	0.114
intercept	-1.01	0.217	0
Cluster 4			
age	0.0534	0.004	4.18e-46
female	0.0676	0.104	0.514
depression	-0.304	0.117	9.43e-3
intercept	-3.22	0.227	0
Cluster 3			
age	0.0273	0.004	8.90e-10
female	0.0164	0.129	0.899
depression	-0.284	0.145	0.049
intercept	-2.47	0.262	0
Cluster 2			
age	0.053	0.005	4.90e-23
female	-0.0332	0.144	0.818
depression	-0.0271	0.16	0.866
intercept	-4.18	0.331	0
Cluster 1			
age	0.0626	0.007	8.01e-18
female	-0.0358	0.186	0.848
depression	-0.694	0.259	7.30e-3
intercept	-5.25	0.46	0

Table 5.2 Multinomial logistic regression of HMM-clustered engagement groups

The regression coefficients are in comparison to Cluster 7

considered. The key differences between RADAR-aRMT and Mass Science were the inclusion of a remote enrolment flow, remote connection of third party wearables, visualisation of participant data, and originally the inclusion of passive data collection from within the same app. Additionally, there was a design decision to allow participants to have fine-grained control over what data they donated. It is therefore possible to individually toggle the collection of Fitbit, Garmin, and previously location data.

Data feedback to participants, through visualisations for example, is a common suggestion in focus groups. It was implemented in a limited fashion in the history screen, providing a view of previously reported happiness and energy scales. Unfortunately there was a balance between implementing features within the app and the need to release the study as early in the pandemic as possible. The purpose of the app is to be the patient-facing interface of a data collection platform, and so features supporting that purpose were the focus of implementation efforts. The development of a research app that, in addition to data collection, directly provides a useful service to participants may help to increase adherence.

5.4.2 Participant engagement

The general pattern of engagement in the Covid Collab study was similar to those in other mHealth citizen science studies. There is a high level of attrition, particularly at the beginning of the study, but with a subset of participants with long term commitment. Imbalances in age and sex are common across studies, but not always in the same direction. Here, there is a predominance of middle-aged and female participants. The survival analysis conducted demonstrated the importance of including multiple predictors. The predictors significantly associated with attrition are age, historic sleep, historic activity, mental commodities, depression, BMI, and smoking status. These may often differ in study groups, particularly in a citizen science study where participants are not directly selected. Some may also be directly or indirectly related to the outcome measure of a study, and therefore it will be important to consider how engagement of participants might affect analysis or cause spurious associations.

The clustering approach used on engagement data in this study produced visually distinct and understandable groupings outside just the total time spent in the study. Clustering of engagement using a HMM was previously carried out in the Cloudy with a Chance of Pain study,[209] hereafter referred to as Cloudy. However, there are a few important differences in this study. Firstly, in Cloudy it appears that the states and associated probabilities of the transition and emission matrices were explicitly set to high, low, and disengaged engagement states. In this study a separate model, also consisting of three states, was fit to each participant's sequence in a non-supervised fashion. Secondly, the observation sequence in Cloudy

appears to be binary, whereas in this study the observation sequence is the number of surveys submitted in a week and therefore a Poisson emission distribution is used. Thirdly, the clustering in Cloudy appears to be achieved through fitting a mixture hidden Markov model whereas in this study the log likelihood of each sequence under each participant's model is used as a distance matrix in a spectral clustering algorithm. Finally, participants were split into four clusters in Cloudy and seven in this study. The increased number of clusters used in this study possibly led to a better visual split between different types of engagement behaviour, for example on the basis of how consistent a participant was, rather than 'high', 'moderate', 'low', and 'tourist' clusters described in Cloudy.

While it seems adequate, the clustering procedure used in this study could be improved. Clustering appeared visually less satisfactory when participants who immediately left the study were included (see Figure A.3 in the appendix). Short duration participants seem littered throughout several clusters without good separation. The architecture of the HMM was very basic, consisting of three states with no constraints on transition. It may not adequately capture all types of engagement behaviour. The observation sequence was also the sum of all surveys completed in a week and therefore the clustering did not consider how active tasks with different demands or frequencies might be engaged with differently.

5.4.3 Implications

The engagement patterns elucidated in this study have some implications on the recruitment practices of future citizen science projects. Under-represented groups may need specific targeting, which may be achieved through strategies such as choosing where the study is publicised or engaging with under-represented groups to understand why they may be less likely to take part. Targets for the proportion of certain groups within the study may also be informed by the expected rate of attrition of those groups, where you may require a higher number of participants from a group that is more likely to drop out of the study. There are further implications for the analysis of data from existing studies, like Covid Collab, that have dramatic differences in adherence and engagement. Active tasks are often used as, or form part of, an outcome variable. How that outcome variable is defined may cause it to become associated with a variable or group purely on the basis of how engagement is correlated with that variable or group. For example, in the paper that forms the basis of the following chapter the method for assigning participants to the long COVID group is to take participants who self-report at least one symptom every week for twelve weeks after a COVID-19 diagnosis. While it is a necessity to define long COVID on something, given that it is under-diagnosed and a true label often does not exist, the group will be biased towards consistent and engaged participants, and it is important to keep that caveat in mind. Similar issues may be present

in other studies. In a survey-based mHealth study long COVID groups were based on the presence of symptoms at certain time points post-diagnosis.[210] Engagement patterns in other studies may not be exactly the same as in Covid Collab, but there is potential for spurious associations to be found because of biases in engagement.

The point that engagement behaviour should be carefully considered is generalisable to other medical areas. Considering psychological mHealth studies, if someone is depressed or has another mental health issue they may be more likely to drop out or have missing data. Although not investigated in this study, it is possible that acute periods of depression, or relapse or remission in symptoms, are responsible for changes in engagement, rather than a baseline change caused by the presence of a comorbidity.

For many analysis problems in mHealth, the power-law-like distribution of data over participants may help motivate the use of models and training paradigms that can make use of the small amount of data that is often available for individual participants, such as the meta-learning techniques discussed earlier in the thesis. Many participants are not engaged for long, and this pattern may carry over to studies that aim to create detection or disease classification models. Any attempt to create personalised models would be limited to methods that can use the small amounts of data provided by many of those individual participants.

Use of passive data, rather than relying on active participant engagement, is a promising avenue. However, it is often still necessary to have self-reported labels to properly understand the passive data, and often passive data is still used as a predictor variable rather than being able to be used as an outcome directly.

5.4.4 Limitations and evaluation

There are a number of limitations to this study. The study cohort may be specific to the disease and condition under study, and the COVID-19 pandemic likely motivated many people to take part in a COVID-19 citizen science project that may not otherwise have engaged in or come across a similar mHealth study. Specific patterns found here, such as the increased proportion of women, may not hold across all studies. However, the general ideas and the techniques used here may be more widely useful.

We did not systematically collect reasons for participant dropout, which is itself logistically challenging in a remote study. Concrete steps to increase engagement, and whether they would affect different clusters of participants differently, are therefore hard to determine. Understanding how study or app design could be changed would require further work with differences in design across randomised groups, or A/B testing as an integral part of future studies.

5.4.5 Future work

The most immediate item of future work will be based around how acute periods of depression, anxiety, and symptoms, affect prior and post engagement. Self-rated scales for depression and anxiety were prompted for every two weeks in the Covid Collab study, and so it will be possible to investigate acute effects within the same dataset. Longer term, further study app development work could be focused around boosting and understanding engagement, rather than purely as a collection platform.

5.5 Conclusion

The Mass Science app was a mobile study app developed for the Covid Collab citizen science project. App design was considered in the context of study engagement and attrition. Within the Covid Collab study, engagement differed dramatically between participants and was significantly associated with sociodemographic factors and the presence of certain comorbidities. Clustering of participants demonstrated groups of different contribution and attrition patterns. Participant engagement will be important to consider in the analysis of mobile health, and particularly citizen science, studies.

Chapter 6

Covid Collab: Protocol Paper

6.1 Preamble

Following the description of the Mass Science app and supporting infrastructure in the previous chapter, the aim here is to give a more detailed look at the Covid Collab study itself, including the collected data, questionnaire protocols, and initial goals, to provide context for the following analysis chapter.

The idea of a remote citizen science mobile health project to collect data during the pandemic was not unique. Several studies specifically collected commercial fitness device wearable data. The largest was the Robert Koch Institute's Corona-Datenspende study[239] which recruited over 540,000 people[240] but initially only included wearable data. There were several studies within the USA: the DETECT study,[241] the Stanford COVID-19 wearables study,[242] CovIdentify,[243] and TemPredict.[244]

All of these citizen science studies looked at the combination of donated wearable data and self-reported COVID-19 symptom surveys. Because the studies were launched roughly concurrently without prior visibility, there was not any deliberate alignment in protocol. This caused minor differences in how self-reported symptoms, diagnosis events, vaccination status, and wearable data were collected. In contrast to Covid Collab, none of the studies collected regular mental health questionnaires. Wearable data was typically collected either only from the period that the participant was enrolled in the study or around specific illness events, except the TemPredict study, which collected twelve months of baseline data. Other citizen science initiatives included self-report-only studies such as the COVID Symptom Study, which attracted over four million participants.[245]

The chapter is included as a published paper. **DOI** 10.2196/32587

Protocol

Investigating the Use of Digital Health Technology to Monitor COVID-19 and Its Effects: Protocol for an Observational Study (Covid Collab Study)

Callum Stewart[1], BSc, MSc; Yatharth Ranjan[1], BSc, MSc; Pauline Conde[1], BSc; Zulqarnain Rashid[1], BSc, PhD; Heet Sankesara[1], BTech; Xi Bai[2], PhD; Richard J B Dobson[1,2,3,4,5], PhD; Amos A Folarin[1,2], PhD

[1]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]Institute of Health Informatics, University College London, London, United Kingdom

[3]Maudsley Biomedical Research Centre, South London and Maudsley NHS Trust and King's College London, London, United Kingdom

[4]Health Data Research UK London, University College London, London, United Kingdom

[5]NIHR Biomedical Research Centre, University College London Hospitals NHS Foundation Trust, London, United Kingdom

Corresponding Author:
Amos A Folarin, PhD
Department of Biostatistics and Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 20 7848 0924
Email: amos.folarin@kcl.ac.uk

Abstract

Background: The ubiquity of mobile phones and increasing use of wearable fitness trackers offer a wide-ranging window into people's health and well-being. There are clear advantages in using remote monitoring technologies to gain an insight into health, particularly under the shadow of the COVID-19 pandemic.

Objective: Covid Collab is a crowdsourced study that was set up to investigate the feasibility of identifying, monitoring, and understanding the stratification of SARS-CoV-2 infection and recovery through remote monitoring technologies. Additionally, we will assess the impacts of the COVID-19 pandemic and associated social measures on people's behavior, physical health, and mental well-being.

Methods: Participants will remotely enroll in the study through the Mass Science app to donate historic and prospective mobile phone data, fitness tracking wearable data, and regular COVID-19–related and mental health–related survey data. The data collection period will cover a continuous period (ie, both before and after any reported infections), so that comparisons to a participant's own baseline can be made. We plan to carry out analyses in several areas, which will cover symptomatology; risk factors; the machine learning–based classification of illness; and trajectories of recovery, mental well-being, and activity.

Results: As of June 2021, there are over 17,000 participants—largely from the United Kingdom—and enrollment is ongoing.

Conclusions: This paper introduces a crowdsourced study that will include remotely enrolled participants to record mobile health data throughout the COVID-19 pandemic. The data collected may help researchers investigate a variety of areas, including COVID-19 progression; mental well-being during the pandemic; and the adherence of remote, digitally enrolled participants.

International Registered Report Identifier (IRRID): DERR1-10.2196/32587

(JMIR Res Protoc 2021;10(12):e32587) doi: 10.2196/32587

KEYWORDS

mobile health; COVID-19; digital health; smartphone; wearable devices; mental health; wearable; data; crowdsourced; monitoring; surveillance; observational; feasibility; infectious disease; recovery; mobile phone

Introduction

Background

The COVID-19 pandemic has brought about widespread and drastic changes to people's lives, work, and health resulting from infection by SARS-CoV-2 as well as the public health and social measures (PHSMs) that were introduced to limit the disease. It is important to not only understand how and under what circumstances the disease itself spreads but also understand the holistic impact of the pandemic.

Although many people are resilient to the conditions imposed by the pandemic, previous instances of disease outbreaks [1] and quarantines [2] have been associated with negative psychological outcomes. Postinfection conditions that followed previous coronavirus outbreaks include posttraumatic stress disorder, depression, anxiety, and confusion, among others. Similarly, quarantine has been associated with several conditions, including stress [3], posttraumatic stress disorder [4,5], and depression [4,6]. A longer duration of quarantine is associated with worse psychological outcomes [2]—a potentially pertinent fact given the protracted period of the COVID-19 pandemic. Additionally, the stigma of disease and the hazards that many face may differ among different people in different occupations or sociodemographic groups [7].

More recently, the presence of persistent symptoms following acute COVID-19 illness has received increased attention. Around 20% of people in an Office for National Statistics survey from the United Kingdom who had a positive COVID-19 test result reported symptoms lasting at least 5 weeks, and 10% reported symptoms lasting at least 12 weeks [8]. The symptomatologic groups, which are formed by people with persistent illness following SARS-CoV-2 infection, have not been fully determined. Preliminary studies show a multitude of symptoms with various levels of co-occurrence, including persistent respiratory issues, fatigue, psychological and neurological symptoms, and fever [9-11]. The presence of these long-term symptoms is often referred to as *long COVID*.

Mobile health (mHealth) as a field is well suited to the unique problems that have been encountered during the COVID-19 pandemic [12,13]. The need for social distancing and wide-scale quarantines precludes many studies that require direct physical contact with participants. Apart from the ability to continue where other study and data collection methods have been limited, mHealth technologies also offer various advantages. The pervasive nature of mobile phones and wearable fitness devices allows for a fine-grain, second-by-second level of detail as well as prolonged periods of continuous monitoring, which are useful because although the pandemic has been long in duration, it has often been punctuated by acute events, such as infection or the introduction of public health measures. Moreover, the fine resolution of such data provides a more comprehensive view of a person's health and behavior. Historic fitness, health, and activity records are often connected to a person's web-based accounts. Participants are able to donate such data, which can be used to better understand changes related to participants' prepandemic activities and health, their preinfection status, and the duration required to recover to preinfection baseline. Finally, passive data sets collected in this manner have the benefit of being in a standardized format, regardless of their country or institution of origin, and larger numbers of potential participants can be quickly reached through digital methods compared to those reached through more traditional recruitment strategies.

Various previous and ongoing studies have demonstrated the ability to monitor long-term mental well-being [14,15] and track the prevalence of flu-like disease [16] through the use of remote monitoring technologies (RMTs). Such technologies therefore appear to be a useful lens through which to investigate the COVID-19 pandemic, and multiple initiatives have been set up by several groups [17-19].

Objectives

To investigate some aspects of the COVID-19 pandemic, we launched the Covid Collab study in April 2020. The study is a crowdsourced initiative [20] that will involve remote enrollment. It will use a cross-platform phone app to deliver surveys; allow for the input of COVID-19–related data; and allow participants to connect to third-party sources of wearable data, such as Fitbit LLC. By prospectively collecting regular mental well-being and COVID-19 survey data alongside historic and ongoing health-related wearable device data, we hope to address the following objectives.

We will determine whether remote monitoring can provide data on COVID-19 states with objective, measurable differences. Wearable device data have previously been used to predict the prevalence of influenza-like illnesses [16] and can therefore potentially be used to better understand levels of infection and persistent postsequelae symptoms. We aim to assess the feasibility of detecting acute infections, wellness, and long COVID symptoms at a personal and population level.

We will also stratify and define patterns of symptoms of COVID-19 and any postacute infection illness. Self-reported symptoms and objective measures of activity from wearables will be used to identify any groups or patterns of symptoms, especially those among the nonhospitalized population, which has been less visible and easy to recruit in many studies.

We also aim to identify risk factors and causes of COVID-19, long COVID, and the severity of illness. The incidence of COVID-19 and the likelihood of a person developing persistent symptoms following infection will be investigated with respect to a person's state prior to enrollment, which will be based on sociodemographic information; participants' prior medical histories; and wearable- and phone-derived information, such as activity levels, heart rates, and sleeping patterns.

Finally, we will investigate mental well-being throughout the pandemic. Alongside measures of SARS-CoV-2 infection, we will also collect regular responses to mental well-being surveys. We will describe trajectories of mental well-being in response to illness and PHSMs during the pandemic as well as identify risk and protective factors.

Methods

Study Design

The Covid Collab study is a crowdsourced observational study that will involve remote enrollment. Covid Collab aims to collect wearable device data, phone data, and survey responses from a large number of self-enrolled participants. This is an observational population study with several structures available for particular objectives. Cross-sectional comparisons will involve drawing cases and controls from participants who have and have not reported illness during the course of the study. By conducting individual longitudinal comparisons and participant-specific models, baseline measurements will be compared against measurements from different stages of COVID-19 (ie, acute infection and postinfection) or from periods of interest (eg, vaccination periods and lockdowns).

Recruitment

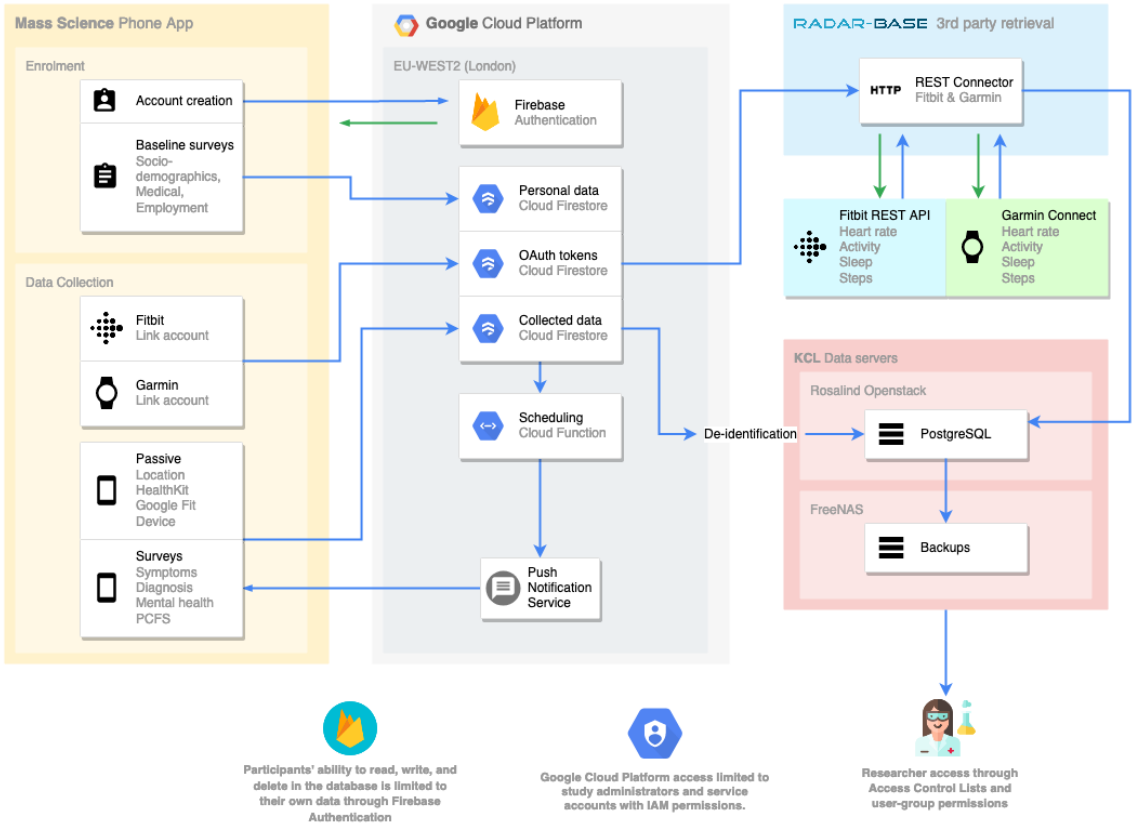
Recruitment started in April 2020 on a small scale, and large-scale recruitment began in June 2020. Given the crowdsourced nature of the study, participants will be able to enroll from anywhere. However, because of the location of our research group, the majority of the promotional activities that have been carried out have targeted people within the United Kingdom. The study is open to enrollment for any person over the age of 18 years who uses a smartphone and, optionally, a wearable fitness device. Participants without a fitness device will still be able to complete COVID-19 and mental health surveys.

Participants will enroll within the Mass Science app—the study app for Covid Collab. During enrollment, the participants will be provided with in-app study information, an in-app consent form, and a basic demographics survey. Directly following enrollment, the participants will go through an onboarding procedure. First, participants will complete a more in-depth demographic survey for collecting information on age, gender, ethnicity, height, weight, previous and existing medical conditions, employment status and whether there has been a change in employment status during the pandemic, and marital status. Second, participants will receive prompts for optionally turning on the location data sharing function in the background of their smartphones throughout their involvement in the study. They will also receive prompts for connecting their wearable device accounts to facilitate wearable device data collection.

Platform and Mass Science App

To facilitate the study, we used pieces of the Remote Assessment of Disease and Relapse (RADAR)–base mHealth data [21] collection platform, alongside services from Google Cloud Platform, as the data collection back end and a custom-built app for remote enrollment and participant interaction (Figure 1).

Figure 1. An overview of the data collection platform that will be used in the Covid Collab study. API: application programming interface; IAM: Identity and Access Management; KCL: King's College London; OAuth: Open Authorization; PCFS: Post–COVID-19 Functional Status; REST: Representational State Transfer.



The Mass Science app is a cross-platform smartphone app that was developed for the Covid Collab study using Flutter. Its key functionalities include providing prospective participants with the ability to enroll in the study; delivering scheduled surveys; allowing participants to input information related to SARS-CoV-2 infection and vaccination; collecting wearable device data either directly from phones or by requesting access to participants' data through third-party application programming interfaces (APIs); and collecting phone data, including location information. The collection of each data type (eg, location) will be optional. This will allow people to provide data that they are comfortable to share.

Google Cloud Platform [22] comprises the majority of the back end. User authentication will be managed through Firebase Authentication (Google LLC), survey scheduling will be managed through Cloud Functions and Firebase Cloud Messaging (Google LLC), and the initial collection of phone data and surveys will be conducted through Firestore (Google LLC).

RADAR-base is a general mHealth data collection platform that has been used in several RMT studies [14,23,24]. It comprises several modular applications. Some wearable device companies provide access to their customers' data through an API (a set of definitions and protocols that ease programmatic access to services). We will use the RADAR-base Representational State Transfer Connector to retrieve wearable device data from the Fitbit Web API (Fitbit LLC) and Garmin Health API (Garmin Ltd).

Procedures and Data Collection

Surveys

A number of baseline, on-demand, and scheduled surveys (Table 1) will be included in the study and completed by participants through the Mass Science app. Sociodemographic and medical information will be collected at baseline. Mental health questionnaires—the Patient Health Questionnaire-8 (PHQ-8) scale [25] for symptoms of depression and the General Anxiety Disorder-7 (GAD-7) scale [26] for symptoms of anxiety—will be made available, and participants will be prompted to complete these questionnaires every 2 weeks. A questionnaire on COVID-19 and long COVID symptoms and a visual analog happiness and energy scale will also be made available. These can be completed on demand, but participants will also be prompted biweekly to complete them. COVID-19 diagnosis and vaccination information can be submitted at any time. Following a reported COVID-19 diagnosis, participants will be prompted to fill in the Post–Covid-19 Functional Status scale [27]. Prompts regarding when a scheduled survey is available will be delivered through Firebase Cloud Messaging push notifications, which will appear as notifications on participants' phones.

Table 1. The surveys that will be collected during the study.

Questionnaires	Purpose	Frequency
Baseline questionnaires		
Covid Collab demographics (Multimedia Appendix 1)	To collect demographic data	Baseline
Covid Collab comorbidities (Multimedia Appendix 1)	To collect data on disorders and comorbidities	Baseline
Scheduled questionnaires		
Post–COVID-19 Functional Status scale [27]	A fast ordinal scale for the evaluation of post–COVID-19 functional status	Fortnightly following diagnosis
COVID-19 symptoms (Multimedia Appendix 1)	To catalog acute-phase and lingering COVID-19 symptoms and long COVID symptoms	Twice weekly and on demand
General Anxiety Disorder-7 [26]	To identify probable cases of anxiety and to determine the severity of symptoms	Fortnightly
Patient Health Questionnaire-8 [25]	To assess the severity and presence of symptoms of depression	Fortnightly
On-demand questionnaires		
Diagnosis (Multimedia Appendix 1)	Self-report diagnosis questionnaire	On demand
Vaccination (Multimedia Appendix 1)	Vaccination survey	On demand

Wearables

Wearable device data will be collected through 2 methods. First, participants can connect their web-based accounts, thereby allowing us to collect data from the wearable vendors' HTTP API. Both Fitbit LLC and Garmin Ltd will provide data access through this method by allowing users to authorize Covid Collab to access their data through the companies' respective APIs. In this case, data can be retrieved directly from a server. Second, we can retrieve data via users' smartphones by using Apple HealthKit (Apple Inc) [28] and Google Fit (Google LLC) [29]. In this case, data will be uploaded to Firestore alongside other phone data.

The exact data types that will be available will depend on the devices that the participants use, what the wearable device manufacturers make available, and what the users choose to authorize when allowing access to their wearable data. Where available, we will collect intraday and summary heart rate, step count, and activity data; sleep data; and other physical and health information, such as height and weight. If a participant does not own a wearable device, they will still be able to provide survey responses and phone data through the Mass Science app.

We expect that some participants will have existing data for the periods of time preceding enrollment and the pandemic. After they connect their wearable device accounts, we will retrospectively collect wearable device data from January 2019, where available. Prospective data will be retrieved as they become available.

Location

Geographic position data will be collected from consenting participants' phones. To reduce battery use, a location point will be recorded only when movement is detected and not when participants are stationary. Raw location data are highly sensitive. As such, they will be stored separately, and only deidentified features and summary statistics will be accessible to researchers. Following a change in stance by the Google Play Store (Google LLC) in January 2021, location collection was discontinued in subsequent updates of the Android app.

Data Enrichment

Analyses will require the enrichment of the data through the incorporation of publicly available data sets. Primarily, this will be performed via the contextualization of location data by using the CORINE (Coordination of Information on the Environment) Land Cover data set [30] and OpenStreetMap (OpenStreetMap Foundation) and via the incorporation of public and social measures from the World Health Organization PHSM database [31].

Data Management

All data will be stored and encrypted, and personal information will be stored in a separate database. Location data will be deidentified via the aggregation of raw geographic coordinates into features. Access to personal information will be limited strictly to study administrators for administration purposes (eg, to delete data at the request of a participant). Researchers' access to the anonymized data set will be limited through access control lists. Participants can choose to allow us to share anonymized versions of their data in a larger public data set, which will be made available at a later date.

Statistical Analysis

Data Exclusion and Absence

As a crowdsourced study involving the optional sharing of different modalities of data, we expect that there may be greater data missingness and participant attrition than those in studies that involve more direct patient contact and engagement. Different objectives may require different exclusion criteria. For example, determining wearable biomarkers for COVID-19 may only require a connected device and a single COVID-19 diagnosis survey, while characterizing trajectories of mental well-being would require multiple PHQ-8 and GAD-7 responses from a single participant.

Rates of participation, adherence, and dropout will be examined with respect to sociodemographics, time points during the pandemic, and participants' concurrent health. Additionally, patterns of user engagement will be characterized to show how participants may interact in similar studies and what drives engagement. User engagement will be determined based on completion rates for the prompted surveys.

Characterizing COVID-19 and Long COVID Symptomology

We will describe and define subgroups for symptoms of COVID-19 and long COVID through the clustering of self-reported symptoms. This will include a time-independent view of all symptoms throughout the illness as well as time-dependent clustering to investigate how the disease progresses. A latent class analysis will be used to group time-independent symptoms. A cluster analysis will be conducted on symptom severity data (4-point Likert scale). The optimal number of latent classes will be selected based on the Bayesian information criterion. Time-dependent symptom clustering will be carried out by using mixture latent Markov models. The classes will be described with respect to the frequency of symptoms and their prevalence in different sociodemographic groups.

Risk Factors for Severe COVID-19 and Long COVID

Risk factors will be assessed by conducting a logistic regression between participants with long COVID symptoms and participants who had COVID-19 but did not experience persistent symptoms. A logistic regression between participants with COVID-19 who self-report severe symptoms (based on a 4-point Likert scale) and those who self-report mild symptoms or are asymptomatic will also be conducted. Predictors will include sociodemographics, smoking status, medical history, and measures of health and behavior derived from the RMT passive data streams (eg, historic and contemporary activity levels and heart rates).

Disease State Classification

By using the identified clusters of symptoms, we will explore RMT parameters that can be used to classify COVID-19. This analysis will involve using conventional machine learning methods, including support vector machines and random forests, in combination with feature selection and fusion approaches, as well as more contemporary deep learning methods.

Trajectories and Classification of Mental Well-being

The primary mental health outcome measures will be the PHQ-8 and GAD-7 for depression and anxiety, respectively, which participants will be prompted to complete every 2 weeks. Additionally, a visual analogue scale for happiness and for energy will be included alongside the biweekly symptoms questionnaire.

Mental well-being will be investigated from several viewpoints. First, we will analyze how mood changes in response to and following a SARS-CoV-2 infection. Second, we will determine how mental well-being has been affected throughout the pandemic for the entire cohort in relation to public health measures and by taking into account levels of activity and information on location (eg, time spent outside, home stay duration, or local population density). Finally, we will assess the feasibility of using machine learning approaches to predict low mood on the basis of passive wearable and phone data.

Results

The Covid Collab study began in April 2020, and large-scale recruitment began in earnest in June 2020. As of June 2021, there are over 17,000 participants. Of those, 11,350 have a connected wearable device, and 16,350 have completed at least 1 survey. An interim analysis is expected to be complete by July 2021. The publication of the final analyses is expected to occur by December 2022 but may depend on the evolution of the COVID-19 pandemic.

Discussion

Remote monitoring is a promising avenue for understanding COVID-19 and the effects of the pandemic. Our study has multiple advantages, including the availability of historic wearable device data, the ability to reach a wide range and large number of people, and the high resolution of data. However, there are also a number of limitations to the study.

Although crowdsourced recruitment is technically open to all, it is likely that there will be bias. The study is only reachable by those who own a smartphone, and a person who already owns a wearable device may be more likely to take part in the study. Both of these populations may be skewed, in some respect, relative to the general population. Moreover, different segments of the population may be more likely to seek out and engage with scientific studies of this kind. For example, within our currently enrolled cohort, 68.6% (11,840/17,255) of participants are female. It will be important to quantify the composition of the cohort and determine how the composition relates to the known COVID-19 incidence rates among different groups, study adherence rates, and data completion rates within the study.

Participant attrition is present in many internet-based studies [32]. As previously mentioned, due to the nature of a study involving remote enrollment and little to no personal interaction, we may expect higher attrition rates than those in studies with different enrollment strategies or methods for promoting participant interaction and engagement [33]. A “history view” screen was implemented in the app. It shows previous mood responses to allow for the direct return of results to participants. However, other studies have used other methods for promoting participant engagement that are not present in our study largely due to development time limitations. Such methods include different notification strategies [34,35] and gamification [36,37].

Another limitation is imposed by the evolving nature of the pandemic and our knowledge of COVID-19; in response to new information, we may be required to change aspects of or add to the protocol. For example, long COVID symptoms and the Post–COVID-19 Functional Status scale were added recently, as more evidence of persistent impairment following SARS-CoV-2 infection has emerged. Time constraints also require us to balance the introduction of features with the need to recruit participants at an earlier stage. For example, the use of the Garmin Health API was recently included in the protocol. This may have limited the prior recruitment of users of Garmin devices. However, current and prospective participants with Garmin devices will still be able to donate historic data connected to their accounts.

There are several similar ongoing studies throughout the world. Although our participants may overlap with those of other studies, each study is fairly well geographically separated. Although we are recruiting participants from throughout the world, as a UK-based group, our outreach and ability to connect with potential participants are greatest in the United Kingdom. Given the similarity of the collected data and the loose alignment of questionnaires, there is potential for collaboration or meta-analysis.

Overall, the introduced study ought to provide an angle through which to view the mental and physical health of a population throughout the COVID-19 pandemic. Using historic and ongoing wearable and mHealth data should allow for more thorough precision health models to be built and enable us to understand how prior lifestyles have affected the risk of developing COVID-19, long COVID symptoms, and mental health issues.

Acknowledgments

The views expressed are those of the authors and not necessarily those of the National Health Service (NHS), National Institute for Health Research (NIHR), or Department of Health and Social Care. This study was supported by (1) the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) the BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement No. 116074. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and European Federation of Pharmaceutical Industries and Associations (EFPIA); it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and the European Society of Cardiology (ESC); (4) the National Institute for Health Research University College London Hospitals Biomedical Research Centre; (5) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; (6) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King’s College Hospital NHS Foundation Trust; and (7) the research computing facility at King’s College London, Rosalind [38]

Conflicts of Interest

None declared.

Multimedia Appendix 1

Definitions for the unpublished questionnaires that will be used in the Covid Collab study.
[DOCX File , 11 KB-Multimedia Appendix 1]

References

1. Rogers JP, Chesney E, Oliver D, Pollak TA, McGuire P, Fusar-Poli P, et al. Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic. Lancet Psychiatry 2020 Jul;7(7):611-627 [FREE Full text] [doi: 10.1016/S2215-0366(20)30203-0] [Medline: 32437679]
2. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar 14;395(10227):912-920 [FREE Full text] [doi: 10.1016/S0140-6736(20)30460-8] [Medline: 32112714]
3. DiGiovanni C, Conley J, Chiu D, Zaborski J. Factors influencing compliance with quarantine in Toronto during the 2003 SARS outbreak. Biosecur Bioterror 2004;2(4):265-272. [doi: 10.1089/bsp.2004.2.265] [Medline: 15650436]
4. Hawryluck L, Gold WL, Robinson S, Pogorski S, Galea S, Styra R. SARS control and psychological effects of quarantine, Toronto, Canada. Emerg Infect Dis 2004 Jul;10(7):1206-1212 [FREE Full text] [doi: 10.3201/eid1007.030703] [Medline: 15324539]
5. Reynolds DL, Garay JR, Deamond SL, Moran MK, Gold W, Styra R. Understanding, compliance and psychological impact of the SARS quarantine experience. Epidemiol Infect 2008 Jul;136(7):997-1007. [doi: 10.1017/S0950268807009156] [Medline: 17662167]
6. Guo Y, Cheng C, Zeng Y, Li Y, Zhu M, Yang W, et al. Mental health disorders and associated risk factors in quarantined adults during the COVID-19 outbreak in China: Cross-sectional study. J Med Internet Res 2020 Aug 06;22(8):e20328 [FREE Full text] [doi: 10.2196/20328] [Medline: 32716899]
7. Pfefferbaum B, North CS. Mental health and the Covid-19 pandemic. N Engl J Med 2020 Aug 06;383(6):510-512. [doi: 10.1056/NEJMp2008017] [Medline: 32283003]
8. Venkatesan P. NICE guideline on long COVID. Lancet Respir Med 2021 Feb;9(2):129 [FREE Full text] [doi: 10.1016/S2213-2600(21)00031-X] [Medline: 33453162]
9. Sykes DL, Holdsworth L, Jawad N, Gunasekera P, Morice AH, Crooks MG. Post-COVID-19 symptom burden: What is long-COVID and how should we manage it? Lung 2021 Apr;199(2):113-119 [FREE Full text] [doi: 10.1007/s00408-021-00423-z] [Medline: 33569660]
10. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of Long-COVID: analysis of COVID cases and their symptoms collected by the Covid Symptoms Study App. medRxiv. Preprint posted online on December 19, 2020 [FREE Full text] [doi: 10.1101/2020.10.19.20214494]
11. Lopez-Leon S, Wegman-Ostrosky T, Perelman C, Sepulveda R, Rebolledo PA, Cuapio A, et al. More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. medRxiv. Preprint posted online on January 30, 2021 [FREE Full text] [doi: 10.1101/2021.01.27.21250617]
12. Amft O, Favela J, Intille S, Musolesi M, Kostakos V. Personalized pervasive health. IEEE Pervasive Comput 2020 Jul 16;19(3):11-13 [FREE Full text] [doi: 10.1109/mprv.2020.3003142]
13. Amft O, Gonzalez LIL, Lukowicz P, Bian S, Burggraf P. Wearables to fight COVID-19: From symptom tracking to contact tracing. IEEE Pervasive Comput 2020 Nov 25;19(4):53-60 [FREE Full text] [doi: 10.1109/mprv.2020.3021321]
14. Matcham F, di San Pietro CB, Bulgari V, de Girolamo G, Dobson R, Eriksson H, RADAR-CNS consortium. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72 [FREE Full text] [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
15. Sun S, Folarin AA, Ranjan Y, Rashid Z, Conde P, Stewart C, RADAR-CNS Consortium. Using smartphones and wearable devices to monitor behavioral changes during COVID-19. J Med Internet Res 2020 Sep 25;22(9):e19992 [FREE Full text] [doi: 10.2196/19992] [Medline: 32877352]
16. Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. Lancet Digit Health 2020 Feb;2(2):e85-e93 [FREE Full text] [doi: 10.1016/S2589-7500(19)30222-5] [Medline: 33334565]
17. Mishra T, Wang M, Metwally AA, Bogu GK, Brooks AW, Bahmani A, et al. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat Biomed Eng 2020 Dec;4(12):1208-1220. [doi: 10.1038/s41551-020-00640-6] [Medline: 33208926]
18. Quer G, Radin JM, Gadaleta M, Baca-Motes K, Ariniello L, Ramos E, et al. Wearable sensor data and self-reported symptoms for COVID-19 detection. Nat Med 2021 Jan;27(1):73-77. [doi: 10.1038/s41591-020-1123-x] [Medline: 33122860]
19. Corona-Datenspende. Robert Koch-Institut. URL: https://corona-datenspende.de/ [accessed 2021-08-02]

20. Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res 2012 Mar 07;14(2):e46 [FREE Full text] [doi: 10.2196/jmir.1988] [Medline: 22397809]
21. Ranjan Y, Rashid Z, Stewart C, Conde P, Begale M, Verbeeck D, Hyve, RADAR-CNS Consortium. RADAR-Base: Open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019 Aug 01;7(8):e11734 [FREE Full text] [doi: 10.2196/11734] [Medline: 31373275]
22. Cloud Computing Services. Google Cloud. URL: https://cloud.google.com/ [accessed 2021-09-10]
23. Bruno E, Biondi A, Böttcher S, Vértes G, Dobson R, Folarin A, et al. Remote assessment of disease and relapse in epilepsy: Protocol for a multicenter prospective cohort study. JMIR Res Protoc 2020 Dec 16;9(12):e21840 [FREE Full text] [doi: 10.2196/21840] [Medline: 33325373]
24. Muurling M, de Boer C, Kozak R, Religa D, Koychev I, Verheij H, RADAR-AD Consortium. Remote monitoring technologies in Alzheimer's disease: design of the RADAR-AD study. Alzheimers Res Ther 2021 Apr 23;13(1):89 [FREE Full text] [doi: 10.1186/s13195-021-00825-4] [Medline: 33892789]
25. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
26. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
27. Klok FA, Boon GJAM, Barco S, Endres M, Geelhoed JJM, Knauss S, et al. The Post-COVID-19 Functional Status scale: a tool to measure functional status over time after COVID-19. Eur Respir J 2020 Jul 02;56(1):2001494 [FREE Full text] [doi: 10.1183/13993003.01494-2020] [Medline: 32398306]
28. HealthKit. Apple Developer Documentation. URL: https://developer.apple.com/documentation/healthkit [accessed 2021-09-13]
29. Google Fit. Google Developers. URL: https://developers.google.com/fit [accessed 2021-09-13]
30. CLC 2018. Copernicus Land Monitoring Service. URL: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018 [accessed 2021-08-02]
31. Tracking public health and social measures. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm [accessed 2021-08-02]
32. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
33. Meyerowitz-Katz G, Ravi S, Arnolda L, Feng X, Maberly G, Astell-Burt T. Rates of attrition and dropout in app-based interventions for chronic disease: Systematic review and meta-analysis. J Med Internet Res 2020 Sep 29;22(9):e20283 [FREE Full text] [doi: 10.2196/20283] [Medline: 32990635]
34. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018 Nov 29;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
35. Mehrotra A, Pejovic V, Vermeulen J, Hendley R, Musolesi M. My phone and me: Understanding people's receptivity to mobile notifications. 2016 May Presented at: CHI'16: CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, California, USA p. 1021-1032. [doi: 10.1145/2858036.2858566]
36. Floryan M, Chow PI, Schueller SM, Ritterband LM. The model of gamification principles for digital health interventions: Evaluation of validity and potential utility. J Med Internet Res 2020 Jun 10;22(6):e16506 [FREE Full text] [doi: 10.2196/16506] [Medline: 32519965]
37. Edney S, Ryan JC, Olds T, Monroe C, Fraysse F, Vandelanotte C, et al. User engagement and attrition in an app-based physical activity intervention: Secondary analysis of a randomized controlled trial. J Med Internet Res 2019 Nov 27;21(11):e14645 [FREE Full text] [doi: 10.2196/14645] [Medline: 31774402]
38. Rosalind: Research computing infrastructure. King's College London. URL: https://rosalind.kcl.ac.uk [accessed 2021-11-23]

Abbreviations

API: application programming interface
CORINE: Coordination of Information on the Environment
GAD-7: General Anxiety Disorder-7
mHealth: mobile health
NHS: National Health Service
NIHR: National Institute for Health Research
PHQ-8: Patient Health Questionnaire-8
PHSM: public health and social measure
RADAR: Remote Assessment of Disease and Relapse
RMT: remote monitoring technology

6.7 Summary

In this chapter the Covid Collab study protocol was presented. While the objective was to set the scene for the following analysis chapter, there are a couple of unique aspects and contributions.

Firstly, consent was sought to share the collected data in a public dataset. The dataset contains extensive wearable recordings and mental health measurements and is therefore potentially useful outside of COVID-19. Several projects have started to use data within and outside of COVID-19 research, including a self-supervised learning approach to building a 'wearables' foundational model, resting heart rate forecasting, and circadian rhythms.

Secondly, the inclusion of historic wearable data and mental health measures is in contrast to otherwise similar citizen science studies. Chapter 4 demonstrated the potential utility of short baseline recordings and limited prior labelled data. As such, I felt it was important to include historic data to better contextualise or correct participant's recordings during the pandemic, as the following chapter will help illustrate.

Chapter 7

Covid Collab: Presentation of Long COVID and Risk Factor Analysis in a Mobile Health Study

7.1 Preamble

As the pandemic progressed, it became clear that long-term sequelae of COVID-19 infection was an important and under researched area. The Covid Collab study was initially set up to monitor acute symptoms of COVID-19, but because it had continuous wearable-based monitoring, did not limit the period in which questionnaires were prompted for, and had frequent mental health questionnaires, it was well-placed to also study longer-term effects.

Similar citizen science wearable data donation studies also reoriented around post-COVID sequelae from an earlier focus on acute detection.[246] In the *DETECT* study, Radin et al. found COVID-postive participants were more likely to have an increased heart rate of five or more beats per minute after 56 days than those without an infection.[15] The Corona-Datenspende study put increased emphasis on long COVID in their 2023 roadmap[240] and have demonstrated the impact of vaccination on the recovery rate from COVID infection through wearables.[247] A traditionally-recruited observational study demonstrated reduced light and deep sleep time, as measured through a Biostrap smartband, in COVID-postive participants compared to a control group.[248] Together these results suggested that monitoring certain long-term damage or dysfunction following infection is possible using wearable technology.

The large self-report mobile health studies also contributed to the understanding of long COVID. The *COVID Symptom Study* reported prevalence rates of 4.5% at eight weeks and 2.6% at twelve weeks, and suggested that the likelihood of developing long COVID could be

related to symptomatology during the acute infection.[210] Later meta-analyses suggests the prevalence rate is much higher, although estimates are highly variable.[249] The most common symptom, fatigue, is estimated to effect around ⅓ of people for at least twelve weeks. [250,251] Other common symptoms include anxiety and depression, cardiological impairments, respiratory issues, and musculoskeletal pain.

The aims primary aims of this chapter are to (i) identify long-term changes to physiological signals, mental health, and reported symptoms following a COVID-19 infection at a group-wide level and (ii) to determine risk-factors for the development of long COVID from a passive- and active-RMT perspective. A few unique aspects of the study, most notably the historic data and the inclusion of mental health measures from the very start of the study, set it apart from the other wearable data donation drives.

This chapter is included as a pre-print of a paper that is in the publication process. The self-report analysis was primarily performed by Yatharth Ranjan and the text relating to that section of work was predominantly written by him. I performed and wrote the other sections of work, including the group-wide and passive-based long COVID cohort analysis, with feedback and advice from the other listed authors.

DOI 10.1101/2022.09.27.22280404

Presentation of long COVID and associated risk factors in a mobile health study

Callum Stewart[1] Yatharth Ranjan[1] Pauline Conde[1]
Shaoxiong Sun[1] Zulqarnain Rashid[1] Heet Sankesara[1]
Nicholas Cummins[1] Petroula Laiou[1] Xi Bai[1]
Richard J B Dobson[1,2] Amos A Folarin[1,2*]

[1]Department of Health Informatics and Biostatistics, Institute of Psychiatry, Psychology and Neuroscience, Kings College London, UK
[2]Institute of Health Informatics, University College London, UK

September 22, 2022

Abstract

Background The Covid Collab study was a citizen science mobile health research project set up in June 2020 to monitor COVID-19 symptoms and mental health through questionnaire self-reports and passive wearable device data.

Methods Using mobile health data, we consider whether a participant is suffering from long COVID in two ways. Firstly, by whether the participant has a persistent change in a physiological signal commencing at a diagnosis of COVID-19 that last for at least twelve weeks. Secondly, by whether a participant has self-reported persistent symptoms for at least twelve weeks. We assess sociodemographic and wearable-based risk factors for the development of long COVID according to the above two categorisations.

Findings Persistent changes to physiological signals measured by commercial fitness wearables, including heart rate, sleep, and activity, are visible following a COVID-19 infection and may help differentiate people who develop long COVID. Anxiety and depression are significantly and persistently affected at a group level following a COVID-19 infection. We found the level of activity undertaken in the year prior to illness was protective against long COVID and that symptoms of depression before and during the acute illness may be a risk factor.

Interpretation Mobile health and wearable devices may prove to be a useful resource for tracking recovery and presence of long-term sequelae to COVID-19. Mental wellbeing is significantly negatively effected on average for an extended period of time following a COVID-19 infection.

*Corresponding Author (amos.folarin@kcl.ac.uk)

Introduction

There have been over 500 million confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections as of April 2022 [1]. Despite the development and successful rollout of vaccinations and treatments, COVID-19 remains a danger both in terms of the acute illness, chance of death, long-term illness following infection, and the possibility of problematic variants developing.

Persistent symptoms following a SARS-CoV-2 infection, often termed *long COVID* or *Post Acute COVID-19 syndrome*, are thought to affect a significant number of patients. The presence of these long-term effects was largely illuminated early in the pandemic through subjective accounts from patients of the disease [2]. While there has been a flurry of research starting to address the prevalence, clinical features, and risk factors [3] of long COVID, our understanding of the condition remains sparse.

Attempts have been made to categorise long COVID (LCOVID) on the basis of symptomatology, time periods, and aetiology, but it remains a loosely defined syndrome with multiple associated terminologies [4]. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) suggest the use of *Acute COVID-19* for symptoms up to four weeks, *Ongoing symptomatic COVID-19* for symptoms from four to twelve weeks, and *Post-COVID-19 syndrome* for symptoms continuing past twelve weeks[5]. Symptoms have been reported across a wide range of organs and body systems, including cardiorespiratory, neurological, psychological, musculatory, gastrointestinal, and systemic[6, 7]. Common symptoms include but are not limited to fatigue, dyspnea, anxiety, sleep disorders, pain, dizziness, and anosmia.

Prevalence estimates for LCOVID vary, partially with respect to study cohorts, terminology, and study design. Many studies recruit from hospitalised populations and therefore select for severe cases of COVID-19. A recent large-scale community study on self-reported symptoms estimated [3.1-5.8]% of participants experienced at least one persistent symptom for over 12 weeks following a COVID infection [3].

The pandemic has been a focal point for the greater emergence of digital health technologies in research and healthcare. Within COVID-19 research, there are multiple large studies using digital health and 'big data' approaches to better understand trajectories of, diagnose, and estimate the prevalence of COVID-19 and its long-term sequelae. Mobile health (mHealth) data modalities can offer an insight into LCOVID complementary to existing studies. As a scalable and continuous data collection method, passive mobile health sensing provides an objective measure of health. Additionally, long periods of historical data is often available from wearable fitness devices. The availability of wearable data outside of medical care pathways grants an avenue to observe people who may otherwise be missed. For example, significant burdens to health and function in people with flu who do not seek medical care because of the sub-clinical nature of their illness have been demonstrated by making use of commercial wearable sensor data [8].

Covid Collab is an observational mHealth study which began in June 2020.

Participants enrolled through the study app, Mass Science, and were prompted to complete regular surveys on COVID symptoms experienced, vaccination and diagnosis status, mood, and mental well-being. Participants were able to share existing and prospective wearable data through their Fitbit and Garmin accounts[9]. We collect wearable data covering a period prior to the pandemic, giving a historic baseline against which to compare. Additionally, we regularly prompted for the completion of self-reported questionnaires on current symptoms and mental health throughout the study, providing a contemporaneous account of mental wellbeing and COVID-19-associated symptoms before, throughout, and after any COVID-19 infection.

We queried Pubmed from inception up until 01 Jul 2022 for studies on LCOVID that use wearable or mHealth technologies using the string '((COVID* OR SARS-COV-2) AND (long OR persistent OR hauler OR post OR sequelae)) AND (mHealth OR wearable OR telemedicine OR app)'. Of the 2144 results, the vast majority of returned studies concerned the role of telemedicine in delivering care of other conditions since the start of the COVID-19 pandemic. A number of studies relate to the monitoring, detection, or diagnosis of acute COVID-19. Others use digital technology or telemedicine in the treatment or to assess the impact of rehabilitation courses. Eight studies investigate LCOVID through remote digital technologies. Of those, six used self-reported questionnaire data collected through televisits or apps to characterise symptomatology and trajectories of LCOVID, including a large-scale community study from the ZOE COVID Symptom app. Three studies use passively collected data from commercial or experimental wearable sensors to describe how heart rate, sleep, and activity change in a COVID-19 positive population following infection. Two of those show a pattern of bradycardia and tachycardia in resting heart rate, with a persistent change in some cases lasting over four months.

Covid Collab provides a unique viewpoint for quantifying the features and risk factors of LCOVID. This study incorporates survey data alongside pervasive wearable sensor data. Participant self-reported questionnaires included regular mental health measures as well as physical COVID-19 related symptoms. The study population was recruited remotely and openly throughout the pandemic. It therefore includes non-hospitalised participants and those with a mild response to acute COVID-19 infection, with data often collected prior to infection which does not rely on participant recall and the inclusion of historic wearable data from prior to enrolment. Software and data collection infrastructure have been open sourced to facilitate the reuse of this system for future digital epidemiology research or monitoring programmes.

The aims of this paper are to: (1) Quantify the prevalence and severity of long-term symptoms across collected mHealth metrics including heart rate, sleep, physical activity, and self-reported symptoms and mood. (2) Identify risk factors for the severity and duration of persistent symptoms. Accordingly LCOVID (LCOVID) is considered here as persistent changes or symptoms at and beyond the twelfth week post-diagnosis.

Methods

Study design and participants

The study is a longitudinal self-enrolled community study administered through a smartphone app. As of August 2022, 17,667 participants had enrolled. Participation was open to any adult greater than or equal to 18 years of age. However, enrolment was skewed towards those based in the United Kingdom (UK) and of female gender (N=12137, 68.7%) [Table 1]. Recruitment was carried out via the study app, media publications, and promotion within the UK version of the Fitbit app between August 2020 and May 2021.

Participants were included in the analysis if they reported a COVID-19 diagnosis before 2022/02/01. In total, 1,743 participants were included. Different numbers of participants were included in different aspects of the analysis because of the differing rates of completion between modalities. A group of age-, sex-, and time-matched controls (N=3,600) were selected from participants with high questionnaire completion rates who had not reported a COVID-19 diagnosis.

A detailed explanation of the study protocol in Covid Collab has previously been published [9]. The development of the pandemic and our changing understanding of the disease and requirements of the study led to some amendments to the protocol, including an extended socio-demographics questionnaire building on the initial registration questionnaire. Additionally, participants were able to donate from different sources of data, but were free to provide as much as they chose. Because of those two reasons, there are differences in data availability between participants. Of the participants included in this study, 759 (43.5%) had completed the extended socio-demographic questionnaire.

Study metrics

There are two major categories of data in this study: passive and active data (Table 2). Active data refers to questionnaires delivered in-app that require active participant engagement. The questionnaires include the PHQ-8 scale of depression[10], the GAD-7 scale of generalised anxiety[11], a visual analog scale for arousal and valence[12], and a COVID-19 symptoms questionnaire. In addition, participants are able to submit COVID-19 diagnosis (antigen, PCR, or symptom determined) and vaccination events. Passive data refers to data collected from instruments that do not require any conscious participant involvement. In this study, the passive metrics consist of heart rate, heart rate variability, sleep, step counts, and activity logs provided through the participant's sharing of commercial wearable sensor data.

Group analysis

A comparison of passive and self-reported metrics between case and control groups was carried out. Group-wide resting heart rate (RHR), heart rate variability (HRV) measured as the Fitbit daily root mean square of successive

Demographic category	Cases	Controls
Initial questionnaire		
Age [min-max]	44.78ś13.18 [15.0-87.0]	48.30ś13.28 [15.00-88.00]
Sex = f	1247 (76.1%)	2682 (74.5%)
Height [min-max] (cm)	168.24ś9.83 [118.0-220.0]	168.59ś9.79 [100.00-212.00]
Weight [min-max] (kg)	78.45ś18.43 [26.50-146.20]	77.14ś17.57 [31.70-148.30]
Smoking		
Non-smoker	917 (56.71%)	2118 (58.87%))
Ex-smoker	461 (28.51%)	1030 (28.63%))
Current smoker	239 (14.78%)	450 (12.51%))
Comorbidities (Extended questionnaire, cases n=759, control n=2229)		
Asthma	173 (22.79%)	395 (17.72%)
Hypertension	83 (10.94%)	256 (11.48%)
Diabetes	31 (4.08%)	88 (3.95%)
Depression	173 (22.79%)	532 (23.87%)
Anxiety	136 (17.92%)	356 (15.97%)
Employment (Extended questionnaire)		
Full time	411 (54.15%)	1079 (48.41%)
Part time	125 (16.47%)	321 (14.40%)
Retired	97 (12.78%)	448 (20.10%)
Student	22 (2.90%)	72 (3.23%)
Unemployed	19 (2.50%)	57 (2.56%)
Marital and family status (Extended questionnaire)		
In a relationship	592 (78.72%)	1697 (77.17%)
Single or separated	156 (20.74%)	498 (22.65%)
Unknown	4 (0.53%)	4 (0.18%)
With children	553 (73.54%)	1413 (64.26%)
Living Situation (Extended questionnaire)		
Alone	61 (8.04%)	306 (13.73%)
With partner	454 (59.82%)	1428 (64.06%)
With family	255 (33.60%)	523 (23.46%)
With children	204 (26.88%)	411 (18.44%)
With adult children	122 (16.07%)	240 (10.77%)
Houseshare	20 (2.64%)	57 (2.56%)

Sociodemographic data from initial enrolment and an extended questionnaire across the COVID-positive cohort and the control group of COVID-negative participants.

Table 1: Sociodemographic statistics in Covid Collab

Category	Metric	Frequency	Description
Questionnaires	PHQ-8	14 days	[10]
	GAD-7	14 days	[11]
	Arousal-Valence	Ad-hoc / twice-weekly	[12]
	Symptoms relating to COVID-19	Ad-hoc / twice-weekly	
	Diagnosis	Ad-hoc	Self report diagnosis by Antigen, PCR, Symptom
Fitbit	Heart rate	Daily	Daily resting heart rate provided by Fitbit Web & API.[13]
	Heart rate variability	Daily	Daily root mean square of successive differences (RMSSD) provided by Fitbit Web API[14]
	Sleep duration	Per-sleep	Estimated duration of a sleep[15]
	Sleep efficiency	Per-sleep	Fitbit calculates an 'efficiency' score for each recorded sleep.[15]
	Step count	Daily	[15]
	Activity log	Ad-hoc	[15]

Table 2: Passive and active mobile health metrics collected in this study.

differences (dailyRMSSD), sleep duration, sleep efficiency, step count, PHQ-8 score, GAD-7 score, and self-rated valence and arousal are compared at acute (<4 weeks), ongoing (4-12 weeks), and post-COVID (>12 weeks) syndrome periods, as defined in NICE guidelines[5], following a self-reported diagnosis of COVID-19 to a time-matched group of control participants. Analysis is carried out in Python. A Brunner-Munzul test was carried out between the two groups for each set of metrics using the statsmodels library[16]. Self-reported symptom counts, severity and durations were visualised to understand the symptomatology of COVID-19 and to explore long lasting symptoms.

Risk factors for LCOVID

To test risk factors for LCOVID it is necessary to define a candidate group of participants who are likely to have LCOVID on the basis of the data that we have available. We consider two approaches.

Firstly, we consider using the change in resting heart rate over a period of 12 weeks post-COVID-19 infection as a proxy for LCOVID (the RHR-LCOVID cohort), where a greater change compared to a baseline would indicate a more likely case or greater severity of LCOVID. In this categorisation, we do not explicitly group participants but instead use the change in heart rate as a continuous outcome variable. To do so, we need to estimate a baseline predicted heart rate that the participant would be expected to have if they did not have a COVID-19 infection. To estimate an expected resting heart rate, we fit a Bayesian structural time series model on each participant's historic resting heart

rate up until the date of diagnosis using the CausalImpact library[17]. The model comprises a local-level model, a seasonal model with a period of 28 days, and a regularised regression on a set of 500 participants who were not otherwise involved in the analysis. The change in resting heart rate at 12 weeks is used as the outcome in a linear regression with age, sex, historic activity, historic sleep duration, and the change in RHR between the baseline and acute phase. The historic activity and sleep duration are taken from Fitbit data between one and two years prior to diagnosis. The historic activity is the average duration of time in minutes spent in the Fitbit 'high activity' level per day. The sleep duration is the average time spent asleep per day. The baseline to acute change in RHR is defined as the difference between the average RHR 1-4 weeks prior to a COVID-19 diagnosis and 0-4 weeks after a COVID-19 diagnosis.

Secondly, we consider participants who have self-reported symptoms for an extended period following a self-reported diagnosis of COVID-19. The self-reported symptoms submitted by all participants who reported a positive diagnosis were used to determine length of illness and split the diagnosed cohort into short- and long- COVID groups. If at least one symptom was reported at least once per week for at least twelve weeks, the participant is assigned to the symptom-based LCOVID group (L_{symp}). Participants were otherwise assigned to the symptom-based short COVID group (S_{symp}). Risk factor assessment using logistic regression was performed on the S_{symp} and L_{symp} groups based on demographics, baseline passive data, and mental health scores during the acute phase of COVID infection.

Results

Group-wide analysis

Passive wearable device metrics and self-reported mental health survey scores were compared between case (COVID-19+) and control groups at three time points (Table 3). The 'acute' period was defined as between the date of diagnosis and four weeks post-diagnosis. The 'ongoing' period took values between four weeks and eight weeks post-diagnosis. The 'post-COVID' period took values between twelve and sixteen weeks post-diagnosis. For each period and metric, a comparison of the case and control distributions of the mean values for each of the constituent participants was carried out. A Brunner-Munzul two group non-parametric test[18] was calculated to compare each distribution. Significance was determined with a p-value cutoff of 0.05 after Benjamini-Hochberg correction[19].

Considering first the passive metrics, resting heart rate significantly increased in the COVID-19 positive case group compared to the control group in every period. The difference in heart rate in the acute phase (0.58bpm) is less than the following 'ongoing' period (1.1bpm) and similar to the post-COVID syndrome period after twelve weeks (0.46bpm). This apparent subdued change during the acute infection is because taking the mean does not properly reflect the non-monotonic changes to resting heart rate during this period. The general

group-wide pattern is a peak during the first week of infection, followed by a trough in heart rate during the second week, and finally another, longer lasting, increase (Figure 1). This may imply two acute sub-phases on a shorter timescale than 4 weeks. Step count is also significantly negatively affected during the acute period, but is not significantly changed afterwards. The two sleep metrics show an increase in sleep duration and decrease in efficiency in both case and control groups. There is not a significant difference in duration, but sleep efficiency is significantly decreased throughout all three periods.

All of the self-reported measures of mental health were significantly negatively affected during every period. The mean difference between case and controls for each mental health metric did decrease over time. For example, from a +2.74 (<4 weeks) to +0.98 (>12 week) difference in the average PHQ-8 score. The increased average level and high variance suggests a subset of people suffer persistent symptoms of depression, anxiety, and fatigue (inferred from arousal) for at least twelve weeks.

Metric	**Acute (<4w)**			**Ongoing (4-12w)**			**Post-COVID (>12w)**		
	Case	**Control**	**p-value**	**Case**	**Control**	**p-value**	**Case**	**Control**	**p-value**
RHR	0.69 ± 3.44	0.11 ± 3.15	**6.97e-05**	1.17 ± 3.35	0.07 ± 3.19	**1.17e-16**	0.63 ± 3.44	0.17 ± 3.28	**3.62e-04**
RMSSD	13.58 ± 0.38	13.58 ± 0.39	0.79	13.67 ± 0.43	13.68 ± 0.44	0.75	13.62 ± 0.39	13.62 ± 0.39	0.99
Steps	−1478 ± 3635	39.44 ± 3572	**6.74e-37**	−49.2 ± 3820	128 ± 3630	0.42	288 ± 3828	464 ± 3691	0.93
Sleep efficiency	−1.34 ± 7.09	−0.83 ± 6.30	**5.05e-06**	−1.11 ± 6.75	−0.81 ± 6.47	**0.03**	−1.23 ± 7.72	−0.98 ± 6.58	**1.29e-03**
Sleep duration	5.98 ± 63.06	4.88 ± 56.07	0.67	6.90 ± 65.34	4.53 ± 59.25	0.73	2.81 ± 70.34	2.49 ± 60.78	0.90
PHQ-8	7.96 ± 6.00	5.22 ± 5.30	**4.90e-29**	6.98 ± 6.00	5.29 ± 5.35	**1.13e-07**	6.21 ± 5.57	5.23 ± 5.48	**1.53e-03**
GAD-7	5.88 ± 5.36	4.49 ± 5.02	**1.03e-10**	5.34 ± 5.27	4.60 ± 5.07	**1.47e-03**	5.10 ± 5.37	4.39 ± 4.97	**0.03**
Arousal	−0.18 ± 0.44	0.13 ± 0.41	**5.57e-68**	0.01 ± 0.45	0.13 ± 0.41	**5.15e-07**	0.05 ± 0.44	0.14 ± 0.44	**1.42e-03**
Valence	−0.003 ± 0.36	0.18 ± 0.39	**3.18e-35**	0.11 ± 0.41	0.17 ± 0.39	**3.07e-03**	0.15 ± 0.40	0.21 ± 0.40	**0.01**

Mean value comparison between case and control groups at acute (up to 4 weeks), ongoing (4-12 weeks after diagnosis), and post-COVID (over 12 weeks after diagnosis) time points. There is no significant difference between case and control groups in any metrics in the period 8 weeks to 4 weeks before diagnosis. Resting heart rate (RHR), sleep efficiency, sleep duration, and step count were calculated relative to a baseline level taken as a mean 12 weeks prior to diagnosis. Uncorrected p-values are reported, but emboldened p-values indicate significance after Benjamini/Hochberg correction.

Table 3: Group-wide Comparisons

Risk factors for LCOVID

LCOVID through passive wearable data

A multiple linear regression was carried out to determine whether pre-pandemic historic fitness wearable data could be used as a risk factor for persistent

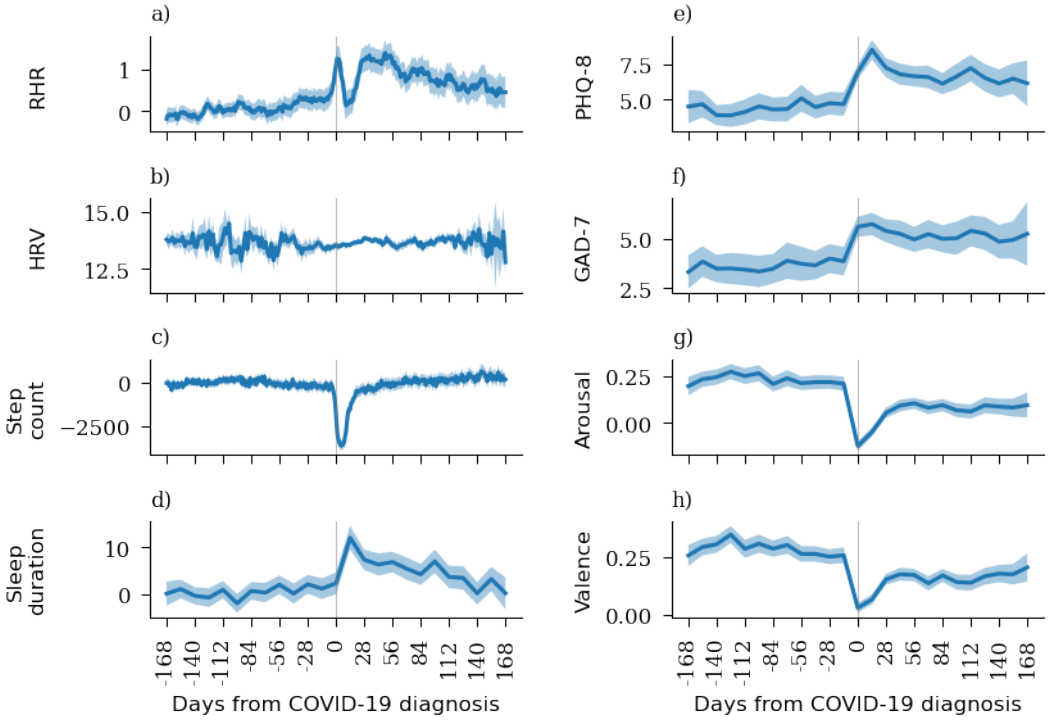


Figure 1: Passive and self-reported measures of mental health across the COVID positive cohort

Dates range from 16 weeks prior to 24 weeks post diagnosis of COVID-19. The shaded area corresponds to the 95% confidence interval taken over 14-day windows.
a) Daily resting heart rate provided through Fitbit Web API. b) Heart rate variability. c) Change to daily step count from baseline. d) Change in sleep duration from baseline. e-f) PHQ-8 and GAD-7 scores respectively. g-h) Arousal and valence scores, which are reported on a visual analogue scale and range from -1 to +1.
HRV: Heart rate variability. RHR: Resting heart rate

elevated RHR at 12 weeks post diagnosis (as a proxy for LCOVID). In total, 597 participants had sufficient passive data available across the whole time period. The outcome variable was the change in resting heart rate between baseline (-4 to -1 weeks) and 12-weeks post-diagnosis among the COVID-positive cohort. The change in RHR between the baseline and acute phase was included as an independent variable to account for the initial change. That is, given a certain change in the acute phase, what variables are significantly associated with that change persisting. Greater historic activity, the mean duration of time spent taking part in heavy activity between one and two years prior to diagnosis, was negatively correlated with an increase in the outcome variable. These results suggest a slight protective effect against LCOVID for younger and more active people. Female sex was positively associated with persistent LCOVID but not significantly so.

Variable	**Coefficient**	**Std Err**	**p-value**	**[0.025 0.975] CI**
Intercept	-1.9786	1.078	0.067	[-4.097 0.139]
Age	0.0274	0.011	**0.017 ***	[0.005 0.050]
Female sex	0.5577	0.337	0.098	[-0.104 1.219]
Historic activity	-0.0166	0.007	**0.015 ***	[-0.030 -0.003]
Historic sleep	0.0027	0.002	0.160	[-0.001 0.007]
Acute ΔRHR	0.2805	0.039	**1.80e-12***	[0.204 0.357]

A linear regression of change in heart rate between baseline and 12 weeks post-diagnosis against age, sex, historic activity, historic sleep, and change in resting heart rate between baseline and acute. P-values are reported uncorrected, but significant results after Benjamini-Hochberg correction are emboldened and marked with an **'*'**

Table 4: Risk factor regression of Long-COVID based on passive wearable data

Long COVID through self-reported symptoms

Self-reported symptoms data are visualised using a heatmap in Figure 2. The colourbar on 0.10, the heatmap represents counts (number of reports) while the severity is denoted by the 3 levels (mild, moderate, severe) on the right y-axis. While most symptoms have highest counts and severity around the diagnosis, some symptoms, such as fatigue, persist for longer with a moderate to high severity. Cough and breathing problems also persist for longer periods but with mild severity.

To further explore how chronic or acute symptoms of COVID-19 and subsequently how these may relate to LCOVID, durations of various symptoms were plotted (Figure 3). This shows fatigue is typically the longest-lasting symptom, with some exceptional cases reporting fatigue for more than 140 days. The duration of the combined category of any symptom shows that some participants have symptoms lasting over 300 days.

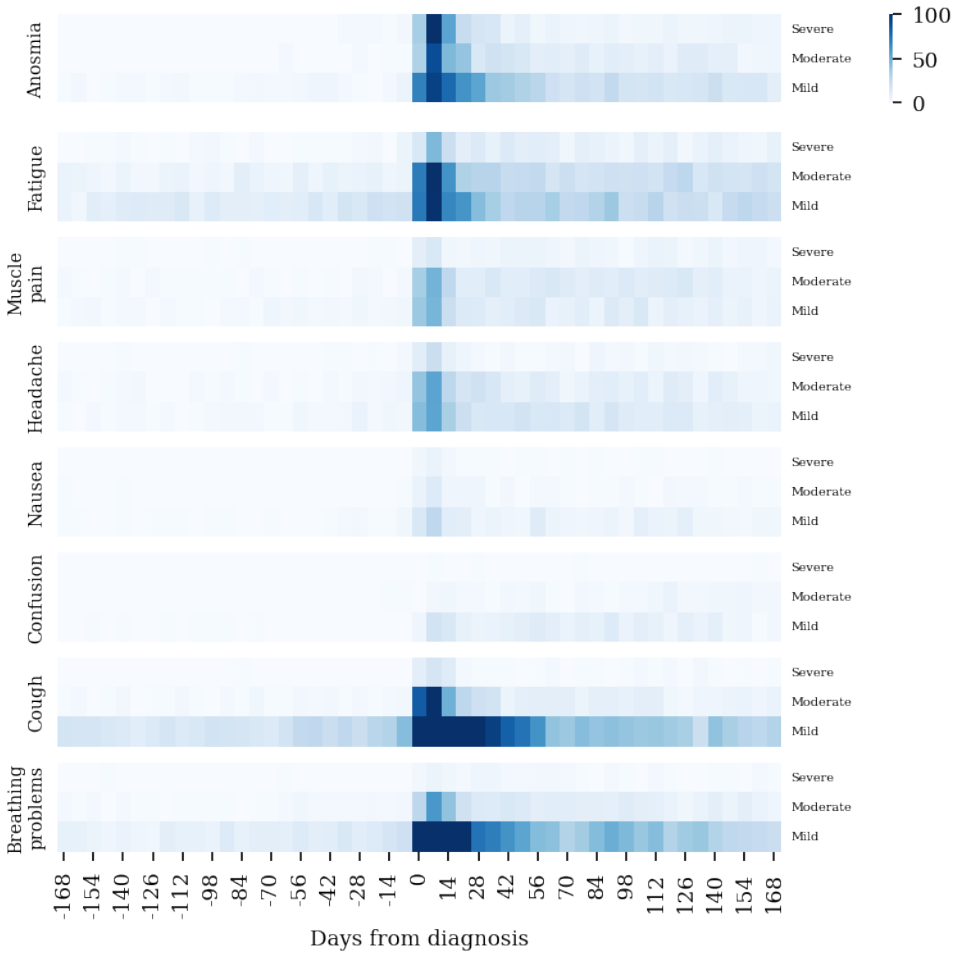


Figure 2: Self-reported symptom heatmap

A heatmap showing counts of self-reported symptom severity among the COVID positive cohort around the date of diagnosis. The count is limited at 100 to better visualise longer-term trends and rarer symptoms.

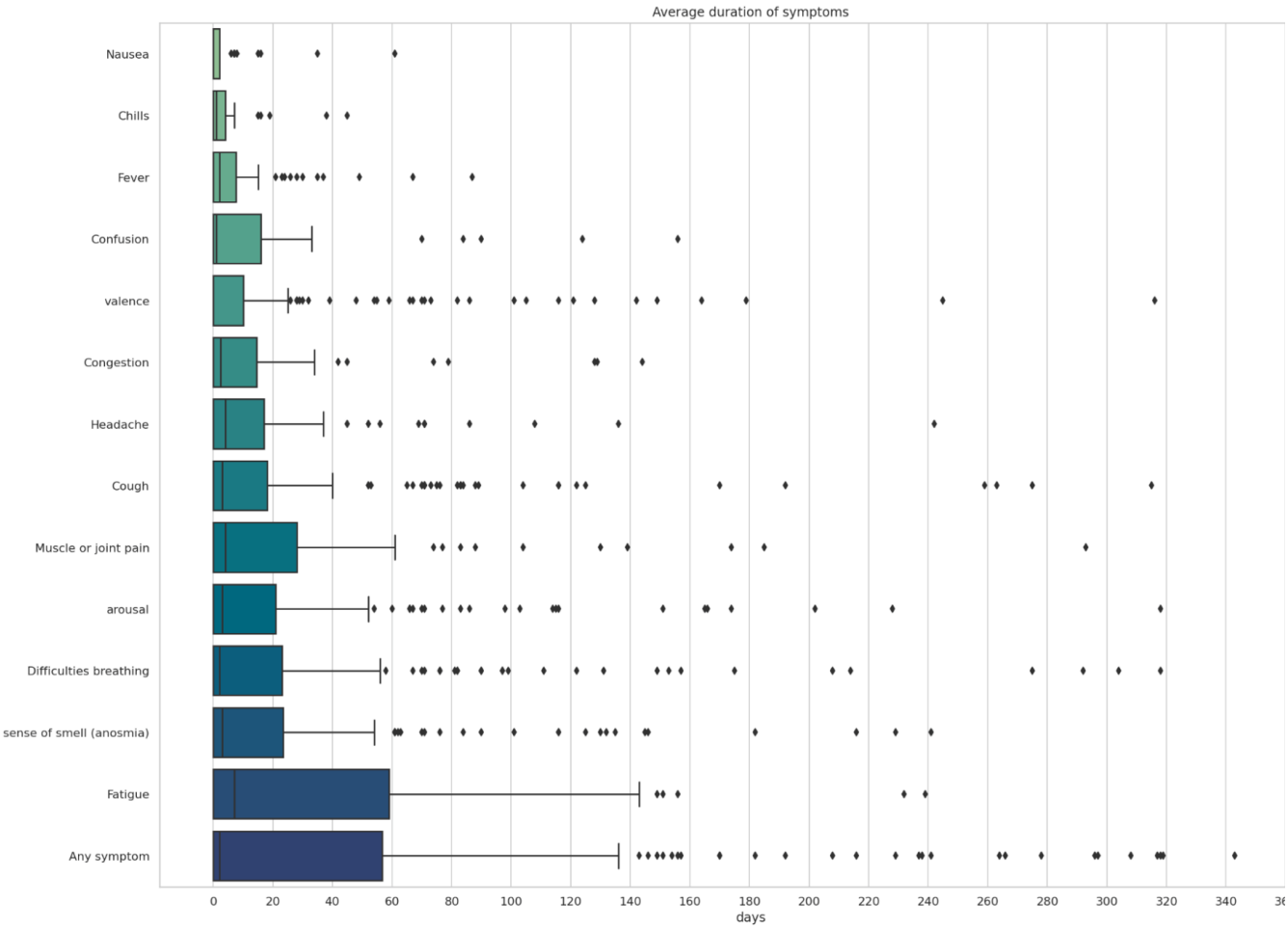


Figure 3: Self-reported symptom duration

A box plot representing average duration of symptoms (from the diagnosis date) for participants diagnosed with COVID-19. For clarity 13 most reported symptoms are included. The Any Symptom is an occurrence of any one of the symptoms. LS

Symptom stratification

As discussed in the methods section, the L_{symp} and S_{symp} groups were derived using symptom data, with L_{symp} participants defined as having reported persistent symptoms for at least 12 weeks. Of the 1327 total diagnosis reported 12.13% (161) reported persistent symptoms and were classified as L_{symp}, with the remaining participants (1104, 83.19%) classified as S_{symp}.

Stratification of the symptom-based cohorts over various socio-demographic factors was carried out (Table 5). There is a significant difference in age between the two cohorts, with the L_{symp} cohort associated with the older age group. The percentage of participants with certain comorbidities, including asthma, hypertension, diabetes and depression, was higher in the L_{symp} group, but not significantly so. A higher percentage of people in the L_{symp} group were employed in part-time roles or were retired while the S_{symp} cohort had a higher percentage of participants with full-time employment. L_{symp} also had a higher percentage of participants who were married and had children compared to S_{symp}. No differences can be seen in smoking between the two cohorts.

To visualise the differences between the S_{symp} and L_{symp} cohorts around diagnosis, various metrics were plotted as shown in Figure 4. Resting Heart Rate (RHR) was significantly higher in the L_{symp} cohort and stayed high for a longer period. Tachycardia and bradycardia are also more pronounced in the L_{symp} cohort. There appear to be some inherent differences in step count between $L_{symp}$and S_{symp} (<100 days before). Interestingly, the L_{symp} cohort has a greater step count and the difference in the rate of drop in steps approaching the diagnosis date, whereas the recovery of steps appear more similar. Sleep duration was higher in the S_{symp} cohorts over the whole time period but the difference in durations between L_{symp} and S_{symp} peaked close to the diagnosis date.

A series of Multiple Logistic Regressions were performed to establish the likelihood of having LCOVID based on the effects of sociodemographic, wearable, and mental health survey covariates. Each regression was on a single explanatory variable and adjusted for age, gender and ethnicity. The p-values were adjusted using the Benjamini-Hochberg correction for multiple testing.

Features for various independent variables were calculated based on the mean value over the year prior to diagnosis, the mean value during the acute phase (diagnosis + 14 days). Of the 1327 participants with a positive COVID diagnosis, 1213 [1060 S_{symp} and 153 L_{symp}] were included after exclusion of participants with missing data for the required variables. The inclusion criteria for continuous variables was based on a completion rate of at least 60% in the baseline and acute phases. More participants may have been excluded in different regressions because of missing data per variable.

The logistic regression results are given in Table 6. The significant Benjamini-Hochberg corrected p-values (p-value < 0.05) are marked with a * and in bold. Furthermore, a forest plot was generated to visualise the effects of the variables (Figure 5) with significant effects shown in orange. Age ranges were used to assess the effect of age on developing LCOVID and these were the most prominent

Variable	Short Covid (>3d)	Long Covid (>12w)
General demographics		
Gender [%F, %M]	[76.23, 23.03]	[75.16, 24.84]
Age (Mean, std, [IQR])	43.85 ± 12.52 [34.0 53.0]	49.51 ± 10.47 [41.3 57.8]
Bmi (Mean, std, [IQR])	27.26 ± 4.84 [23.5 30.9]	28.29 ± 5.48 [24.3 32.9]
Mental Health		
Depression (%)	19.87	25.66
Anxiety, nerves, or generalised anxiety disorder (%)	8.27	6.64
Physical Health		
Asthma (%)	17.01	23.45
Hypertension or high blood pressure (%)	6.04	8.41
Obesity (%)	3.66	3.98
Diabetes (Type 2) (%)	2.86	4.87
Cancer (%)	3.66	0.44
Employment		
Full time (%)	54.21	47.79
Part time (%)	12.08	19.03
Retired (%)	9.54	15.49
Self employed (%)	5.88	7.08
At home carer (%)	2.54	3.54
Unemployed (%)	2.23	2.21
Marriage and children		
Married (%)	53.42	61.06
Single (%)	13.67	17.7
Living with partner (%)	13.99	10.62
Living apart from partner (%)	3.34	3.1
Divorced (%)	3.82	2.21
Separated (%)	3.18	1.77
Children - y (%)	65.59	74.34
Children - n (%)	28.46	22.12
Smoking		
Smoker - never (%)	55.03	56.52
Smoker - ex (%)	30.04	31.06
Smoker - 1-10 (%)	2.93	2.48
Smoker - ecig (%)	3.02	2.48
Smoker - 11-20 (%)	1.46	1.86
Smoker - <1 (%)	1.74	0.62

Table 5: Sociodemographic stratification for S-LCOVID and S-SCOVID cohorts

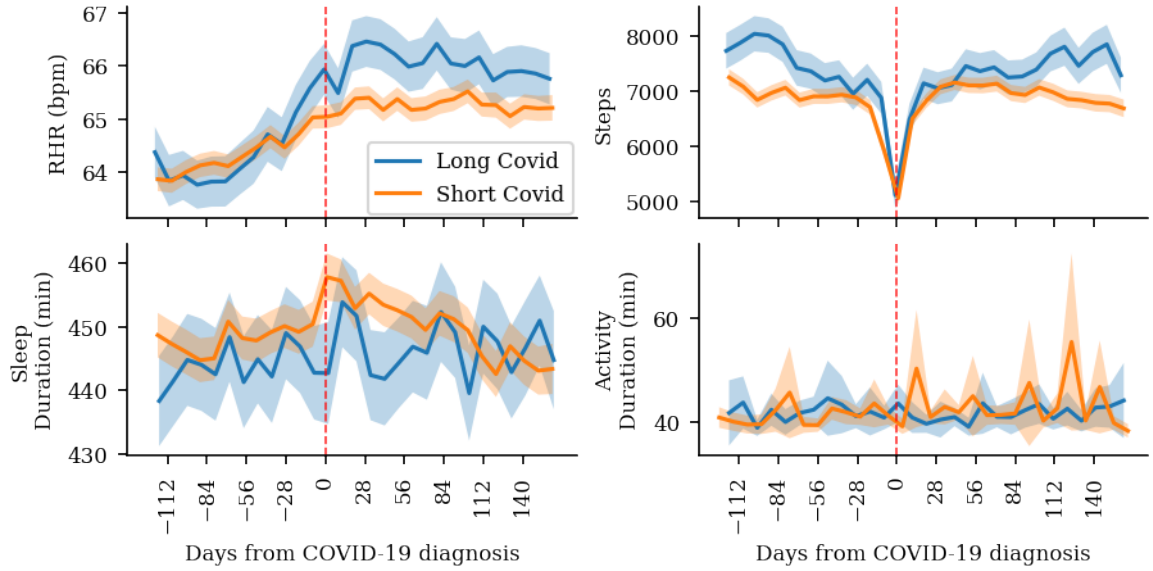


Figure 4: Passive metrics across symptom-based short and LCOVID cohorts

Comparison of passive and self-reported symptom measures for S_{symp} and L_{symp} cohorts, ranging from 16 weeks prior to 24 weeks post diagnosis of COVID-19. The shaded area corresponds to the 95% confidence interval taken over 10-day windows.

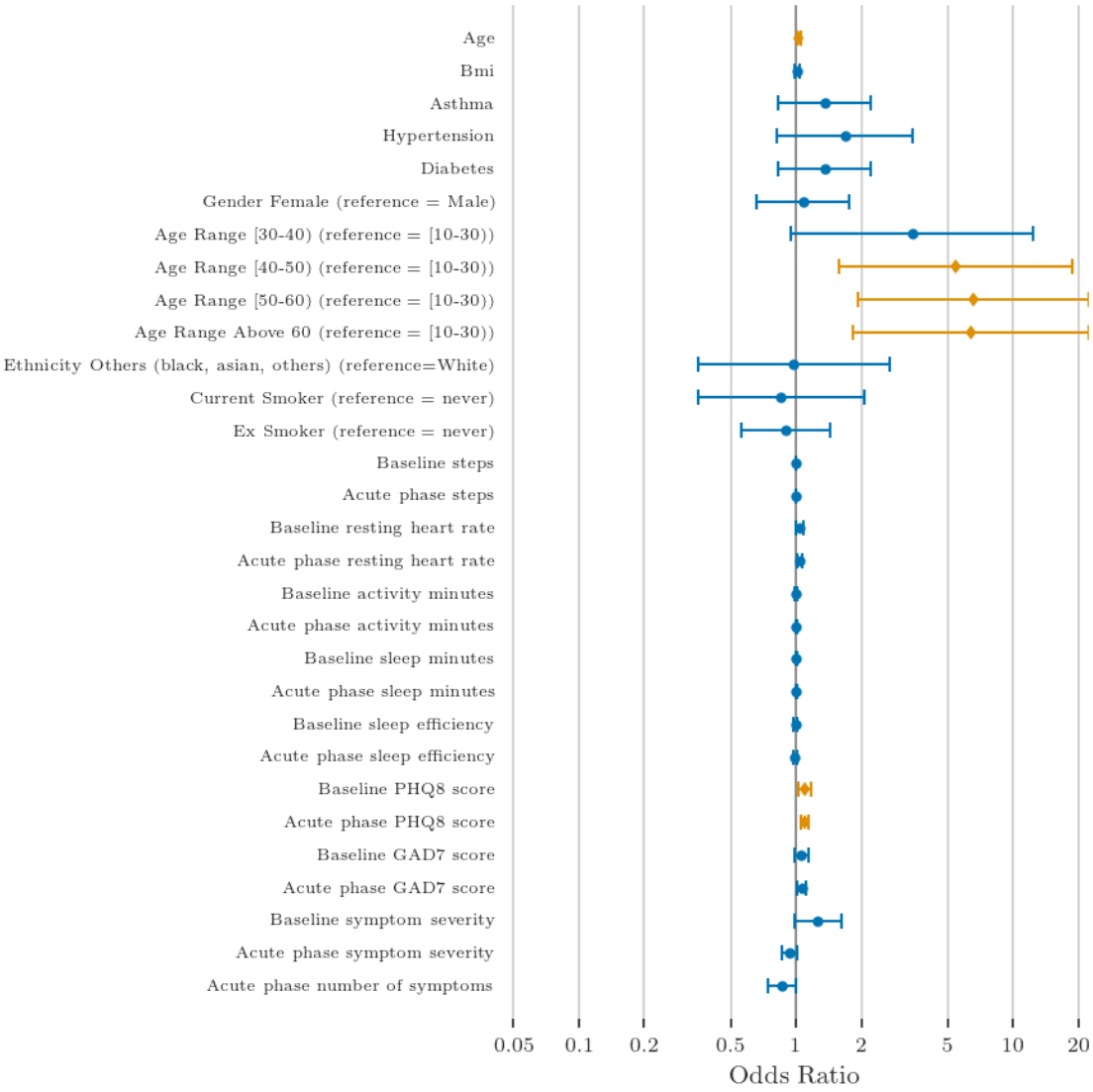


Figure 5: Logistic regression odds ratio for variables across symptom-based short and long COVID cohorts

Odds Ratio with 95% confidence interval, adjusted for age. The red colours show that the dependent variable has a significant effect (p-value < 0.05) on the independent variable.

Bmi	1.01509	0.47034	0.98541	1.04567
Asthma	1.35638	0.37109	0.83104	2.21382
Hypertension	1.68335	0.29906	0.82024	3.45471
Diabetes	1.35638	0.37109	0.83104	2.21382
Gender Female (reference = Male)	1.08099	0.8061	0.66169	1.766
Age Range [30-40) (reference = [10-30))	3.4188	0.16243	0.9475	12.33587
Age Range [40-50) (reference = [10-30))	5.43478	**0.04354***	1.57941	18.70117
Age Range [50-60) (reference = [10-30))	6.57051	**0.02487***	1.94017	22.25143
Age Range Above 60 (reference = [10-30))	6.41026	**0.02745***	1.83102	22.44179
Ethnicity Others (black, asian, others) (reference=White)	0.97608	0.96288	0.35207	2.70609
Current Smoker (reference = never)	0.85398	0.8061	0.35404	2.05987
Ex Smoker (reference = never)	0.89694	0.8061	0.56322	1.42839
Baseline steps	0.99999	0.8061	0.99993	1.00005
Acute phase steps	0.99997	0.47034	0.99991	1.00003
Baseline resting heart rate	1.03403	0.20396	0.99418	1.07547
Acute phase resting heart rate	1.03569	0.07045	1.00583	1.06644
Baseline activity minutes	0.99497	0.5177	0.98343	1.00665
Acute phase activity minutes	1.001	0.8061	0.99404	1.00801
Baseline sleep minutes	1.00386	0.15988	1.00003	1.00771
Acute phase sleep minutes	1.00165	0.47814	0.99818	1.00514
Baseline sleep efficiency	0.99664	0.8061	0.97753	1.01612
Acute phase sleep efficiency	0.98888	0.47034	0.96697	1.01129
Baseline PHQ8 score	1.09338	**0.04436***	1.02267	1.16898
Acute phase PHQ8 score	1.09489	**0.00301***	1.046	1.14606
Baseline GAD7 score	1.05617	0.29906	0.97874	1.13972
Acute phase GAD7 score	1.06283	0.05935	1.01249	1.11568
Baseline symptom severity	1.26064	0.16243	0.98573	1.61222
Acute phase symptom severity	0.93129	0.20396	0.85717	1.01181
Acute phase number of symptoms	0.86327	0.16243	0.74269	1.00342

The values in bold and marked with * show that the variable has a significant effect on the outcome of having long COVID.

Table 6: Multiple logistic regression for L_{symp}.

risk factors for LCOVID. Age Range (10-30) was used as the reference for other categories. All other age ranges had a significant effect with a rise in the odds ratio at each level, suggesting that the odds of developing LCOVID increases with an increase in age. The 50-60 age group was 6.5 times more likely and the oldest group (60+) was 6.4 more likely to have LCOVID than the reference group. Female gender (with reference as male) did not show a significant effect. Comorbidities, such as asthma, hypertension, and diabetes, did not show a significant effect (p-value < 0.05) on the presence of LCOVID, but the odds ratio were greater than one. Passive features and self-reported questionnaires were also considered as risk factors for LCOVID. Acute and Baseline phase PHQ8 scores had a significant effect.

Discussion

In this study we investigate persistent symptoms of and recovery from COVID-19 through the lens of mobile health data. We find a signal for LCOVID in both passive and active metrics, as well as associates to comorbidities, physiological metrics, and prior behaviour.

At a group-wide level, several wearable and mental health metrics are significantly changed during the acute COVID infection, some of which remain significantly different from the control group for longer than twelve weeks (Figure 1). Resting heart rate is the wearable metric with the longest lasting noticeable change. We estimate the proportion of participants with a long-term change to their heart rate coinciding with COVID-19 infection to be 7% on the basis of the BSTS model fit of heart rate data prior to infection, somewhat less than another study in which 13.4% of participants had a RHR of five bpm or more at twelve weeks [20]. Depression, anxiety, and self-rated arousal-valence remain negatively affected in the LCOVID phase.

As has previously been reported[20, 21], we observe a pattern of transient tachycardia from COVID-19 infection onset for a week, followed by a period of transient bradycardia from the second to third week, and finally a chronic or long-lasting increase in resting heart rate in some participants which can last several months or longer (Fig 4. This pattern of acute tachycardia and subsequent bradycardia is more prominent in the LCOVID cohort compared to the short COVID cohort, as shown in Figure 4.

The changing nature of the pandemic not only led to a changing understanding of what should be monitored in a study of this type, but importantly led to large societal and public health interventions. Many of those will also have had an effect on mental well-being and physical health. For instance, lockdown measures coincide with increased infection levels and also have an effect on activity[22, 23], sleep[24, 25], heart rate[25], and mental health[24, 26]. Therefore, when monitoring recovery in COVID-19 through mobile health, we should also consider the wider societal context. By time matching the control group we try to account for those changes, however, the impact of events concurrent to COVID-19 infection could be investigated in more detail.

An advantage of requesting existing wearable data from users of commercial fitness wearables is the ability to create a longitudinal dataset covering a period prior to enrolment, in some cases for many years. A higher level of historic physical activity is negatively correlated with development of the passive marker of LCOVID, the persistent increased RHR at twelve weeks post-diagnosis. To our knowledge, no other study considers the effect of historic activity or fitness level on LCOVID, but it has been demonstrated to reduce the severity and risk of hospitalisation in acute COVID[27]. Sleep duration was not significantly associated, but may be worth further investigation alongside other markers of historic health in larger datasets. Age was significantly positively associated, in agreement with multiple other studies[3, 28, 29] and the symptom-based findings in this study. Female sex was not significantly different to male sex, but did have a positive coefficient and a fairly low p-value. It seems likely that in a larger or

more powerful dataset, female sex would be a risk factor for LCOVID, in line with previously published research[3, 28–30].

Symptom Findings

The estimated prevalence of LCOVID in the literature is diverse. Our finding of the proportion of people in the L_{symp} group, after reporting a diagnosis and having persistent symptoms for 12 weeks (12.13%), is consistent with results from UK government Office for National Statistics (13.7%) in a study involving 20,000 participants[31]. However, studies have reported different prevalence rates for LCOVID at twelve weeks, from 2.6%[29] to 14.8% [32] and 37% [3]. Variance could be explained through sociodemographic differences across cohorts, methodological differences in the collection of symptom data, or how LCOVID is defined based on collected symptom data.

Our results show fatigue is the longest lasting symptom, with several participants experiencing fatigue for more than 140 days, which is consistent with previously published research[3, 29, 33, 34].

In agreement with previous studies[3, 28, 29] and the RHR-based regression, age was found to be a significant risk factor for LCOVID (L_{symp} cohort) with ages greater than 50 at very high risk. BMI (and obesity) was not a significant factor in our study. Other studies have shown that the female sex had a positive association with LCOVID[3, 28–30] but this was inconclusive in our study. A lack of power in this cohort reduces our ability to educe significance. Comorbidities like asthma, hypertension, and diabetes are potential risk factors for LCOVID, with non-significant p-values, which would agree with previous findings[28, 29]. This is not conclusive, as the L_{symp} cohort was defined through persistent symptoms which could also be caused by chronic illnesses, not necessarily COVID-19.

Investigation of passive metrics from wearables and self-reported questionnaires revealed that while a depression comorbidity was not significantly associated, the average PHQ8 score over the year prior to a COVID-19 diagnosis and during the acute phase of the disease was positively associated with LCOVID, indicating that a period of low mood before and during the disease could be a risk factor for LCOVID. Further resting heart rate in the acute phase of the disease also had a positive relation to LCOVID, with a low but non-significant p-value of 0.07, and could be a potential risk factor. The persistently increased RHR in the L_{symp} cohort, as visualised in Figure 4, also demonstrates a level of coherence between the two approaches to determining LCOVID.

Strengths and weaknesses

This study brings together COVID-19 self-reported symptoms, passive wearable data, and regular mental health surveys in a population who are not necessarily hospitalised. Both the symptom and passive approaches demonstrate that age is a risk factor for LCOVID. The availability of historic wearable data enabled the development of a long duration baseline with which to compare subsequent changes during COVID-19 infection. Increased historic activity, which suggests

a participant who had previously engaged in more exercise, is protective against the passive-based proxy for LCOVID.

The passive data approach uses existing data which is highly available among those who own wearable devices, is unbiased by subjective rating and identification of symptoms, and doesn't burden participants, but is limited in symptom scope to what a wearable device can measure. Meanwhile, the symptom self-report based methodology allows reporting of a wider range of symptoms than would be captured through wearable sensors and more concrete labels in the absence of a robust LCOVID classification algorithm for passive data, but relies on engaged and persistent participants.

There are multiple limitations to this study. Firstly, the definition of a LCOVID group on the basis of self-reported symptoms or by using resting heart rate as a proxy measure are weak approximations of a true diagnosis label. While we were able to show group-wide differences during the post-COVID period in self-reported mental health measures and passive wearable device biosignals, the use of the change in resting heart rate in the post-COVID period is non-specific and the effect of COVID-19 can be overwhelmed by natural variability in the individual case. Self-reported symptom monitoring requires time and commitment on the part of the person monitoring their COVID-19 recovery, which may be unrealistic to expect, especially given the increased prevalence of depression and fatigue. In both cases we assume a consistent deviation from a healthy baseline. However, symptoms of LCOVID may fluctuate or show signs of relapse and remission. Developing a model to identify LCOVID in mobile health data using another, labelled, dataset to then stratify COVID-19 recovery in datasets without explicit labelling may be an approach worth following, with a similar approach demonstrated recently in an electronic health records study of LCOVID[35].

Secondly, the nature of mobile health studies in general and of a community-sourced study, which relies on motivation and interest by participants, can lead to sporadic completion rates. Data completeness is reliant on what a participant is able and willing to share and on their continued engagement with the study. Meanwhile, the open remote enrolment paradigm biases the groups participating to those who have the studies published in a way that reaches them and that are self-motivated to take part. For example, the proportion of female participants is notably higher, a pattern that is seen across similar studies [23, 36]. This may be partially addressed through larger-scale studies or a meta-analysis including the similar studies that are running across various countries.

Conclusion

In conclusion, we demonstrate a measurable difference in measures of mental wellbeing and in biosignals from commercial wearable devices between COVID-19 positive and non-diagnosed participants during the sixteen weeks following diagnosis. Two methods of inferring the presence of LCOVID are compared. One method is based on persistent changes in resting heart rate and the other on persistent self-reported symptoms of COVID-19. For the self-reported symptoms

method, the LCOVID risk factors were explored using demographics, self-reports and passive wearable data, and compared with results from existing literature. In the future we plan to assess the feasibility of combining studies to create larger datasets or meta-analyses, to develop a LCOVID detection algorithm for use in mobile health data, and to investigate the additional effect of public health and safety measures.

References

1. H. Ritchie et al.: Coronavirus Pandemic (COVID-19). *Our World in Data* (2020). https://ourworldindata.org/coronavirus.
2. P. H. Roth and M. Gadebusch-Bondio: The contested meaning of long COVID Patients, doctors, and the politics of subjective evidence. *Social Science &: Medicine* **292** (Jan. 2022), 114619. doi: 10.1016/j.socscimed.2021.114619.
3. M. Whitaker et al.: Persistent COVID-19 symptoms in a community study of 606,434 people in England. *Nature Communications* **13**(1) (Apr. 2022). doi: 10.1038/s41467-022-29521-z.
4. D. Munblit et al.: Long COVID: aiming for a consensus. *The Lancet Respiratory Medicine* **10**(7) (July 2022), 632–634. doi: 10.1016/s2213-2600(22)00135-7.
5. *COVID-19 rapid guideline: managing the long-term effects of COVID-19*. Mar. 2022. url: https://www.nice.org.uk/guidance/ng188/resources/covid19-rapid-guideline-managing-the-longterm-effects-of-covid19-pdf-51035515742 (visited on 09/19/2022).
6. T. M. Schou, S. Joca, G. Wegener, and C. Bay-Richter: Psychiatric and neuropsychiatric sequelae of COVID-19 A systematic review. *Brain, Behavior, and Immunity* **97** (Oct. 2021), 328–348. doi: 10.1016/j.bbi.2021.07.018.
7. M. Gavriatopoulou et al.: Organ-specific manifestations of COVID-19 infection. *Clinical and Experimental Medicine* **20**(4) (July 2020), 493–506. doi: 10.1007/s10238-020-00648-x.
8. A. Mezlini et al.: Estimating the Burden of Influenza-like Illness on Daily Activity at the Population Scale Using Commercial Wearable Sensors. *JAMA Network Open* **5**(5) (May 2022), e2211958. doi: 10.1001/jamanetworkopen.2022.11958.
9. C. Stewart et al.: Investigating the Use of Digital Health Technology to Monitor COVID-19 and Its Effects: Protocol for an Observational Study (Covid Collab Study). *JMIR Research Protocols* **10**(12) (Dec. 2021), e32587. doi: 10.2196/32587.
10. K. Kroenke et al.: The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* **114**(1-3) (Apr. 2009), 163–173. doi: 10.1016/j.jad.2008.06.026.
11. R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe: A Brief Measure for Assessing Generalized Anxiety Disorder. *Archives of Internal Medicine* **166**(10) (May 2006), 1092. doi: 10.1001/archinte.166.10.1092.
12. J. A. Russell: A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6) (Dec. 1980), 1161–1178. doi: 10.1037/h0077714.

13. A. Russell, C. Heneghan, and S. Venkatraman: Investigation of an estimate of daily resting heart rate using a consumer wearable device (Oct. 2019). doi: 10.1101/19008771.

14. A. Natarajan, A. Pantelopoulos, H. Emir-Farinas, and P. Natarajan: Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study. *The Lancet Digital Health* **2**(12) (Dec. 2020), e650–e657. doi: 10.1016/s2589-7500(20)30246-6.

15. *Web API*. 2022. url: https://dev.fitbit.com/build/reference/web-api/ (visited on 09/19/2022).

16. S. Seabold and J. Perktold: Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the Python in Science Conference*. SciPy, 2010. doi: 10.25080/majora-92bf1922-011.

17. K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S. L. Scott: Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* **9**(1) (Mar. 2015). doi: 10.1214/14-aoas788.

18. E. Brunner and U. Munzel: The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal* **42**(1) (2000), 17–25. doi: https://doi.org/10.1002/(SICI)1521-4036(200001)42:1<17::AID-BIMJ17>3.0.CO;2-U.

19. Y. Benjamini and Y. Hochberg: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**(1) (Jan. 1995), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

20. J. M. Radin et al.: Assessment of Prolonged Physiological and Behavioral Changes Associated With COVID-19 Infection. *JAMA Network Open* **4**(7) (July 2021), e2115959. doi: 10.1001/jamanetworkopen.2021.15959.

21. A. Natarajan, H.-W. Su, and C. Heneghan: Occurrence of Relative Bradycardia and Relative Tachycardia in Individuals Diagnosed With COVID-19. *Frontiers in Physiology* **13** (May 2022). doi: 10.3389/fphys.2022.898251.

22. B. Constandt et al.: Exercising in Times of Lockdown: An Analysis of the Impact of COVID-19 on Levels and Patterns of Exercise among Adults in Belgium. *International Journal of Environmental Research and Public Health* **17**(11) (June 2020), 4144. doi: 10.3390/ijerph17114144.

23. S. Sun et al.: Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *Journal of Medical Internet Research* **22**(9) (Sept. 2020), e19992. doi: 10.2196/19992.

24. A. S. Kochhar et al.: Lockdown of 1.3 billion people in India during Covid-19 pandemic: A survey of its impact on mental health. *Asian Journal of Psychiatry* **54** (Dec. 2020), 102213. doi: 10.1016/j.ajp.2020.102213.

25. J. L. Ong, T. Lau, M. Karsikas, H. Kinnunen, and M. W. L. Chee: A longitudinal analysis of COVID-19 lockdown stringency on sleep and resting heart rate measures across 20 countries. *Scientific Reports* **11**(1) (July 2021). doi: 10.1038/s41598-021-93924-z.

26. K. F. Ahrens et al.: Differential impact of COVID-related lockdown on mental health in Germany. *World Psychiatry* **20**(1) (Jan. 2021), 140–141. doi: 10.1002/wps.20830.

27. J. P. Brandenburg, I. A. Lesser, C. J. Thomson, and L. V. Giles: Does Higher Self-Reported Cardiorespiratory Fitness Reduce the Odds of Hospitalization From COVID-19? *Journal of Physical Activity and Health* **18**(7) (July 2021), 782–788. doi: 10.1123/jpah.2020-0817.

28. A. Subramanian et al.: Symptoms and risk factors for long COVID in non-hospitalized adults. *Nature Medicine* **28**(8) (July 2022), 1706–1714. doi: 10.1038/s41591-022-01909-w.

29. C. H. Sudre et al.: Attributes and predictors of long COVID. *Nature Medicine* **27**(4) (Mar. 2021), 626–631. doi: 10.1038/s41591-021-01292-y.

30. R. A. Evans et al.: Clinical characteristics with inflammation profiling of long COVID and association with 1-year recovery following hospitalisation in the UK: a prospective observational study. *The Lancet Respiratory Medicine* **10**(8) (Aug. 2022), 761–775. doi: 10.1016/s2213-2600(22)00127-8.

31. *Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 1 April 2021*. Apr. 2021. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/1april2021 (visited on 09/19/2022).

32. E. T. Cirulli et al.: Long-term COVID-19 symptoms in a large unselected population (Oct. 2020). doi: 10.1101/2020.10.07.20208702.

33. H. E. Davis et al.: Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *eClinicalMedicine* **38** (Aug. 2021), 101019. doi: 10.1016/j.eclinm.2021.101019.

34. *Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 4 August 2022*. Aug. 2022. url: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/4august2022 (visited on 09/19/2022).

35. E. R. Pfaff et al.: Identifying who has long COVID in the USA: a machine learning approach using N3C data. *The Lancet Digital Health* **4**(7) (July 2022), e532–e541. doi: 10.1016/s2589-7500(22)00048-6.

36. S. Reade et al.: Cloudy with a Chance of Pain: Engagement and Subsequent Attrition of Daily Data Entry in a Smartphone Pilot Study Tracking Weather, Disease Severity, and Physical Activity in Patients With Rheumatoid Arthritis. *JMIR mHealth and uHealth* **5**(3) (Mar. 2017), e37. doi: 10.2196/mhealth.6496.

7.7 Summary

In this final analysis chapter I firstly investigated the presentation of post-COVID syndrome through wearable data and self-reported mood questionnaires and secondly looked at whether historic wearable metrics were predictive of developing long COVID according to a persistent change in resting heart rate. Group-wide differences between COVID-positive and negative participants showed significant negative effects in the infected group in wearable signals (increased heart rate, decreased sleep efficiency), mental health (increased PHQ-8 and GAD-7), and affective state (reduced valence and arousal) at 12 weeks. The primary novelty was in the use of historic wearable data, which allowed us to show that increased activity in the years prior to a COVID-19 infection was protective against developing post-COVID syndrome. We also noted a particular pattern of relative resting heart rate elevation, reduction, and subsequent elevation during the acute phase of the disease, which had also been reported in other studies.[15,16]

The findings of this study must be viewed through the biased engagement and adherence reported in chapter 5. Particularly when considering self-reported data, increased completion rates and higher consistency among older aged adults, increased enrolment rates of women, and the interplay between engagement and health status all potentially confound findings relating to those factors. In line with prior studies,[210] the self-report-based long COVID definition classified participants based on sustained symptom reports and subsequently looked for risk factors. However, putative risk factors such as age and sex are also related to consistency of participation, which will have a confounding effect. The passively-defined resting heart rate definition of long COVID may be less biased. However, not all cases of post-COVID syndrome will include cardiac symptoms or be reflected in heart rate.

Baseline recordings were used in a very different way to the meta-learning approach of chapter 4, but benefits are still present. Conditioning resting heart rate forecasting on the participant's long term data improved the heart rate prediction at 12-weeks and accounted for the effects of lockdown and seasonality. Similar work has taken a simple difference between mean heart rate directly prior to infection and at the later timepoint.[15] However, the group-wide effect sizes are reasonably small and could have been explained through seasonal heart rate trends given that COVID-19 infection rates also follow a seasonal pattern.

As will be discussed further in the following chapter, commercial data donation initiatives have drawbacks, but provide consistent monitoring and unique objective, if basic, historic measures of fitness and health, and is a source of information that would be impossible to derive outside unreliable retrospective self-reports.

Chapter 8

Discussion

Through this thesis I have tried to demonstrate the value of mHealth in scientific understanding, characterise common features of mHealth studies, and argue for the importance of contextual data in broad strokes across different parts of the mobile health study lifecycle. If nothing else, I hope to have argued for using all the data available to a problem and for a focus on generalisation.

8.1 Contributions

Can a multi-modal remote sensing system be used to detect focal seizures?

Chapter 3 discusses detection of epileptic seizures with a motor component. Although the dataset used was fairly small, good classification performance was demonstrated in generalised tonic-clonic seizures, in line with existing research, with variable performance across different participants and different focal seizures. The performance of the model trained across all participants in a leave-on-participant-out cross-validation was compared to a model trained on only a single participant in a leave-on-seizure-out paradigm. The personal model had a higher performance but was unviable for many participants because of there was an insufficient number of seizure recorded per person.

How can contextual baseline data be used to make more accurate predictions?

Motivated by a potential use-case in seizure detection, but somewhat limited by data availability, I wanted to assess the viability of applying a meta-learning approach to few-shot

learning with the goal of personalising classification models that use physiological signal data.

The desire for personalised models in seizure detection led to the consideration of few-shot learning and the evaluation of a neural process model in Chapter 4. Classification of stress was more accurate in a NP that had been provided with a labelled example of a participant's data from the beginning of the recording than traditional machine learning models. Looking slightly further ahead, it would be useful to build on the stress classification work and validate it in free-living conditions. There is a tendency to try and model complex medical and biological phenomena, such as depression, directly from low-level data. Building up intermediate-level features, such as recognising periods of stress and quantifying the degree of autonomic arousal, could be a more interpretable approach. Having a persistent long-term objective measure of stress could greatly aid in understanding the relationship and causality between disease and stress. This approach can be seen to a certain extent already, in areas like sleep classification.

A different use of baseline data is used in Chapter 7. Rather than conditioning a machine learning model, prior heart rate data is used to fit a structural time series model to forecast heart rate for use as a proxy severity measure of post-COVID syndrome. The strength of resting heart rate variance and seasonal trends differed between participants. When calculating the counterfactual heart rate at twelve weeks, the person-specific data is vital in ensuring any apparent effects are not due to those trends.

Can a citizen science study with 'opportunistic' historic wearable sensor data provide novel insights into COVID-19?

Chapter 7 takes a first look at long COVID in the Covid Collab study. At a group level, people who had an acute infection of COVID-19 were found to have a persistent significant change in physiological function and reduced mental wellbeing. Looking at long COVID through the change in resting heart rate following infection suggested that activity level prior to infection may be protective and demonstrated the utility of historic wearable data collected by commercial third-parties.

What are the implications of participant engagement on analysis in citizen science and mHealth studies?

Chapter 5 introduces the software underlying the Covid Collab study and describes the patterns of engagement and attrition of participants enrolled in it. Engagement patterns of

participant in Covid Collab varied. Attrition was high, in line with similar citizen science projects;[209] and there were relationships to sociodemographic characteristics, comorbidities, and potentially outcome measures. These should be considered when analysing data coming from Covid Collab and similar mHealth studies. Most participant engagement clustering in mHealth studies have focused on the total duration of engagement. I put forward a method of clustering based on fitting hidden Markov models to each participant's engagement sequence and using the probability of other participants' engagement under that model as a distance metric for clustering. It was able to distinguish groups, visually and by according to descriptive statistics of the clusters, according to length of engagement, consistency of engagement, and the quantity of completed tasks per week. The different ways in which a participant engages causes different patterns of data fragmentation and missingness. Several sociodemographic factors and mental health comorbidities were found to be significantly correlated with the rate of dropping out of the study, suggesting potential bias and confounding effects on the downstream analysis of citizen science project data.

8.2 Evaluation, limitations, and future direction

8.2.1 Analytical approaches

Given the breadth of this thesis' topic, a limited number of analytical approaches were evaluated. The multimodal seizure detection algorithm used a classic analysis pipeline. While I suggested the inclusion of euclidean angle-based features and the model performed competively;[252] it formed the argument for a personalisation approach that was then applied in another classification task. Given the highly variable symptoms of a seizure, it would be interesting to see whether a few-shot learning model would perform well.

The approach to model personalisation using neural processes in Chapter 4 provided a little confirmatory evidence for itself, but was very limited in scope. Only a NP with fully-connected layers was considered. However, more suitable architectures such as the sequential NP;[253] which can explicitly model dynamic stochastic processes, or the inclusion of recurrent[254] or attentive[255] units could improve performance. Indeed, NPs are not the most popular meta-learning method, and should also be compared against gradient-based methods.[256] However, NPs had the interesting property of being a meta-learner that approximated a stochastic process. The ability to include a level of uncertainty in predictions seemed attractive when it came to medical use-cases. Additionally, the update of context parameters at test-time only requires a computationally cheap forward pass of task-specific data through an encoder network, which may be advantageous in low-compute environments.

Looking forward, the preliminary work presented here should be built on. The meta-learning approach should be validated on a larger number of health and affective computing tasks, and a fuller comparison of meta-learning and few-shot learning algorithms could be undertaken, taking into account the recent advances in the field.[257]

Further afield, meta-learning is not the only deep learning approach to few-shot or domain adaption. Large pretrained *foundation models*[258] form the backbone for downstream tasks in speech,[259] natural language processing,[260] and computer vision,[261] and are increasingly used in health informatics applications within these domains.[262,263]

Work has been done on self-supervised learning and transfer learning with remote sensing data. Tang et al. produced a self-supervised model trained on Fenland study accelerometry data,[264] which has been followed up by Yuan et al.'s model trained on 7-day accelerometry recordings from 100,000 participants in UK Biobank.[265] Meanwhile, the multimodal nature of sensor data was exploited in a self-supervised contrastive learning approach[266] and including affective and sleep tasks, rather than just the activity recognition common to accelerometry models. However, it was trained on three relatively small datasets. A lack of large public multimodal remote sensing datasets may be hampering progress.

Meta-learning algorithms not often directly compared to fine-tuning on self-supervised models, but there is some overlap in the rationale for their use and recent papers have suggested fine-tuning out-performs meta-learning in few-shot image tasks while being easier to implement.[267,268] However, other studies have used the two approaches in combination outperform either method individually.[190,269] In the future, it may not be unreasonable to use a meta-learning personalisation approach on top of a large pretrained model.

8.2.2 COVID-19 and Covid Collab

The Covid Collab study was, to my knowledge, the largest UK wearables study monitoring COVID-19. It therefore provides a unique look at both passive wearable data and self-reported mental health during the pandemic.

Due to limited researchers on the project, there was very little ability to interact with participants or run engagement campaigns. As such, the participation and retention rates may have suffered. Retention was actually favourable compared to some prior citizen science studies,[270] with around 20% of participants remaining after one year, but this was likely boosted by increased sense of obligation to COVID-19 research among the public during the pandemic.

Only the surface of the data collected in the Covid Collab study has been scratched. Several Master's projects have considered symptomatology and classification in acute and long COVID using the dataset. As well as directly looking at illness, we will consider how

public health interventions, such as lockdowns, were reflected in mobile health and mental health data.

More generally it is also an mHealth collection with mental health outcomes over the past two years, with over 100,000 PHQ-8 and GAD-7 surveys and over 300,000 arousal-valence scores, and raw geolocation. One of the near-term goals for that data is to compare the performance of different meta-learning models in a prediction model for depressive symptoms using wearable device data.

Additionally, consent was sought to publish a public anonymised dataset from the Covid Collab study. The problem of community-level over-optimism in public datasets mentioned above notwithstanding, I hope it will be a useful resource for remote sensing health research.

The software supporting the project, such as the Mass Science app, will be made public, and has also been used in the Convalescence study.[194] The Convalescence study is a national long COVID project that combines clinical functional tests with a long-term wearable follow-up in a set of established population cohorts. The gold-standard testing and well characterised cohort will therefore be ideal to validate the exploratory findings from Covid Collab and other citizen science projects.

8.2.3 Mobile health datasets

Sample size was an issue across all studies. For example, Seizure detection was limited because only certain types of seizure were present in the dataset and only a few participants had multiple seizures, greatly reducing the ability to train individual or personalised models. Despite a fairly large cohort at first glance (n=17,500), the Covid Collab study was arguably under-powered when looking at the risk factors for long COVID, an issue caused in part by high rates of attrition and fragmented contribution of different types of data depending on participants. It may be possible to partially address data missingness through imputation, but given the potential correlation of engagement with disease state and mental health it may be problematic. A better approach could be to combine analyses with the similar studies that have been carried out across various countries.

There was a lack of long term monitoring in the physiological classification datasets. Without long term use in free-living conditions, it is hard to assess whether the models produced would generalise outside a clinic or research setting. It is very likely that increased activity and real-life stressors would increase motion artefacts, increase the false-positive rate, and have greater data missingness. The stress-evoking events used in WESAD and DriveDB can be expected to reliably activate the physiological stress response,[271] but they are in an extremely constrained environment.

Additionally, WESAD and DriveDB are relatively small and very popular affective computing datasets. As optimised hyperparameters selected through cross-validation can create over-optimistic performance predictions when compared to out-of-sample data, so to can large-scale collective evaluation from the research community lead to an over-optimistic bias in published results.[272] If stress classification is to be used as an intermediary step, it will be important to keep this in mind and validate potential algorithms in true out-of-sample data.

Bring-your-own-device citizen science and large longitudinal studies

One of the key findings in the long COVID study was the potential protective effect of higher rates of activity prior to diagnosis on the likelihood of developing persistent sequelae. It also demonstrates the potential utility of commercial repositories of historic wearable data that people are able to share with researchers. A lot of current mHealth research, even when including commercial fitness wearables, restrict the period of data that they collect to the duration of the study.[273–275] However, for some participants there are long-term objective records of the rate of exercise, sleep duration and quality, heart rate, and potentially participant-entered records of weight, height, or food consumption. This data can be used, as it was here, as an explanatory variable outcomes measured during a study, but there are other potential uses. Pulling in data collected prior to the study could be especially useful in training unsupervised anomaly detection models, providing contextual data to meta-learning models, or normalising features. It would also have a use in situations in which participants can provide post-hoc labels or integration with other data sources, such as electronic health records.

A bring-your-own-device study has a number of advantages. Costs can be significantly lower, but there are also potential data quality improvements. Participant compliance may be expected to be higher due to device familiarity and because it is already used in everyday life. That the device is their own may also limit the Hawthorne effect, the modification of behaviour due to the awareness of being watched.[276] However, there are limitations. Studies of this type, including Covid Collab, have all recruited highly skewed study populations. Wearable ownership rates and interest in contributing to research studies both vary across sociodemographic factors.[276] The reliance on wearable fitness device companies also creates a dependence on agents with a commercial interest and data that has gone through vendor-specific proprietary processing, which is often brought to market as a fitness rather than a medical device. Generalisation to new devices, future devices, or even differently-processed data may be compromised if a limited set of devices are used.

Adherence and retention are also very common problems. In an analysis of 100,000 participants across 8 studies, Pratap et al. show an 80% drop-off by the 40th day of the study.

[270] This was modified by disease state and study referral method, as well as age and sex as in Covid Collab. Certain methods of engagement, such as questionnaire timing and monetary incentives, were effective. However, no matter the approach to participant recruitment, engagement, and the design of self-reports, it is likely that a subset of people will have a greater drive to take part in scientific studies. Minimising the importance of adherence and self-reported data through linking donated wearable data to other sources, for instance health outcomes in electronic health records, may be a more constructive avenue.

There is a limit to what can be learnt and validated through the observational design that follows citizen science data donation studies. Going forward, it will be important to follow up on any exploratory findings with smaller randomised control studies with gold-standard outcomes. Additionally, citizen science projects, such as Covid Collab, and donated commercial wearable data could be a useful source of data for self-supervised learning. Keeping in mind the above concerns of dependence on a particular vendor, ideally as part of a larger, broader whole.

8.3 Conclusion

Mobile health and remote sensing have the opportunity to improve health outcomes through access, deeper insight, real-time intervention, but care must be taken for that potential to be realised. An increased focus on generalisable results and representative data, particularly reflecting real-world environments, would help the translation from research to application.

References

1. D. C. Mohr, M. Zhang, and S. M. Schueller: Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual Review of Clinical Psychology* **13**(1) (May 2017), 23–47. doi: 10.1146/annurev-clinpsy-032816-044949.
2. July 2023.
3. A. Zinzuwadia and J. P. Singh: Wearable devices—addressing bias and inequity. *The Lancet Digital Health* **4**(12) (Dec. 2022), e856–e857. doi: 10.1016/s2589-7500(22)00194-7.
4. B. Munos et al.: Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Annals of the New York Academy of Sciences* **1375**(1) (July 2016), 3–18. doi: 10.1111/nyas.13117.
5. J. Zhang, R. Chiodini, A. Badr, and G. Zhang: The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38**(3) (Mar. 2011), 95–109. doi: 10.1016/j.jgg.2011.02.003.
6. T. Quisel, L. Foschini, A. Signorini, and D. C. Kale: Collecting and Analyzing Millions of mHealth Data Streams. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2017. doi: 10.1145/3097983.3098201.
7. C. Menni et al.: Symptom prevalence, duration, and risk of hospital admission in individuals infected with SARS-CoV-2 during periods of omicron and delta variant dominance: a prospective observational study from the ZOE COVID Study. *The Lancet* **399**(10335) (Apr. 2022), 1618–1624. doi: 10.1016/s0140-6736(22)00327-0.
8. Y. Ranjan et al.: RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. *JMIR mHealth and uHealth* **7**(8) (2019), e11734.
9. M. Kozik, N. Isakadze, and S. S. Martin: Mobile health in preventive cardiology: current status and future perspective. *Current Opinion in Cardiology* **36**(5) (July 2021), 580–588. doi: 10.1097/hco.0000000000000891.
10. R. W. Treskes, E. T. van der Velde, R. Barendse, and N. Bruining: Mobile health in cardiology: a review of currently available medical apps and equipment for remote monitoring. *Expert Review of Medical Devices* **13**(9) (Aug. 2016), 823–830. doi: 10.1080/17434440.2016.1218277.
11. N. Singh et al.: Heart Rate Variability: An Old Metric with New Meaning in the Era of Using mHealth technologies for Health and Exercise Training Guidance. Part Two: Prognosis and Training. *Arrhythmia & Electrophysiology Review* **7**(4) (2018), 1. doi: 10.15420/aer.2018.30.2.
12. M. C. Ortega, E. Bruno, and M. P. Richardson: Electrodermal activity response during seizures: A systematic review and meta-analysis. *Epilepsy & Behavior* **134** (Sept. 2022), 108864. doi: 10.1016/j.yebeh.2022.108864.

13. P. N. Pfeiffer et al.: Mobile health monitoring to characterize depression symptom trajectories in primary care. *Journal of Affective Disorders* **174** (Mar. 2015), 281–286. doi: 10.1016/j.jad.2014.11.040.

14. W. G. Dixon et al.: How the weather affects the pain of citizen scientists using a smartphone app. *npj Digital Medicine* **2**(1) (Oct. 2019). doi: 10.1038/s41746-019-0180-3.

15. J. M. Radin et al.: Assessment of Prolonged Physiological and Behavioral Changes Associated With COVID-19 Infection. *JAMA Network Open* **4**(7) (July 2021), e2115959. doi: 10.1001/jamanetworkopen.2021.15959.

16. A. Natarajan, H.-W. Su, and C. Heneghan: Occurrence of Relative Bradycardia and Relative Tachycardia in Individuals Diagnosed With COVID-19. *Frontiers in Physiology* **13** (May 2022). doi: 10.3389/fphys.2022.898251.

17. A. M. De Livera, S. Zaloumis, and J. A. Simpson: Models for the Analysis of Repeated Continuous Outcome Measures in Clinical Trials. *Respirology* **19**(2) (Feb. 2014), 155–161. doi: 10.1111/resp.12217.

18. A. Coravos, S. Khozin, and K. D. Mandl: Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digital Medicine* **2**(1) (Mar. 2019). doi: 10.1038/s41746-019-0090-4.

19. C.-Y. Wu et al.: Reproducibility and Replicability of High-frequency, In-home Digital Biomarkers in Reducing Sample Sizes for Clinical Trials. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **7**(1) (Jan. 2021), e12220. doi: 10.1002/trc2.12220.

20. F. A. Curro et al.: Person-Centric Clinical Trials: Defining the N-of-1 Clinical Trial Utilizing a Practice-Based Translational Network. *Clinical Investigation* **5**(2) (Feb. 2015), 145–159. doi: 10.4155/cli.14.126.

21. K. Guk et al.: Evolution of Wearable Devices with Real-Time Disease Monitoring for Personalized Healthcare. *Nanomaterials* **9**(6) (May 2019), 813. doi: 10.3390/nano9060813.

22. A. Head et al.: Inequalities in Incident and Prevalent Multimorbidity in England, 2004–19: A Population-Based, Descriptive Study. *The Lancet Healthy Longevity* **2**(8) (Aug. 2021), e489–e497. doi: 10.1016/S2666-7568(21)00146-X.

23. J. A. Levine: The Application of Wearable Technologies to Improve Healthcare in the World's Poorest People. *Technology and Investment* **08**(02) (2017), 83–95. doi: 10.4236/ti.2017.82007.

24. S. Majumder, T. Mondal, and M. Deen: Wearable Sensors for Remote Health Monitoring. *Sensors* **17**(12) (Jan. 2017), 130. doi: 10.3390/s17010130.

25. S. C. Mathews et al.: Digital health: a path to validation. *npj Digital Medicine* **2**(1) (May 2019). doi: 10.1038/s41746-019-0111-3.

26. K. Kolasa and G. Kozinski: How to Value Digital Health Interventions? A Systematic Literature Review. *International Journal of Environmental Research and Public Health* **17**(6) (Mar. 2020), 2119. doi: 10.3390/ijerph17062119.

27. R. Syed et al.: Digital Health Data Quality Issues: Systematic Review. *Journal of Medical Internet Research* **25** (Mar. 2023), e42615. doi: 10.2196/42615.

28. A. Ghandeharioun et al.: Objective Assessment of Depressive Symptoms with Machine Learning and Wearable Sensors Data. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. San Antonio, TX: IEEE, Oct. 2017, 325–332. doi: 10.1109/ACII.2017.8273620.

29. H. Motahari-Nezhad et al.: Digital Biomarker Based Studies: Scoping Review of Systematic Reviews. *JMIR mHealth and uHealth* **10**(10) (Oct. 2022), e35722. doi: 10.2196/35722.
30. S. B. Goldberg, D. M. Bolt, and R. J. Davidson: Data Missing Not at Random in Mobile Health Research: Assessment of the Problem and a Case for Sensitivity Analyses. *Journal of Medical Internet Research* **23**(6) (June 2021), e26749. doi: 10.2196/26749.
31. S. Cho, I. Ensari, C. Weng, M. G. Kahn, and K. Natarajan: Factors Affecting the Quality of Person-Generated Wearable Device Data and Associated Challenges: Rapid Systematic Review. *JMIR mHealth and uHealth* **9**(3) (Mar. 2021), e20738. doi: 10.2196/20738.
32. B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn: Investigating Sources of Inaccuracy in Wearable Optical Heart Rate Sensors. *npj Digital Medicine* **3**(1) (Feb. 2020), 18. doi: 10.1038/s41746-020-0226-6.
33. E. Dogan, C. Sander, X. Wagner, U. Hegerl, and E. Kohls: Smartphone-Based Monitoring of Objective and Subjective Data in Affective Disorders: Where Are We and Where Are We Going? Systematic Review. *Journal of Medical Internet Research* **19**(7) (July 2017), e262. doi: 10.2196/jmir.7006.
34. N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara: Addressing Bias in Big Data and AI for Health Care: A Call for Open Science. *Patterns* **2**(10) (Oct. 2021), 100347. doi: 10.1016/j.patter.2021.100347.
35. C. R. Lesko et al.: Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology* **28**(4) (July 2017), 553–561. doi: 10.1097/EDE.0000000000000664.
36. L. A. Celi et al.: Sources of Bias in Artificial Intelligence That Perpetuate Healthcare Disparities-A Global Review. *PLOS Digital Health* **1**(3) (Mar. 2022). Ed. by H. S. Fraser, e0000022. doi: 10.1371/journal.pdig.0000022.
37. R. Challen et al.: Artificial Intelligence, Bias and Clinical Safety. *BMJ Quality & Safety* **28**(3) (Mar. 2019), 231–237. doi: 10.1136/bmjqs-2018-008370.
38. A.-F. Näher et al.: Secondary Data for Global Health Digitalisation. *The Lancet Digital Health* **5**(2) (Feb. 2023), e93–e101. doi: 10.1016/S2589-7500(22)00195-9.
39. A. Zinzuwadia and J. P. Singh: Wearable Devices—Addressing Bias and Inequity. *The Lancet Digital Health* **4**(12) (Dec. 2022), e856–e857. doi: 10.1016/S2589-7500(22)00194-7.
40. T.-W. Guu et al.: Wearable Devices: Underrepresentation in the Ageing Society. *The Lancet Digital Health* **5**(6) (June 2023), e336–e337. doi: 10.1016/S2589-7500(23)00069-9.
41. D. J. Stein et al.: What Is a Mental/Psychiatric Disorder? From DSM-IV to DSM-V. *Psychological Medicine* **40**(11) (Nov. 2010), 1759–1765. doi: 10.1017/S0033291709992261.
42. Y. Bengio, A. Courville, and P. Vincent: Representation Learning: A Review and New Perspectives (2012). doi: 10.48550/ARXIV.1206.5538.
43. Y. Song, T. Wang, S. K. Mondal, and J. P. Sahoo: A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities (2022). doi: 10.48550/ARXIV.2205.06743.
44. X. Liu et al.: Self-Supervised Learning: Generative or Contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. doi: 10.1109/TKDE.2021.3090866.
45. A. Adadi: A Survey on Data-efficient Algorithms in Big Data Era. *Journal of Big Data* **8**(1) (Jan. 2021), 24. doi: 10.1186/s40537-021-00419-9.
46. *Web API*. 2022. url: https://dev.fitbit.com/build/reference/web-api/ (visited on 09/19/2022).
47. *iOS - Health - Apple*. Oct. 2023. url: https://www.apple.com/ios/health/ (visited on 10/10/2023).

48. E. Garcia-Ceja et al.: Mental Health Monitoring with Multimodal Sensing and Machine Learning: A Survey. *Pervasive and Mobile Computing* **51** (Dec. 2018), 1–26. doi: 10.1016/j.pmcj.2018.09.003.

49. F. S. S. Leijten and the Dutch TeleEpilepsy Consortium: Multimodal Seizure Detection: A Review. *Epilepsia* **59**(S1) (June 2018), 42–47. doi: 10.1111/epi.14047.

50. H. Ma, W. Li, X. Zhang, S. Gao, and S. Lu: AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, 3109–3115. doi: 10.24963/ijcai.2019/431.

51. V. Radu et al.: Multimodal Deep Learning for Activity and Context Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1**(4) (Jan. 2018), 1–27. doi: 10.1145/3161174.

52. G. Muhammad et al.: A Comprehensive Survey on Multimodal Medical Signals Fusion for Smart Healthcare Systems. *Information Fusion* **76** (Dec. 2021), 355–375. doi: 10.1016/j.inffus.2021.06.007.

53. C. Sun, S. Hong, M. Song, and H. Li: A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data (2020). doi: 10.48550/ARXIV.2010.12493.

54. S. Deldari et al.: Latent Masking for Multimodal Self-supervised Learning in Health Timeseries (2023). doi: 10.48550/ARXIV.2307.16847.

55. E. Bruno et al.: Remote Assessment of Disease and Relapse in Epilepsy: Protocol for a Multicenter Prospective Cohort Study. *JMIR Research Protocols* **9**(12) (Dec. 2020), e21840. doi: 10.2196/21840.

56. M. Garnelo et al.: *Neural Processes*. 2018. doi: 10.48550/ARXIV.1807.01622.

57. R. Gordan, J. K. Gwathmey, and L.-H. Xie: Autonomic and endocrine control of cardiovascular function. *World Journal of Cardiology* **7**(4) (2015), 204. doi: 10.4330/wjc.v7.i4.204.

58. J. Tsao et al.: Heart rate variability as a biomarker for autonomic nervous system response differences between children with chronic pain and healthy control children. *Journal of Pain Research* (June 2013), 449. doi: 10.2147/jpr.s43849.

59. M. Hassani, A. F. Jouzdani, S. Motarjem, A. Ranjbar, and N. Khansari: How COVID-19 can cause autonomic dysfunctions and postural orthostatic syndrome? A Review of mechanisms and evidence. *Neurology and Clinical Neuroscience* **9**(6) (Oct. 2021), 434–442. doi: 10.1111/ncn3.12548.

60. M. Dani et al.: Autonomic dysfunction in 'long COVID': rationale, physiology and management strategies. *Clinical Medicine* **21**(1) (Nov. 2020), e63–e67. doi: 10.7861/clinmed.2020-0896.

61. R. D. Thijs, P. Ryvlin, and R. Surges: Autonomic manifestations of epilepsy: emerging pathways to sudden death? *Nature Reviews Neurology* **17**(12) (Oct. 2021), 774–788. doi: 10.1038/s41582-021-00574-w.

62. R. D. Thijs: The autonomic signatures of epilepsy: diagnostic clues and novel treatment avenues. *Clinical Autonomic Research* **29**(2) (Mar. 2019), 131–133. doi: 10.1007/s10286-019-00603-1.

63. E. Won and Y.-K. Kim: Stress, the Autonomic Nervous System, and the Immune-kynurenine Pathway in the Etiology of Depression. *Current Neuropharmacology* **14**(7) (Aug. 2016), 665–673. doi: 10.2174/1570159x14666151208113006.

64. *Primer on the Autonomic Nervous System*. Elsevier, 2012. doi: 10.1016/c2010-0-65186-8.

65. L. K. McCorry: Physiology of the Autonomic Nervous System. *American Journal of Pharmaceutical Education* **71**(4) (Sept. 2007), 78. doi: 10.5688/aj710478.

66. D. S. Goldstein: Dysautonomias: Clinical Disorders of the Autonomic Nervous System. *Annals of Internal Medicine* **137**(9) (Nov. 2002), 753. doi: 10.7326/0003-4819-137-9-200211050-00011.

67. J. Hadaya and J. L. Ardell: Autonomic Modulation for Cardiovascular Disease. *Frontiers in Physiology* **11** (Dec. 2020). doi: 10.3389/fphys.2020.617459.

68. M. Cella et al.: Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. *Schizophrenia Research* **195** (May 2018), 537–542. doi: 10.1016/j.schres.2017.09.028.

69. R. M. Carney, K. E. Freedland, and R. C. Veith: Depression, the Autonomic Nervous System, and Coronary Heart Disease. *Psychosomatic Medicine* **67** (May 2005), S29–S33. doi: 10.1097/01.psy.0000162254.61556.d5.

70. Y. Wang et al.: Altered cardiac autonomic nervous function in depression. *BMC Psychiatry* **13**(1) (July 2013). doi: 10.1186/1471-244x-13-187.

71. G. D. Femminella et al.: Autonomic Dysfunction in Alzheimer's Disease: Tools for Assessment and Review of the Literature. *Journal of Alzheimer's Disease* **42**(2) (Aug. 2014), 369–377. doi: 10.3233/jad-140513.

72. B. B. Wannamaker: Autonomic Nervous System and Epilepsy. *Epilepsia* **26**(s1) (June 1985), S31–S39. doi: 10.1111/j.1528-1157.1985.tb05722.x.

73. S. Vieluf et al.: Autonomic nervous system changes detected with peripheral sensors in the setting of epileptic seizures. *Scientific Reports* **10**(1) (July 2020). doi: 10.1038/s41598-020-68434-z.

74. C. L. Cooper and J. C. Quick, eds.: *The Handbook of Stress and Health*. John Wiley & Sons, Ltd, Apr. 2017. doi: 10.1002/9781118993811.

75. L. Levi: Occupational stress: Spice of life or kiss of death? *American Psychologist* **45**(10) (1990), 1142–1145. doi: 10.1037/0003-066x.45.10.1142.

76. B. Wang et al.: Wearable aptamer-field-effect transistor sensing system for noninvasive cortisol monitoring. *Science Advances* **8**(1) (Jan. 2022). doi: 10.1126/sciadv.abk0967.

77. O. Parlak: Portable and wearable real-time stress monitoring: A critical review. *Sensors and Actuators Reports* **3** (Nov. 2021), 100036. doi: 10.1016/j.snr.2021.100036.

78. P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven: Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ACM, Oct. 2018. doi: 10.1145/3242969.3242985.

79. J. Healey and R. Picard: Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems* **6**(2) (June 2005), 156–166. doi: 10.1109/tits.2005.848368.

80. F. R. Ihmig, A. Gogeascoechea, S. Schäfer, J. Lass-Hennemann, and T. Michael: *Electrocardiogram, skin conductance and respiration from spider-fearful individuals watching spider video clips*. 2020. doi: 10.13026/SQ6Q-ZG04.

81. J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani: A Non-EEG Biosignals Dataset for Assessment and Visualization of Neurological Status. *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, Oct. 2016. doi: 10.1109/sips.2016.27.

82. C. Espinosa-Garcia, H. Zeleke, and A. Rojas: Impact of Stress on Epilepsy: Focus on Neuroinflammation—A Mini Review. *International Journal of Molecular Sciences* **22**(8) (Apr. 2021), 4061. doi: 10.3390/ijms22084061.

83. C. Hammen: Stress and depression. *Annual Review of Clinical Psychology(2005)* **1**(1) (2005), 293–319.

84. N. M. H. GRAHAM, R. M. DOUGLAS, and P. RYAN: STRESS AND ACUTE RESPIRATORY INFECTION. *American Journal of Epidemiology* **124**(3) (Sept. 1986), 389–401. doi: 10.1093/oxfordjournals.aje.a114409.

85. S. L. Moshé, E. Perucca, P. Ryvlin, and T. Tomson: Epilepsy: new advances. *The Lancet* **385**(9971) (Mar. 2015), 884–898. doi: 10.1016/s0140-6736(14)60456-6.

86. R. S. Fisher et al.: Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology. *Epilepsia* **58**(4) (Mar. 2017), 522–530. doi: 10.1111/epi.13670.

87. O. Devinsky: Effects of Seizures on Autonomic and Cardiovascular Function. *Epilepsy Currents* **4**(2) (Feb. 2004), 43–46. doi: 10.1111/j.1535-7597.2004.42001.x.

88. J. M. van BUREN: SOME AUTONOMIC CONCOMITANTS OF ICTAL AUTOMATISM. *Brain* **81**(4) (1958), 505–528. doi: 10.1093/brain/81.4.505.

89. I. Hubbard, S. Beniczky, and P. Ryvlin: The Challenging Path to Developing a Mobile Health Device for Epilepsy: The Current Landscape and Where We Go From Here. *Frontiers in Neurology* **12** (Oct. 2021). doi: 10.3389/fneur.2021.740743.

90. T. Singhal: A Review of Coronavirus Disease-2019 (COVID-19). *The Indian Journal of Pediatrics* **87**(4) (Mar. 2020), 281–286. doi: 10.1007/s12098-020-03263-6.

91. H. Ritchie et al.: Coronavirus Pandemic (COVID-19). *Our World in Data* (2020). https://ourworldindata.org/coronavirus.

92. B. Hu, H. Guo, P. Zhou, and Z.-L. Shi: Characteristics of SARS-CoV-2 and COVID-19. *Nature Reviews Microbiology* **19**(3) (Oct. 2020), 141–154. doi: 10.1038/s41579-020-00459-7.

93. *Symptoms of COVID-19 | CDC*. Aug. 2022. url: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (visited on 09/29/2022).

94. E. Mahase: Covid-19: Sore throat, fatigue, and myalgia are more common with new UK variant. *BMJ* (Jan. 2021), n288. doi: 10.1136/bmj.n288.

95. M. Alene et al.: Magnitude of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLOS ONE* **16**(3) (Mar. 2021). Ed. by K. O. Kwok, e0249090. doi: 10.1371/journal.pone.0249090.

96. F. Callard and E. Perego: How and why patients made Long Covid. *Social Science & Medicine* **268** (Jan. 2021), 113426. doi: 10.1016/j.socscimed.2020.113426.

97. R. Pellegrino, E. Chiappini, A. Licari, L. Galli, and G. L. Marseglia: Prevalence and clinical presentation of long COVID in children: a systematic review. *European Journal of Pediatrics* (Sept. 2022). doi: 10.1007/s00431-022-04600-x.

98. E. Xu, Y. Xie, and Z. Al-Aly: Long-term neurologic outcomes of COVID-19. *Nature Medicine* (Sept. 2022). doi: 10.1038/s41591-022-02001-z.

99. N. W. Larsen, L. E. Stiles, and M. G. Miglis: Preparing for the long-haul: Autonomic complications of COVID-19. *Autonomic Neuroscience* **235** (Nov. 2021), 102841. doi: 10.1016/j.autneu.2021.102841.

100. A. Azizi et al.: Post-COVID-19 mental health and its associated factors at 3-months after discharge: A case-control study. *Clinical Epidemiology and Global Health* **17** (Sept. 2022), 101141. doi: 10.1016/j.cegh.2022.101141.

101. B. Becerra-Canales, H. M. Campos-Martínez, M. Campos-Sobrino, and G. A. Aquije-Cárdenas: Trastorno de estrés postraumático y calidad de vida del paciente post-COVID-19 en Atención Primaria. *Atención Primaria* **54**(10) (Oct. 2022), 102460. doi: 10.1016/j.aprim.2022.102460.

102. A. García-Molina et al.: Neuropsychological rehabilitation for post–COVID-19 syndrome: results of a clinical programme and six-month follow up. *Neurología (English Edition)* (Sept. 2022). doi: 10.1016/j.nrleng.2022.06.007.

103. E. Fraser: Long term respiratory complications of covid-19. *BMJ* (Aug. 2020), m3001. doi: 10.1136/bmj.m3001.

104. J. K. Hennigs et al.: Respiratory muscle dysfunction in long-COVID patients. *Infection* (May 2022). doi: 10.1007/s15010-022-01840-9.

105. B. A. Satterfield, D. L. Bhatt, and B. J. Gersh: Cardiac involvement in the long-term implications of COVID-19. *Nature Reviews Cardiology* **19**(5) (Oct. 2021), 332–341. doi: 10.1038/s41569-021-00631-3.

106. B. Siripanthong et al.: The Pathogenesis and Long-Term Consequences of COVID-19 Cardiac Injury. *JACC: Basic to Translational Science* **7**(3) (Mar. 2022), 294–308. doi: 10.1016/j.jacbts.2021.10.011.

107. H. E. Davis et al.: Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *eClinicalMedicine* **38** (Aug. 2021), 101019. doi: 10.1016/j.eclinm.2021.101019.

108. J. Allen: Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement* **28**(3) (Feb. 2007), R1–R39. doi: 10.1088/0967-3334/28/3/r01.

109. A. Samol et al.: Single-Lead ECG Recordings Including Einthoven and Wilson Leads by a Smartwatch: A New Era of Patient Directed Early ECG Differential Diagnosis of Cardiac Diseases? *Sensors* **19**(20) (Oct. 2019), 4377. doi: 10.3390/s19204377.

110. S. Jongstra et al.: Cognitive Testing in People at Increased Risk of Dementia Using a Smartphone App: The iVitality Proof-of-Principle Study. *JMIR mHealth and uHealth* **5**(5) (May 2017), e68. doi: 10.2196/mhealth.6939.

111. P. L. Enright: The six-minute walk test. *Respiratory care* **48**(8) (2003), 783–785.

112. N. Cummins et al.: Diagnosis of depression by behavioural signals. *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, Oct. 2013. doi: 10.1145/2512530.2512535.

113. K. Wu, D. Zhang, G. Lu, and Z. Guo: Joint learning for voice based disease detection. *Pattern Recognition* **87** (Mar. 2019), 130–139. doi: 10.1016/j.patcog.2018.09.013.

114. D. Rhonda J. Holmes Jennifer M. Oates: Voice characteristics in the progression of Parkinsons disease. *International Journal of Language & Communication Disorders* **35**(3) (Jan. 2000), 407–418. doi: 10.1080/136828200410654.

115. D. E. Webster et al.: The Mole Mapper Study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Scientific Data* **4**(1) (Feb. 2017). doi: 10.1038/sdata.2017.5.

116. L. Blom: mHealth for image-based diagnostics of acute burns in resource-poor settings: studies on the role of experts and the accuracy of their assessments. *Global Health Action* **13**(1) (Aug. 2020), 1802951. doi: 10.1080/16549716.2020.1802951.

117. T. Mazzu-Nascimento et al.: Mobile Health (mHealth) and Advances in Noninvasive Diagnosis of Anemia: An Overview. *International Journal of Nutrology* **13**(02) (Sept. 2020), 042–047. doi: 10.1055/s-0040-1716497.

118. *E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors*. 2022. url: https://www.empatica.com/en-gb/research/e4/ (visited on 09/24/2022).

119. A. L. Goldberger et al.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **101**(23) (2000 (June 13)). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215, e215–e220.

120. A. Miles et al.: *zarr-developers/zarr-python: v2.13.0*. 2022. doi: 10.5281/ZENODO.7104413.

121. J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel: When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology* **17**(1) (Dec. 2017). doi: 10.1186/s12874-017-0442-1.

122. J. A. C. Sterne et al.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**(jun29 1) (June 2009), b2393–b2393. doi: 10.1136/bmj.b2393.

123. J. L. Schafer: *Analysis of incomplete multivariate data*. CRC press, 1997. doi: 10.1201/9780367803025.

124. *pymhealth/pymhealth: A python package for mHealth data processing and feature extraction*. Mar. 2020. url: https://github.com/pymhealth/pymhealth (visited on 09/20/2022).

125. S. K. Lam, A. Pitrou, and S. Seibert: Numba. *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC - LLVM '15*. ACM Press, 2015. doi: 10.1145/2833157.2833162.

126. P. Virtanen et al.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17** (2020), 261–272. doi: 10.1038/s41592-019-0686-2.

127. P. S. Hamilton and W. J. Tompkins: Quantitative Investigation of QRS Detection Rules Using the MIT/BIH Arrhythmia Database. *IEEE Transactions on Biomedical Engineering* **BME-33**(12) (Dec. 1986), 1157–1165. doi: 10.1109/tbme.1986.325695.

128. J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe: A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* **49**(1) (2013), 1017–1034.

129. A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi: cvxEDA: a Convex Optimization Approach to Electrodermal Activity Processing. *IEEE Transactions on Biomedical Engineering* (2016), 1–1. doi: 10.1109/tbme.2015.2474131.

130. *Tilt sensing using Linear Accelerometers*. 2013. url: https://www.nxp.com/docs/en/application-note/AN3461.pdf (visited on 09/24/2022).

131. L. McManus, G. D. Vito, and M. M. Lowery: Analysis and Biophysics of Surface EMG for Physiotherapists and Kinesiologists: Toward a Common Language With Rehabilitation Engineers. *Frontiers in Neurology* **11** (Oct. 2020). doi: 10.3389/fneur.2020.576729.

132. J. D. Blood et al.: The variable heart: High frequency and very low frequency correlates of depressive symptoms in children and adolescents. *Journal of Affective Disorders* **186** (Nov. 2015), 119–126. doi: 10.1016/j.jad.2015.06.057.

133. B. Hjorth: EEG analysis based on time domain properties. *Electroencephalography and Clinical Neurophysiology* **29**(3) (Sept. 1970), 306–310. doi: 10.1016/0013-4694(70)90143-4.

134. A. Delgado-Bonal and A. Marshak: Approximate Entropy and Sample Entropy: A Comprehensive Tutorial. *Entropy* **21**(6) (May 2019), 541. doi: 10.3390/e21060541.

135. F. Shaffer and J. P. Ginsberg: An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health* **5** (Sept. 2017). doi: 10.3389/fpubh.2017.00258.

136. W. Boucsein: *Electrodermal Activity*. Springer US, 2012. doi: 10.1007/978-1-4614-1126-0.

137. E. Lutin, R. Hashimoto, W. D. Raedt, and C. V. Hoof: Feature Extraction for Stress Detection in Electrodermal Activity. *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications, 2021. doi: 10.5220/0010244601770185.

138. Y.-h. Sheu: Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry* **11** (Oct. 2020). doi: 10.3389/fpsyt.2020.551299.

139. S. Seabold and J. Perktold: statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*. 2010.

140. P. McCullagh and J. Nelder: *Generalized Linear Models*. Routledge, Jan. 2019. doi: 10.1201/9780203753736.

141. W. S. Noble: What is a support vector machine? *Nature Biotechnology* **24**(12) (Dec. 2006), 1565–1567. doi: 10.1038/nbt1206-1565.

142. F. Pedregosa et al.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12** (2011), 2825–2830.

143. C.-C. Chang and C.-J. Lin: LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2** (3 2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 27:1–27:27.

144. W.-Y. Loh: Classification and regression trees. *WIREs Data Mining and Knowledge Discovery* **1**(1) (Jan. 2011), 14–23. doi: 10.1002/widm.8.

145. L. Breiman: Random forests. *Machine learning* **45**(1) (2001), 5–32.

146. *Deep Learning Models for Medical Imaging*. Elsevier, 2022. doi: 10.1016/c2020-0-00344-0.

147. D. P. Kingma and M. Welling: *Auto-Encoding Variational Bayes*. 2013. doi: 10.48550/ARXIV.1312.6114.

148. C. E. Elger and C. Hoppe: Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. *The Lancet Neurology* **17**(3) (Mar. 2018), 279–288. doi: 10.1016/s1474-4422(18)30038-3.

149. B. Blachut et al.: Counting seizures: The primary outcome measure in epileptology from the patients' perspective. *Seizure* **29** (July 2015), 97–103. doi: 10.1016/j.seizure.2015.03.004.

150. A. Shah and S. Mittal: Invasive electroencephalography monitoring: Indications and presurgical planning. *Annals of Indian Academy of Neurology* **17**(5) (2014), 89. doi: 10.4103/0972-2327.128668.

151. C. Baumgartner and J. P. Koren: Seizure detection using scalp-EEG. *Epilepsia* **59** (June 2018), 14–22. doi: 10.1111/epi.14052.

152. A. V. de Vel et al.: Non-EEG seizure detection systems and potential SUDEP prevention: State of the art. *Seizure* **41** (Oct. 2016), 141–153. doi: 10.1016/j.seizure.2016.07.012.

153. C. H. Wong et al.: Risk factors for complications during intracranial electrode recording in presurgical evaluation of drug resistant partial epilepsy. *Acta Neurochirurgica* **151**(1) (Jan. 2009), 37–50. doi: 10.1007/s00701-008-0171-7.

154. R. Ortiz and J. Liporace: "Seizure-alert dogs": Observations from an inpatient video/EEG unit. *Epilepsy & Behavior* **6**(4) (June 2005), 620–622. doi: 10.1016/j.yebeh.2005.02.012.

155. *RADAR-CNS: Remote Assessment of Disease and Relapse – Central Nervous System | Radar-CNS*. 2022. url: https://www.radar-cns.org/ (visited on 09/25/2022).

156. R. Martinek et al.: Advanced Bioelectrical Signal Processing Methods: Past, Present, and Future Approach—Part III: Other Biosignals. *Sensors* **21**(18) (Sept. 2021), 6064. doi: 10.3390/s21186064.

157. A. Carbone, G. Castelli, and H. Stanley: Time-dependent Hurst exponent in financial time series. *Physica A: Statistical Mechanics and its Applications* **344**(1-2) (Dec. 2004), 267–271. doi: 10.1016/j.physa.2004.06.130.

158. B. H. Menze et al.: A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**(1) (July 2009). doi: 10.1186/1471-2105-10-213.

159. S. Beniczky, I. Conradsen, O. Henning, M. Fabricius, and P. Wolf: Automated real-time detection of tonic-clonic seizures using a wearable EMG device. *Neurology* **90**(5) (Jan. 2018), e428–e434. doi: 10.1212/wnl.0000000000004893.

160. K. Cuppens et al.: Accelerometry-Based Home Monitoring for Detection of Nocturnal Hypermotor Seizures Based on Novelty Detection. *IEEE Journal of Biomedical and Health Informatics* **18**(3) (May 2014), 1026–1033. doi: 10.1109/jbhi.2013.2285015.

161. P. Meritam, P. Ryvlin, and S. Beniczky: User-based evaluation of applicability and usability of a wearable accelerometer device for detecting bilateral tonic-clonic seizures: A field study. *Epilepsia* **59** (June 2018), 48–52. doi: 10.1111/epi.14051.

162. M. Nasseri et al.: Non-invasive wearable seizure detection using long–short-term memory networks with transfer learning. *Journal of Neural Engineering* **18**(5) (Apr. 2021), 056017. doi: 10.1088/1741-2552/abef8a.

163. F. Onorati et al.: Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. *Epilepsia* **58**(11) (Oct. 2017), 1870–1879. doi: 10.1111/epi.13899.

164. J. van Andel et al.: Multimodal, automated detection of nocturnal motor seizures at home: Is a reliable seizure detector feasible? *Epilepsia Open* **2**(4) (Sept. 2017), 424–431. doi: 10.1002/epi4.12076.

165. T. De Cooman, E. Carrette, P. Boon, A. Meurs, and S. Van Huffel: Online seizure detection in adults with temporal lobe epilepsy using single-lead ECG. *2014 22nd European Signal Processing Conference (EUSIPCO)*. 2014, 1532–1536.

166. A. V. de Vel et al.: Long-term accelerometry-triggered video monitoring and detection of tonic–clonic and clonic seizures in a home environment: Pilot study. *Epilepsy & Behavior Case Reports* **5** (2016), 66–71. doi: 10.1016/j.ebcr.2016.03.005.

167. M. Fawzy and H. Mostafa: High Accuracy Epileptic Seizure Detection System Based on Wearable Devices Using Support Vector Machine Classifier. *2021 International Conference on Microelectronics (ICM)*. IEEE, Dec. 2021. doi: 10.1109/icm52667.2021.9664898.

168. B. E. Heldberg et al.: Using wearable sensors for semiology-independent seizure detection - towards ambulatory monitoring of epilepsy. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2015. doi: 10.1109/embc.2015.7319660.

169. M. Mursalin, Y. Zhang, Y. Chen, and N. V. Chawla: Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. *Neurocomputing* **241** (June 2017), 204–214. doi: 10.1016/j.neucom.2017.02.053.

170. J. Jeppesen, S. Beniczky, P. Johansen, P. Sidenius, and A. Fuglsang-Frederiksen: Detection of epileptic seizures with a modified heart rate variability algorithm based on Lorenz plot. *Seizure* **24** (Jan. 2015), 1–7. doi: 10.1016/j.seizure.2014.11.004.

171. I. Conradsen, S. Beniczky, K. Hoppe, P. Wolf, and H. B. D. Sorensen: Automated Algorithm for Generalized Tonic–Clonic Epileptic Seizure Onset Detection Based on sEMG Zero-Crossing Rate. *IEEE Transactions on Biomedical Engineering* **59**(2) (Feb. 2012), 579–585. doi: 10.1109/tbme.2011.2178094.

172. H. Joo et al.: Spectral Analysis of Acceleration Data for Detection of Generalized Tonic-Clonic Seizures. *Sensors* **17**(3) (Feb. 2017), 481. doi: 10.3390/s17030481.

173. T. Trithipkaiwanpon and U. Taetragool: Sensitivity Analysis of Random Forest Hyperparameters. *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, May 2021. doi: 10.1109/ecti-con51831.2021.9454885.

174. B. F. Huang and P. C. Boutros: The parameter sensitivity of random forests. *BMC Bioinformatics* **17**(1) (Sept. 2016). doi: 10.1186/s12859-016-1228-x.

175. D. Cogan, J. Birjandtalab, M. Nourani, J. Harvey, and V. Nagaraddi: Multi-Biosignal Analysis for Epileptic Seizure Monitoring. *International Journal of Neural Systems* **27**(01) (Nov. 2016), 1650031. doi: 10.1142/s0129065716500313.

176. A. Mariotti: The Effects of Chronic Stress on Health: New Insights into the Molecular Mechanisms of Brain–Body Communication. *Future Science OA* **1**(3) (Nov. 2015), fso.15.21. doi: 10.4155/fso.15.21.

177. D. R. Garfin, R. R. Thompson, and E. A. Holman: Acute Stress and Subsequent Health Outcomes: A Systematic Review. *Journal of Psychosomatic Research* **112** (Sept. 2018), 107–113. doi: 10.1016/j.jpsychores.2018.05.017.

178. M. Helander: Applicability of Drivers' Electrodermal Response to the Design of the Traffic Environment. *Journal of Applied Psychology* **63**(4) (1978), 481–488. doi: 10.1037/0021-9010.63.4.481.

179. G. F. Wilson, J. D. Lambert, and C. A. Russell: Performance Enhancement with Real-Time Physiologically Controlled Adaptive Aiding. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **44**(13) (July 2000), 61–64. doi: 10.1177/154193120004401316.

180. G. F. Wilson: An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology* **12**(1) (Jan. 2002), 3–18. doi: 10.1207/S15327108IJAP1201_2.

181. J. Veltman and A. Gaillard: Physiological Indices of Workload in a Simulated Flight Task. *Biological Psychology* **42**(3) (Feb. 1996), 323–342. doi: 10.1016/0301-0511(95)05165-1.

182. G. Giannakakis et al.: Review on Psychological Stress Detection Using Biosignals. *IEEE Transactions on Affective Computing* **13**(1) (Jan. 2022), 440–460. doi: 10.1109/TAFFC.2019.2927337.

183. S. Gedam and S. Paul: A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques. *IEEE Access* **9** (2021), 84045–84066. doi: 10.1109/ACCESS.2021.3085502.

184. O. M. Mozos et al.: Stress Detection Using Wearable Physiological and Sociometric Sensors. *International Journal of Neural Systems* **27**(02) (Mar. 2017), 1650041. doi: 10.1142/S0129065716500416.

185. F. I. Indikawati and S. Winiarti: Stress Detection from Multimodal Wearable Sensor Data. *IOP Conference Series: Materials Science and Engineering* **771**(1) (Mar. 2020), 012028. doi: 10.1088/1757-899X/771/1/012028.

186. A. Saeed and S. Trajanovski: Personalized Driver Stress Detection with Multi-task Neural Networks Using Physiological Signals (2017). doi: 10.48550/ARXIV.1711.06116.

187. Y. S. Can et al.: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory Stress Detection With Smartwatches. *IEEE Access* **8** (2020), 38146–38163. doi: 10.1109/ACCESS.2020.2975351.

188. C. Finn, P. Abbeel, and S. Levine: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (2017). doi: 10.48550/ARXIV.1703.03400.

189. Y. Feng et al.: Meta-Learning as a Promising Approach for Few-Shot Cross-Domain Fault Diagnosis: Algorithms, Applications, and Prospects. *Knowledge-Based Systems* **235** (Jan. 2022), 107646. doi: 10.1016/j.knosys.2021.107646.

190. H.-y. Lee, S.-W. Li, and N. T. Vu: Meta Learning for Natural Language Processing: A Survey (2022). doi: 10.48550/ARXIV.2205.01500.

191. S. Jha, D. Gong, X. Wang, R. E. Turner, and L. Yao: The Neural Process Family: Survey, Applications and Perspectives (2022). doi: 10.48550/ARXIV.2209.00517.

192. *Homepage - Covid Collab*. 2020. url: https://covid-collab.org/ (visited on 08/20/2022).

193. *Flutter - Build apps for any screen*. 2022. url: https://flutter.dev/ (visited on 08/20/2022).

194. *COVID-19 Longitudinal Health and Wellbeing National Core Study*. Mar. 2022. url: https://www.ucl.ac.uk/covid-19-longitudinal-health-wellbeing/convalescence-long-covid-study (visited on 08/20/2022).

195. K. L. Druce, W. G. Dixon, and J. McBeth: Maximizing Engagement in Mobile Health Studies. *Rheumatic Disease Clinics of North America* **45**(2) (May 2019), 159–172. doi: 10.1016/j.rdc.2019.01.004.

196. C. N. Harrington, L. Ruzic, and J. A. Sanford: Universally Accessible mHealth Apps for Older Adults: Towards Increasing Adoption and Sustained Engagement. *Universal Access in Human–Computer Interaction. Human and Technological Environments*. Springer International Publishing, 2017, 3–12. doi: 10.1007/978-3-319-58700-4_1.

197. E. L. Murnane, D. Huffaker, and G. Kossinets: Mobile health apps. *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers - UbiComp 15*. ACM Press, 2015. doi: 10.1145/2800835.2800943.

198. M. R. Lunn et al.: Using mobile technology to engage sexual and gender minorities in clinical research. *PLOS ONE* **14**(5) (May 2019). Ed. by M. H. Withers, e0216282. doi: 10.1371/journal.pone.0216282.

199. S. Callier and S. M. Fullerton: Diversity and Inclusion in Unregulated mHealth Research: Addressing the Risks. *Journal of Law, Medicine & Ethics* **48**(S1) (2020), 115–121. doi: 10.1177/1073110520917036.

200. L. Maenhout et al.: Nonusage Attrition of Adolescents in an mHealth Promotion Intervention and the Role of Socioeconomic Status: Secondary Analysis of a 2-Arm Cluster-Controlled Trial. *JMIR mHealth and uHealth* **10**(5) (May 2022), e36404. doi: 10.2196/36404.

201. M. R. Hoque: An empirical study of mHealth adoption in a developing country: the moderating effect of gender concern. *BMC Medical Informatics and Decision Making* **16**(1) (May 2016). doi: 10.1186/s12911-016-0289-0.

202. F. R. T. van Elburg, N. S. Klaver, A. P. Nieboer, and M. Askari: Gender differences regarding intention to use mHealth applications in the Dutch elderly population: a cross-sectional study. *BMC Geriatrics* **22**(1) (May 2022). doi: 10.1186/s12877-022-03130-3.

203. S. Simblett et al.: Barriers to and Facilitators of Engagement With mHealth Technology for Remote Measurement and Management of Depression: Qualitative Analysis. *JMIR mHealth and uHealth* **7**(1) (Jan. 2019), e11325. doi: 10.2196/11325.

204. S. M. Gold et al.: Comorbid depression in medical diseases. *Nature Reviews Disease Primers* **6**(1) (Aug. 2020). doi: 10.1038/s41572-020-0200-2.

205. *ResearchKit*. 2022. url: https://researchkit.org/ (visited on 08/20/2022).

206. M. V. McConnell et al.: Feasibility of Obtaining Measures of Lifestyle From a Smartphone App. *JAMA Cardiology* **2**(1) (Jan. 2017), 67. doi: 10.1001/jamacardio.2016.4395.

207. Y.-F. Y. Chan et al.: The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. *Nature Biotechnology* **35**(4) (Mar. 2017), 354–362. doi: 10.1038/nbt.3826.

208. G. Quer et al.: Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine* **27**(1) (Oct. 2020), 73–77. doi: 10.1038/s41591-020-1123-x.

209. K. L. Druce et al.: Recruitment and Ongoing Engagement in a UK Smartphone Study Examining the Association Between Weather and Pain: Cohort Study. *JMIR mHealth and uHealth* **5**(11) (Nov. 2017), e168. doi: 10.2196/mhealth.8162.

210. C. H. Sudre et al.: Attributes and predictors of long COVID. *Nature Medicine* **27**(4) (Mar. 2021), 626–631. doi: 10.1038/s41591-021-01292-y.

211. A. L. Beatty et al.: The COVID-19 Citizen Science Study: Protocol for a Longitudinal Digital Health Cohort Study. *JMIR Research Protocols* **10**(8) (Aug. 2021), e28169. doi: 10.2196/28169.

212. J. Prince, S. Arora, and M. de Vos: Big data in Parkinson's disease: using smartphones to remotely detect longitudinal disease phenotypes. *Physiological Measurement* **39**(4) (Apr. 2018), 044005. doi: 10.1088/1361-6579/aab512.

213. D. A. Drew et al.: Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science* **368**(6497) (June 2020), 1362–1367. doi: 10.1126/science.abc0473.

214. Y. Wei et al.: Design Features for Improving Mobile Health Intervention User Engagement: Systematic Review and Thematic Analysis. *Journal of Medical Internet Research* **22**(12) (Dec. 2020), e21687. doi: 10.2196/21687.

215. S. Amagai, S. Pila, A. J. Kaat, C. J. Nowinski, and R. C. Gershon: Challenges in Participant Engagement and Retention Using Mobile Health Apps: Literature Review. *Journal of Medical Internet Research* **24**(4) (Apr. 2022), e35120. doi: 10.2196/35120.

216. J. McCambridge et al.: Impact of Length or Relevance of Questionnaires on Attrition in Online Trials: Randomized Controlled Trial. *Journal of Medical Internet Research* **13**(4) (Nov. 2011), e96. doi: 10.2196/jmir.1733.

217. A. C. Villanti et al.: Impact of Baseline Assessment Modality on Enrollment and Retention in a Facebook Smoking Cessation Study. *Journal of Medical Internet Research* **17**(7) (July 2015), e179. doi: 10.2196/jmir.4341.

218. A. S. Mustafa, N. Ali, J. S. Dhillon, G. Alkawsi, and Y. Baashar: User Engagement and Abandonment of mHealth: A Cross-Sectional Survey. *Healthcare* **10**(2) (Jan. 2022), 221. doi: 10.3390/healthcare10020221.

219. Z. Khadjesari et al.: Impact and Costs of Incentives to Reduce Attrition in Online Trials: Two Randomized Controlled Trials. *Journal of Medical Internet Research* **13**(1) (Mar. 2011), e26. doi: 10.2196/jmir.1523.

220. M. Mitchell et al.: Uptake of an Incentive-Based mHealth App: Process Evaluation of the Carrot Rewards App. *JMIR mHealth and uHealth* **5**(5) (May 2017), e70. doi: 10.2196/mhealth.7323.

221. N. Bidargaddi et al.: To Prompt or Not to Prompt? A Microrandomized Trial of Time-Varying Push Notifications to Increase Proximal Engagement With a Mobile Health App. *JMIR mHealth and uHealth* **6**(11) (Nov. 2018), e10123. doi: 10.2196/10123.

222. E. Kanjo, D. J. Kuss, and C. S. Ang: NotiMind: Utilizing Responses to Smart Phone Notifications as Affective Sensors. *IEEE Access* **5** (2017), 22023–22035. doi: 10.1109/access.2017.2755661.

223. A. Hampshire et al.: Cognitive deficits in people who have recovered from COVID-19. *EClinicalMedicine* **39** (Sept. 2021), 101044. doi: 10.1016/j.eclinm.2021.101044.

224. *Web API*. 2022. url: https://dev.fitbit.com/build/reference/web-api/ (visited on 09/25/2022).

225. *Health Api | Garmin Connect Developer Program | Garmin Developers*. 2022. url: https://developer.garmin.com/gc-developer-program/health-api/ (visited on 09/25/2022).

226. *Cloud Computing Services | Google Cloud*. Sept. 2022. url: https://cloud.google.com (visited on 09/10/2022).

227. *RADAR-base/RADAR-REST-Connector: A Kafka Source connector to receive data from REST APIs and publish them to Kafka. It has an extended version to support FitBit APIs*. Mar. 2021. url: https://github.com/RADAR-base/RADAR-REST-Connector (visited on 09/25/2022).

228. *Fitbit App*. 2020. url: https://www.fitbit.com/gb/app (visited on 09/25/2022).

229. J. P. Klein and M. L. Moeschberger: *Survival Analysis*. Springer New York, 2003. doi: 10.1007/b97377.

230. E. L. Kaplan and P. Meier: Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**(282) (June 1958), 457–481. doi: 10.1080/01621459.1958.10501452.

231. D. R. Cox: Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2) (Jan. 1972), 187–202. doi: 10.1111/j.2517-6161.1972.tb00899.x.

232. B. George, S. Seals, and I. Aban: Survival analysis and regression models. *Journal of Nuclear Cardiology* **21**(4) (May 2014), 686–694. doi: 10.1007/s12350-014-9908-2.

233. L. Rabiner: A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2) (1989), 257–286. doi: 10.1109/5.18626.

234. *hmmlearn 0.2.7.post20gd16c7c8 documentation*. Feb. 2022. url: https://hmmlearn.readthedocs.io (visited on 08/20/2022).

235. U. von Luxburg: A tutorial on spectral clustering. *Statistics and Computing* **17**(4) (Aug. 2007), 395–416. doi: 10.1007/s11222-007-9033-z.

236. L. Zelnik-Manor and P. Perona: Self-Tuning Spectral Clustering. *Proceedings of the 17th International Conference on Neural Information Processing Systems*. NIPS'04. Vancouver, British Columbia, Canada: MIT Press, 2004, 1601–1608.

237. P. Virtanen et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**(3) (Feb. 2020), 261–272. doi: 10.1038/s41592-019-0686-2.

238. M. Kerz: Biomedical applications in the age of mHealth. PhD thesis. King's College London, 2017.

239. R. Koch-Institut: *Corona-Datenspende | Robert Koch-Institut - Corona-Datenspende*. 2023. url: https://corona-datenspende.de (visited on 07/14/2023).

240. D. Brockmann, M. Wiedermann, R. W. Bruckmann, and A. Rose: *Our data donation roadmap for 2023: Corona Data Donation Project: News Analyses*. Feb. 2023. url: https://corona-datenspende.de/science/en/reports/kickoff-2023/ (visited on 07/14/2023).

241. *DETECT | Join the Study*. Dec. 2021. url: https://detect.scripps.edu/ (visited on 07/14/2023).

242. *Stanford COVID-19 Wearables Project — Stanford Healthcare Innovation Lab*. 2023. url: https://innovations.stanford.edu/wearables (visited on 07/14/2023).

243. *CovIdentify - A Duke University Study*. 2020. url: https://covidentify.covid19.duke.edu/ (visited on 07/14/2023).

244. A. E. Mason et al.: Detection of COVID-19 Using Multimodal Data from a Wearable Device: Results from the First TemPredict Study. *Scientific Reports* **12**(1) (Mar. 2022), 3463. doi: 10.1038/s41598-022-07314-0.

245. *ZOE Health Study*. July 2023. url: https://health-study.zoe.com/ (visited on 08/01/2023).

246. C. P. Adans-Dester et al.: Can mHealth Technology Help Mitigate the Effects of the COVID-19 Pandemic? *IEEE Open Journal of Engineering in Medicine and Biology* **1** (2020), 243–248. doi: 10.1109/OJEMB.2020.3015141.

247. M. Wiedermann et al.: Evidence for Positive Long- and Short-Term Effects of Vaccinations against COVID-19 in Wearable Sensor Metrics. *PNAS Nexus* **2**(7) (July 2023). Ed. by B. Levine, pgad223. doi: 10.1093/pnasnexus/pgad223.

248. M. Mekhael et al.: Studying the Effect of Long COVID-19 Infection on Sleep Quality Using Wearable Health Devices: Observational Study. *Journal of Medical Internet Research* **24**(7) (July 2022), e38000. doi: 10.2196/38000.

249. M. Woodrow et al.: Systematic Review of the Prevalence of Long COVID. *Open Forum Infectious Diseases* **10**(7) (July 2023), ofad233. doi: 10.1093/ofid/ofad233.

250. A. Natarajan et al.: A Systematic Review and Meta-Analysis of Long COVID Symptoms. *Systematic Reviews* **12**(1) (May 2023), 88. doi: 10.1186/s13643-023-02250-0.

251. F. Ceban et al.: Fatigue and Cognitive Impairment in Post-COVID-19 Syndrome: A Systematic Review and Meta-Analysis. *Brain, Behavior, and Immunity* **101** (Mar. 2022), 93–135. doi: 10.1016/j.bbi.2021.12.020.

252. F. Chen, I. Chen, M. Zafar, S. R. Sinha, and X. Hu: Seizures Detection Using Multimodal Signals: A Scoping Review. *Physiological Measurement* **43**(7) (July 2022), 07TR01. doi: 10.1088/1361-6579/ac7a8d.

253. G. Singh, J. Yoon, Y. Son, and S. Ahn: *Sequential Neural Processes*. 2019. doi: 10.48550/ARXIV.1906.10264.

254. S. Qin, J. Zhu, J. Qin, W. Wang, and D. Zhao: *Recurrent Attentive Neural Process for Sequential Data*. 2019. doi: 10.48550/ARXIV.1910.09323.

255. H. Kim et al.: *Attentive Neural Processes*. 2019. doi: 10.48550/ARXIV.1901.05761.

256. C. Finn, P. Abbeel, and S. Levine: Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks (2017). doi: 10.48550/ARXIV.1703.03400.

257. J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner: Fast and Flexible Multi-Task Classification Using Conditional Neural Adaptive Processes (2019). doi: 10.48550/ARXIV.1906.07697.

258. C. Zhou et al.: A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT (2023). doi: 10.48550/ARXIV.2302.09419.

259. A. Baevski, H. Zhou, A. Mohamed, and M. Auli: Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations (2020). doi: 10.48550/ARXIV.2006.11477.

260. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. doi: 10.48550/ARXIV.1810.04805.

261. K. He, X. Zhang, S. Ren, and J. Sun: *Deep Residual Learning for Image Recognition*. 2015. doi: 10.48550/ARXIV.1512.03385.

262. H. E. Kim et al.: Transfer Learning for Medical Image Classification: A Literature Review. *BMC Medical Imaging* **22**(1) (Dec. 2022), 69. doi: 10.1186/s12880-022-00793-7.

263. I. Li et al.: Neural Natural Language Processing for Unstructured Data in Electronic Health Records: A Review. *Computer Science Review* **46** (Nov. 2022), 100511. doi: 10.1016/j.cosrev.2022.100511.

264. C. I. Tang et al.: SelfHAR: Improving Human Activity Recognition through Self-training with Unlabeled Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **5**(1) (Mar. 2021), 1–30. doi: 10.1145/3448112.

265. H. Yuan et al.: Self-Supervised Learning for Human Activity Recognition Using 700,000 Person-days of Wearable Data (2022). doi: 10.48550/ARXIV.2206.02909.

266. S. Deldari et al.: Latent Masking for Multimodal Self-supervised Learning in Health Timeseries (2023). doi: 10.48550/ARXIV.2307.16847.

267. A. Shysheya, J. Bronskill, M. Patacchiola, S. Nowozin, and R. E. Turner: FiT: Parameter Efficient Few-shot Transfer Learning for Personalized and Federated Image Classification (2022). doi: 10.48550/ARXIV.2206.08671.

268. M. Patacchiola, M. Sun, K. Hofmann, and R. E. Turner: Comparing the Efficacy of Fine-Tuning and Meta-Learning for Few-Shot Policy Imitation (2023). doi: 10.48550/ARXIV.2306.13554.

269. S. X. Hu, D. Li, J. Stühmer, M. Kim, and T. M. Hospedales: Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference (2022). doi: 10.48550/ARXIV.2204.07305.

270. A. Pratap et al.: Indicators of Retention in Remote Digital Health Studies: A Cross-Study Evaluation of 100,000 Participants. *npj Digital Medicine* **3**(1) (Feb. 2020), 21. doi: 10.1038/s41746-020-0224-8.

271. A. P. Allen et al.: The Trier Social Stress Test: Principles and Practice. *Neurobiology of Stress* **6** (Feb. 2017), 113–126. doi: 10.1016/j.ynstr.2016.11.001.

272. A. F. Mendelson, M. A. Zuluaga, M. Lorenzi, B. F. Hutton, and S. Ourselin: Selection Bias in the Reported Performances of AD Classification Pipelines. *NeuroImage: Clinical* **14** (2017), 400–416. doi: 10.1016/j.nicl.2016.12.018.

273. F. Matcham et al.: Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC Psychiatry* **19**(1) (Feb. 2019). doi: 10.1186/s12888-019-2049-z.

274. S. Kitsiou et al.: Development of an innovative mHealth platform for remote physical activity monitoring and health coaching of cardiac rehabilitation patients. *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017. doi: 10.1109/bhi.2017.7897223.

275. S. Shin et al.: Activity monitoring using a mHealth device and correlations with psychopathology in patients with chronic schizophrenia. *Psychiatry Research* **246** (Dec. 2016), 712–718. doi: 10.1016/j.psychres.2016.10.059.

276. C. Demanuele et al.: Considerations for Conducting Bring Your Own "Device" (BYOD) Clinical Studies. *Digital Biomarkers* **6**(2) (July 2022), 47–60. doi: 10.1159/000525080.

277. K. Kroenke et al.: The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* **114**(1-3) (Apr. 2009), 163–173. doi: 10.1016/j.jad.2008.06.026.

278. R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe: A Brief Measure for Assessing Generalized Anxiety Disorder. *Archives of Internal Medicine* **166**(10) (May 2006), 1092. doi: 10.1001/archinte.166.10.1092.

279. M. Herdman et al.: Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research* **20**(10) (Apr. 2011), 1727–1736. doi: 10.1007/s11136-011-9903-x.

280. F. A. Klok et al.: The Post-COVID-19 Functional Status scale: a tool to measure functional status over time after COVID-19. *European Respiratory Journal* **56**(1) (May 2020), 2001494. doi: 10.1183/13993003.01494-2020.

Appendix A

Supporting Figures

	log HR	log HR SE	HR	t	P> |t|	[0.025	0.975]
sex[T.Female]	-0.0047	0.0178	0.9953	-0.2634	0.7923	0.9612	1.0306
age	-0.0201	0.0006	0.9801	-34.7562	0.0000	0.9790	0.9812
sex[T.Female]	-0.0336	0.0320	0.9669	-1.0494	0.2940	0.9081	1.0296
employment[T.oow]	-0.0083	0.0350	0.9917	-0.2381	0.8118	0.9260	1.0621
employment[T.student]	0.0098	0.2163	1.0098	0.0452	0.9639	0.6609	1.5430
age	-0.0285	0.0012	0.9719	-24.4073	0.0000	0.9697	0.9741
sex[T.Female]	-0.0444	0.0321	0.9565	-1.3848	0.1661	0.8982	1.0186
has_phys_comorbid	0.0380	0.0308	1.0387	1.2339	0.2173	0.9779	1.1034
has_mental_comorbid	0.0792	0.0309	1.0824	2.5672	0.0103	1.0189	1.1499
age	-0.0285	0.0011	0.9719	-25.8346	0.0000	0.9698	0.9740
sex[T.Female]	-0.0008	0.0286	0.9992	-0.0290	0.9768	0.9447	1.0568
age	-0.0226	0.0010	0.9777	-23.2883	0.0000	0.9758	0.9795
historic_sleep_st	-0.0370	0.0130	0.9637	-2.8544	0.0043	0.9395	0.9885
historic_heart_rate_st	-0.0167	0.0133	0.9834	-1.2553	0.2094	0.9581	1.0094
historic_activity_st	-0.0541	0.0139	0.9473	-3.9001	0.0001	0.9219	0.9734
sex[T.Female]	-0.0465	0.0322	0.9546	-1.4414	0.1495	0.8961	1.0169
depression[T.True]	0.0804	0.0360	1.0838	2.2327	0.0256	1.0099	1.1631
anxiety[T.True]	0.0432	0.0418	1.0441	1.0327	0.3017	0.9620	1.1333
age	-0.0281	0.0011	0.9723	-25.3302	0.0000	0.9702	0.9744
sex[T.Female]	-0.0084	0.0178	0.9917	-0.4705	0.6380	0.9577	1.0269
smoker[T.Ex]	0.0788	0.0185	1.0820	4.2648	0.0000	1.0435	1.1219
smoker[T.Current]	0.1798	0.0226	1.1970	7.9506	0.0000	1.1451	1.2512
age	-0.0204	0.0006	0.9798	-34.7679	0.0000	0.9787	0.9809
sex[T.Female]	-0.0323	0.0220	0.9682	-1.4692	0.1418	0.9274	1.0108
age	-0.0252	0.0007	0.9752	-34.1469	0.0000	0.9737	0.9766
bmi	0.0041	0.0015	1.0041	2.6815	0.0073	1.0011	1.0071

Table A.1 Proportional hazard regression results for duration of engagement

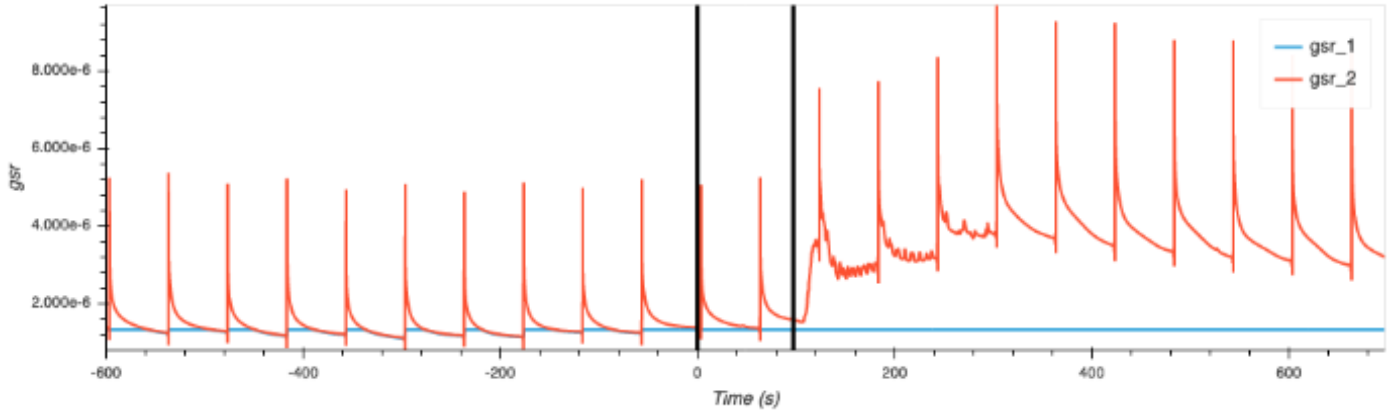


Fig. A.1 EDA polarity change artefact at a 1 minute frequency
The polarity change causes a large spike followed by a rapid and then gradual return to a baseline level. At the one-minute frequency changes to the tonic level are still visible. Some higher frequency behaviour is also visible between 150 and 300 seconds. At a one-hour frequency the artefact is much larger, both in amplitude and duration, causing much of the tonic level to be lost.

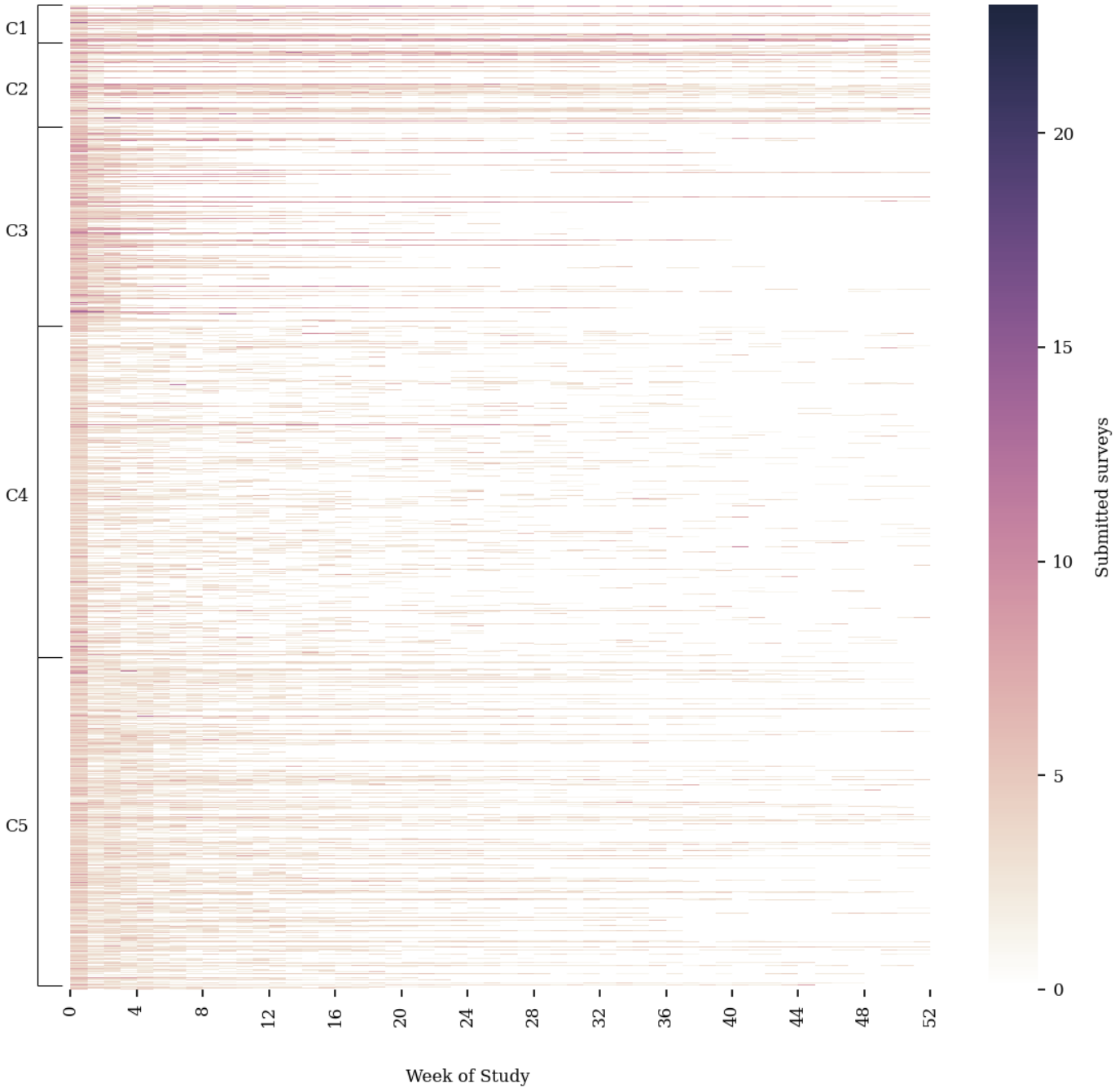


Fig. A.2 Participant subset clustered into 5 engagement groups

This figure contains the result of clustering the engagement sequences of all participants with more than one weeks worth of engagement in to 5 groups.

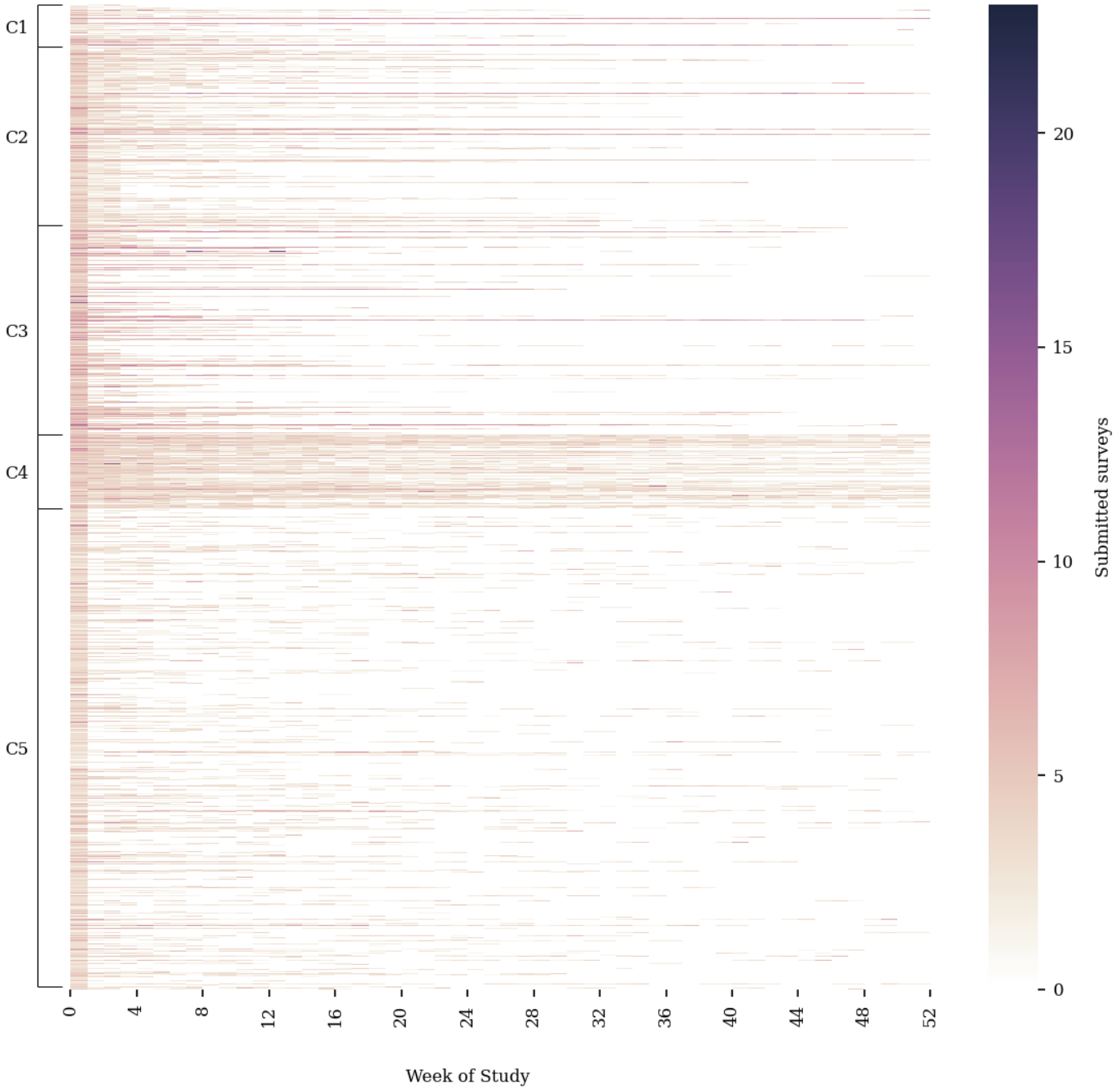


Fig. A.3 Participants clustered into 5 engagement groups

This figure contains the result of clustering the engagement sequences of all participants in to 5 groups, including those with data only during the first week of enrolment.

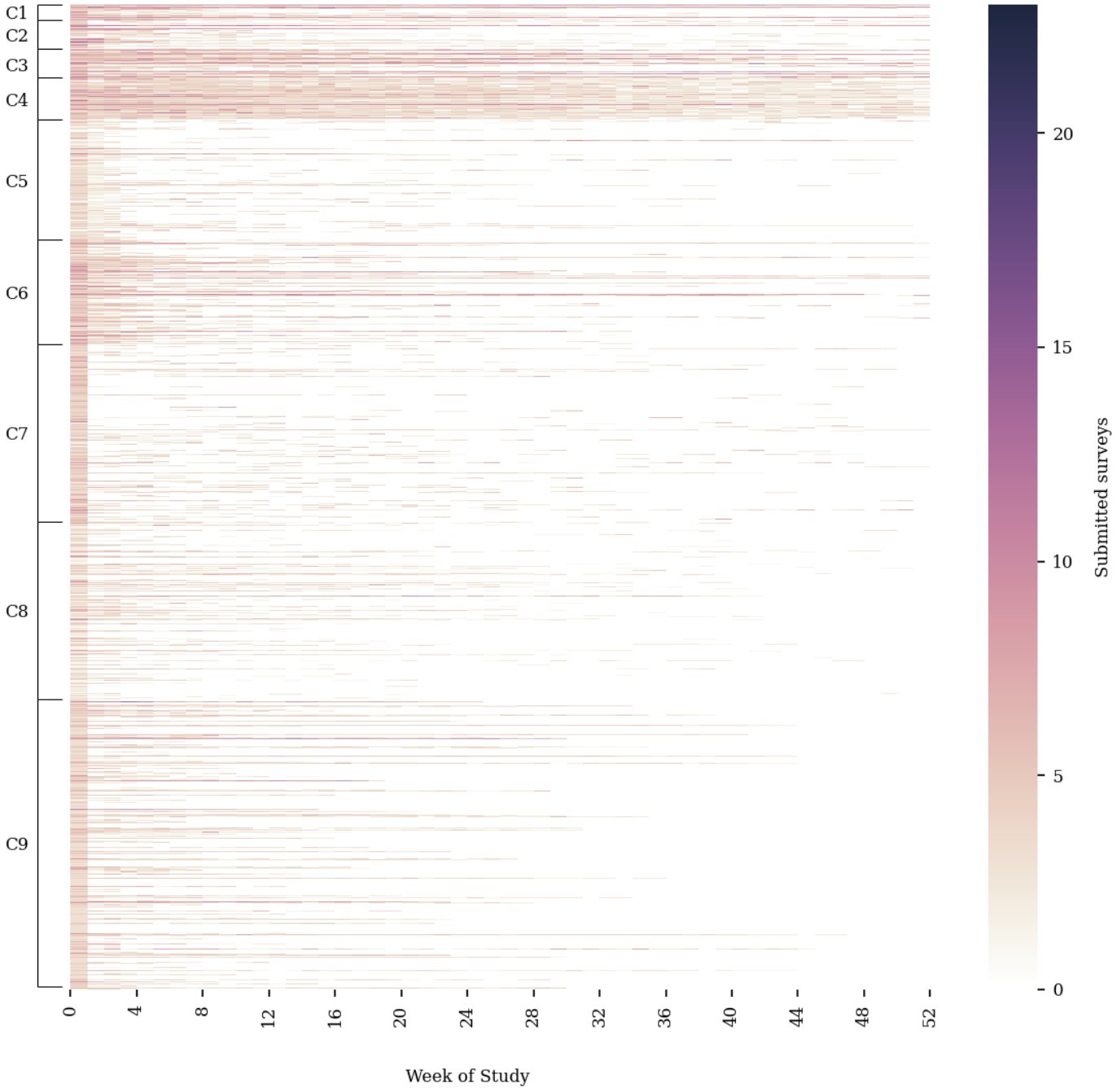


Fig. A.4 Participants clustered into 9 engagement groups

This figure contains the result of clustering the engagement sequences of all participants in to 9 groups, including those with data only during the first week of enrolment.

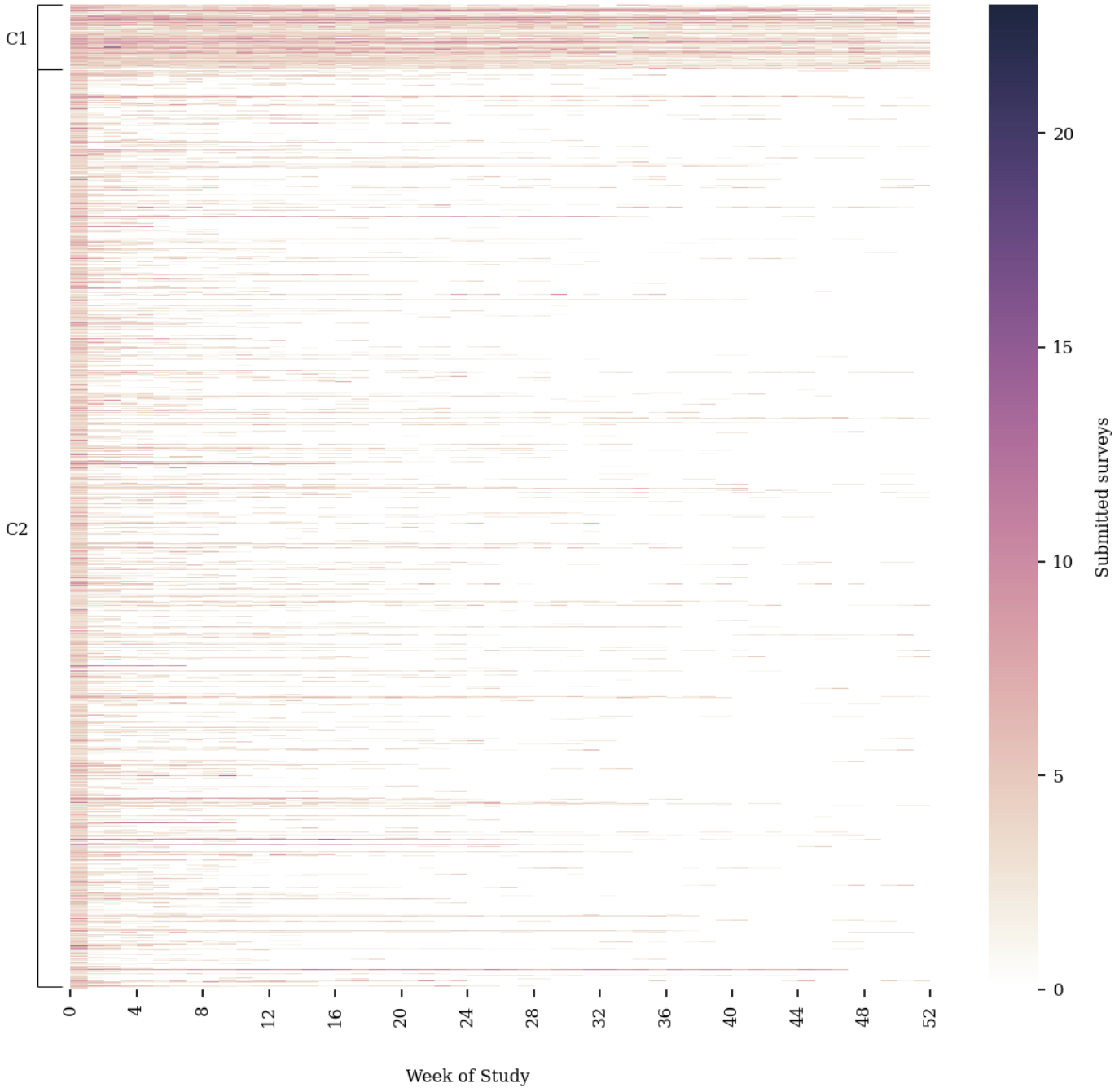


Fig. A.5 Participants clustered into 2 engagement groups

This figure contains the result of clustering the engagement sequences of all participants in to 2 groups, including those with data only during the first week of enrolment.

Appendix B

Mass Science Active Tasks

Sociodemographics

A sociodemographics questionnaire given after enrolment which builds on the initial enrolment form (which includes age, sex, gender, and smoking status).

Task name	Task type	Response	Task description
ethnicity	Listbuilder	[optionally multiple] ethnicity	A dropdown box with the option to choose multiple options and to enter arbitrary text. The default choices are 'Prefer not to say', 'Arab', 'Black', 'Central Asian', 'East Asian', 'South Asian', 'South-east Asian', 'White',

employment	Tickbox	'full_time', 'part_time', 'unemployed', 'zero_hours', 'at_home_carer', 'self_employed', 'freelance', 'small_business', 'state_income', 'retired', 'school', 'university'	Employment status of the participant
employment_change	Tickbox	'unchanged', 'furlough' ,'unemployed', 'hours_increase', 'hours_reduced', 'salary_increase', 'salary_reduced', 'benefits_increased', 'benefits_reduced', 'change_duty', 'change_job'	Whether the participant's employment status has changed since the pandemic began
marital_status	Dropdown	'single', 'separated', 'married', 'relationship_living', 'relationship_apart', 'divorced', 'widowed', 'other', 'pnts'	Marital status
children	Dropdown	'yes', 'no'	Whether the participant has children
living_situation	Tickbox	'alone', 'family', 'partner', 'family_partner', 'housemates', 'childrenlt18', 'childrengt18', 'nonnormal', 'different_country', 'other', 'pnts'	Where and with whom the participant lives

height_weight	Number entry	2 floats	The height and weight of the participant
health_physical	Listbuilder	List of physical comorbidities	Pre-existing physical health ailments. The participant can choose from a dropdown menu or enter their own.
health_mental	Listbuilder	List of mental comorbidities	Pre-existing mental health ailments. The participant can choose from a dropdown menu or enter their own.

Symptoms

A questionnaire for participants to submit symptoms of COVID-19, with separate questions for acute and long related symptoms. There is an arousal-valence scale attached as a more regular and shorter measure of mental wellbeing than the PHQ-8 or GAD-7 questionnaires.

Task name	Task type	Response	Task description
mood	2 Slider tasks	float -1 to 1	Two sliders correspond to 'happiness' and 'energy' (or valence and arousal).
symptoms	Listbuilder with severity	List of symptom severities	A list of acute COVID-19 related symptoms to be rated 'None', 'Mild', 'Moderate', or 'Severe'. By default the list includes 'Fever', 'Cough', 'Difficulties breathing', 'Loss of sense of smell (anosmia)' but the participant can add in arbitrary symptoms.
lcovid_symptoms	Listbuilder with severity	List of symptom severities	A list of Long COVID related symptoms to be rated 'None', 'Mild', 'Moderate', or 'Severe'. By default the list includes 'Fatigue', 'Difficulty thinking (brain fog)', 'Difficulty sleeping (insomnia)' but the participant can add in arbitrary symptoms.

Diagnosis

An ad-hoc questionnaire where participants can report the date that they fell ill or were diagnosed with COVID-19

Task name	Task type	Response	Task description
who	Radio task	'I have been diagnosed', 'A person I live with has been diagnosed'	Whether the participant or somebody they live with has been diagnosed with COVID-19
how	Radio task	'PCR', 'Antibody', 'Self-diagnosed or symptom based', 'Lateral flow test'	By what method the participant has been diagnosed
diagnosis_date	Date picker	Date	The date that the participant received the diagnosis
illness_date	Date picker	Date	The date that the participant first noticed symptoms or believe they fell ill

Vaccination

An ad-hoc questionnaire where participants can fill when they have received a vaccination for COVID-19

Task name	Task type	Response	Task description
vaccine_type	Dropdown	'pfizer', 'moderna', 'oxford', 'janssen' ,'unknown', 'other'	The type or producer of the vaccine
vaccine_dose	Radio task	'Initial dose', 'Booster shot'	Whether the vaccination was the initial dose or a booster
date_received	Datepicker	Date	The date the vaccine was received

6 Minute Walk Test

The Six Minute Walk Test is a standard walking exercise test used to measure fitness[111]. The protocol here is preceded by an information screen which includes a list of steps to be taken by the participant and a instructional video

Task name	Task type	Response	Task description
smwt_completed	Dropdown	'yes', 'no'	Task to determine whether the 6MWT was completed.
smwt_not_completed_reason	Dropdown	'forgot', 'fatigue', 'shortness_breath', 'too_busy', 'too_complicated', 'injury', 'poor_weather', 'other'	The reason for non-completion if *smwt_completed* is 'no'
smwt_datetime	Datepicker	Datetime	Date of the test if *smwt_completed* is 'yes'
smwt_completed_10	Dropdown	'yes', 'no'	Whether the participant was able to walk for 6 minutes
smwt_number_completed	Dropdown	'1', '2', '3', '4', '5'	Number of minutes completed if *smwt_completed_10* is 'no'
smwt_partial_complete_reason	Dropdown	'fatigue', 'shortness_breath', 'too_busy', 'too_complicated', 'injury', 'poor_weather', 'other'	The reason for partial completion if *smwt_completed_10* is 'no'
smwt_borg_scale	Dropdown	1 - 10	Rating of perceived exertion from 1-10 if *smwt_completed* is 'yes'

Chair Rises test

The Chair Rises test is an exercise test in which participants stand up and sit down on a chair for ten repetitions. The protocol here is preceded by an information screen which includes a list of steps to be taken by the participant and a instructional video

Task name	Task type	Response	Task description
chair_rises_completed	Dropdown	'yes', 'no'	A yes/no question to determine whether the Chair Rises test was completed.
chair_rises_not_completed_reason	Dropdown	'forgot', 'fatigue', 'shortness_breath', 'too_busy', 'too_complicated', 'injury', 'poor_weather', 'other'	The reason for non-completion if *chair_rises_completed* is 'no'
chair_rises_datetime	Datepicker	Datetime	Date of the test if *chair_rises_completed* is 'yes'
chair_rises_completed_10	Dropdown	'yes', 'no'	Whether the participant was able to walk for 6 minutes
chair_rises_number_completed	Dropdown	'1', '2', '3', '4', '5'	Number of minutes completed if *chair_rises_completed_10* is 'no'
chair_rises_partial_completion_reason	Dropdown	'fatigue', 'shortness_breath', 'too_busy', 'too_complicated', 'injury', 'poor_weather', 'other'	The reason for partial completion if *chair_rises_completed_10* is 'no'
chair_rises_borg_scale	Dropdown	1 - 10	Rating of perceived exertion from 1-10 if *chair_rises_completed* is 'yes'

PHQ-8

The PHQ-8 test is an 8-item questionnaire which measures symptoms of depression[277]. It is based on the PHQ-9 but without the final question on suicidal ideation. The eight questions ask the participant how many days they have noticed bothered by an indicator of depression, given in the description column below, over the last two weeks.

Task name	Task type	Response	Task description
phq8_1	Dropdown	'Not at all', 'Several days', 'More than half the days', 'Nearly every day'	Little interest or pleasure in doing things
phq8_2	Dropdown		Feeling depressed, or hopeless
phq8_3	Dropdown		Trouble falling asleep or staying asleep or sleeping too much
phq8_4	Dropdown		Feeling tired of having little energy
phq8_5	Dropdown		Poor appetite or over eating
phq8_6	Dropdown		Feeling bad about yourself
phq8_7	Dropdown		Trouble concentrating
phq8_8	Dropdown		Moving or speaking slowly - or the opposite

GAD-7

The GAD-7 test is a 7-item questionnaire which measures symptoms of generalised anxiety[278]. The seven questions ask the participant how many days they have noticed bothered by an indicator of anxiety, given in the description column below, over the last two weeks.

Task name	Task type	Response	Task description
gad7_1	Dropdown	'Not at all', 'Several days', 'More than half the days', 'Nearly every day'	Feeling nervous, anxious or on edge
gad7_2	Dropdown		Not being able to stop or control worrying
gad7_3	Dropdown		Worrying too much about different things
gad7_4	Dropdown		Having trouble relaxing
gad7_5	Dropdown		Being restless
gad7_6	Dropdown		Becoming annoyed or irritable
gad7_7	Dropdown		Feeling afraid

Quality of Life - EQ5D5L

A 5 item questionnaire on quality of life each wit h 5 levels of severity[279]. Each question asks the participant to describe an aspect of quality of life on the particular day they take the test.

Task name	Task type	Response	Task description
eq5d5l_1	Dropdown	'I have no problems in walking about', 'I have slight problems in walking about', 'I have moderate problems in walking about', 'I have severe problems in walking about', 'I am unable to walk about'	Mobility question
eq5d5l_2	Dropdown	'I have no problems washing or dressing myself', 'I have slight problems washing or dressing myself', 'I have moderate problems washing or dressing myself', 'I have severe problems washing or dressing myself', 'I am unable to wash or dress myself'	Self-care question
eq5d5l_3	Dropdown	'I have no problems doing my usual activities', 'I have slight problems doing my usual activities', 'I have moderate problems doing my usual activities', 'I have severe problems doing my usual activities', 'I am unable to do my usual activities'	Ability to complete normal activities
eq5d5l_4	Dropdown	'I have no pain or discomfort', 'I have slight pain or discomfort', 'I have moderate pain or discomfort', 'I have severe pain or discomfort', 'I have extreme pain or discomfort'	Pain severity
eq5d5l_5	Dropdown	'I am not anxious or depressed', 'I am slightly anxious or depressed', 'I am moderately anxious or depressed', 'I am severely anxious or depressed', 'I am extremely anxious or depressed'	Presence of anxiety or depression

ONS-2

A 2 item questionnaire in which participants rate life satisfaction from 0 to 10, where 0 is 'not at all' and 10 is 'completely'

Task name	Task type	Response	Task description
ons2_1	Slider task	integer 0-10	How satisfied the person is with their life
ons2_2	Slider task	integer 0-10	To what extent they feel their life is worthwhile

Post Covid Functional Scale

The Post Covid Functional Scale is a flowchart or questionnaire to help determine the taker's function and recovery following a COVID-19 infection[280]

Task name	Task type	Response	Task description
pcfs_1	Radio task	'Yes', 'No'	Whether the participant can live alone
pcfs_2	Radio task	'Yes', 'No'	Whether there are duties or activities the participant can no longer do. Asked conditional on *pcfs_1* = 'Yes'
pcfs_3	Radio task	'Yes', 'No'	Whether there are persistent symptoms, pain, depression, or anxiety Asked conditional on *pcfs_2* = 'No'
pcfs_4	Radio task	'Yes', 'No'	Whether it was necessary to reduce activities or duties. Asked conditional on *pcfs_3* = 'Yes'

Appendix C

Additional Published Content

C.1 RADAR-base: Major Depressive Disorder and Epilepsy Case Studies

RADAR-base: Major Depressive Disorder and Epilepsy Case Studies

Callum L Stewart
Institute of Psychiatry, Psychology & Neuroscience (IOPPN) King's College London
callum.stewart@kcl.ac.uk

Richard JB Dobson
Institute of Psychiatry, Psychology & Neuroscience (IOPPN) King's College London
richard.j.dobson@kcl.ac.uk

Zulqarnain Rashid
IOPPN, King's College London
zulqarnain.rashid@kcl.ac.uk

Amos A Folarin
IOPPN, King's College London
amos.folarin@kcl.ac.uk

Yatharth Ranjan
IOPPN, King's College London
yatharth.ranjan@kcl.ac.uk

The RADAR-CNS Consortium
https://www.radar-cns.org/

Shaoxiong Sun
IOPPN, King's College London
shaoxiong.sun@kcl.ac.uk

UbiComp/ISWC '18 Adjunct, October 8-12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10...$15.00
https://doi.org/10.1145/3267305.3267540

Abstract
Emerging mobile health (mHealth) and eHealth technology could provide opportunities for remote monitoring and interventions for people with mental health and neurological disorders. RADAR-base is a modern mHealth data collection platform built around Confluent and Apache Kafka. Here we report progress on studies into two brain disorders: major depressive disorder and epilepsy. For depression an ambulatory study is being conducted with patients recruited to three sites and for epilepsy an in-hospital study is being carried out at two sites. Initial results show smartphones and wearable devices have potential to improve care for patients with depression and epilepsy.

Author Keywords
mHealth; mobile context sensing; wearable sensors; data collection platform; mental health

ACM Classification Keywords
H.5.m [Human-centered computing (HCC)]: Ubiquitous and mobile computing.

Introduction
There has been an enormous increase in the capability to monitor individuals via smartphones and wearable devices during the last decade, with a growing range of parameters

offered by such technologies for continuous measurement [17].

The € 22 million Innovative Medicines Initiative (IMI2) Remote Assessment of Disease and Relapse - Central Nervous System (RADAR-CNS) is a major research programme aimed at developing novel methods and infrastructure for measuring Major Depressive Disorder (MDD), Epilepsy (EPI), and Multiple Sclerosis (MS) using wearable devices and smartphone technology [5].

The RADAR-base platform [1] is developed to support the three initial goals of RADAR-CNS, but importantly it has been developed such that it can easily be adapted for the needs of other mental and physical health disorders. The platform enables study design and set up, active and passive remote data collection. It provides secure data transmission and scalable solutions for data storage, management and access. This paper will focus on the MDD (RADAR-MDD) and EPI (RADAR-EPI) studies which use remote and in-hospital deployments of the RADAR-base platform respectively.

Major depressive disorder, sometimes called "clinical depression" or "depression", can be triggered by a life event, or result from stress, or happen without a specific cause. It is the most severe form of depression where people exhibit a sense of hopelessness and despair along with low mood and negative thoughts. This can affect the way people eat, sleep, feel about themselves, and think about things. Without treatment, the symptoms can last for weeks, months, or even years.

The RADAR-base platform has been deployed centrally to collect active (questionnaires) and passively generated (wearable and smartphone sensor) data remotely for patients recruited to 3 sites of MDD study. The sites include King's College Hospital (KCH) London, Centro de Investigacion Biomedica en Red (CIBER) Barcelona and VU University Medical Center Netherland. The objective being to collect regular self reported symptoms and metrics such as sleep and ambulatory behaviour. High resolution data is being collected over a period of up to two years for each participant.

Epilepsy is a neurological condition characterised by a person's tendency to have epileptic seizures. The global prevalence of epilepsy is between 4-10 per 1000 people. Those with epilepsy have a reduced life expectancy; people with symptomatic epilepsy have a life expectancy 18 years shorter [6]. Our hypothesis is that consumer type wearable devices have the potential to provide continuous seizure detection which may enable more informed use of anti-epileptic drugs, generating a more objective view of a person's condition.

Though current hospital observational systems (Video/ EEG/ ECG) are used in home monitoring they are not practical for long term epilepsy seizure detection within home based settings. We are using the RADAR-base platform to explore the feasibility of three wearable devices to detect seizures in an ambulatory settings. Data is being collected for a maximum of 14 days per patient.

These two studies expose the versatility of the RADAR-base platform and generate data with very different complexity, volume, velocity and durations.

Related Work

A number of relevant studies and mHealth platforms for remote monitoring in mental health are discussed here [17].

HORYZONS is a web based interface and feedback system to study people with first episode psychosis (FEP), a

1736

one month pilot study with 20 participants was conducted [2]. The study aim was to provide an Internet-based intervention to young people with psychosis, to provide cost-effective long-term treatment to sustain the benefits of early intervention. The majority (75%) reported that they had a positive and constructive experience using the system, however this was a short term pilot with limited participants focused on young population with FEP.

Another study involves the naturalistic follow up of responders from the study entitled "Integrated biological markers for the prediction of treatment response in depression", or the **CBN-WeII** study. In this study, participants who are currently responding to an oral antidepressant treatment regimen and/or therapeutic intervention were monitored over a minimum period of 13 months, providing an important opportunity to discover near-term biomarkers of relapse [12].

OBSERVEMDD a prospective, multicenter, longitudinal, single-cohort, observational study with MDD participants was performed using accelerometers and smartphone delivered questionnaires [18]. MDD patients who responded to, and continue to respond to an oral antidepressant treatment regimen were selected. The study consisted of 2 parts: a screening phase of up to 2 weeks, and an observational phase of variable duration. A total of 350 participants were recruited.

The RADAR-CNS programme advances the field in a number of ways. Other studies to date have made little or limited use of multi-parametric remote monitoring (RMT) by combining different sensors to detect signatures helpful for predicting outcomes in MDD. RADAR-MDD will take advantage of the combination of multiple sensor types along with remote data collection from a clinical population.

Detection of seizures using non-EEG wearable devices has been reasonably well studied over the last decade [7]. However, performance of the proposed models has often been unsatisfactory, particularly in terms of specificity, although some studies do show some promising results for seizures with a large motor component [3]. Additionally, few studies have been conducted outside of an in-patient environment, so the performance of these models in the real-world is unknown. To address the accuracy issues of models that only use a single sensor type, usually an accelerometer, there has been a movement within the field towards detection using multiple modalities[7]. A few studies have used using multiple sensors. Poh et al. used electrodermal activity (EDA) and an accelerometer, and showed increased GTCS detection performance when using both as opposed to only acceleration in 7 patients[16]. Heldberg et al. also used EDA and an accelerometer, looking at both convulsive and non-convulsive seizures in 8 patients [11]. Other studies have looked at the combination of acceleration and ECG-derived cardiac features[19, 9]. The use of multiple sensors does not always uniformly lead to better performance; Milosevic et al. report improved seizure detection but lower specificity when using both accelerometers and electromyography[15]. Finally, through the RADAR-base platform we have developed a well engineered open source platform with highly generalizable capabilities.

Methods

Remote Data Collection for Major Depression

RADAR-MDD, the major depression clinical substudy of RADAR-CNS, makes use of a range of data collection instruments as discussed below.

Passive RMT (pRMT) app

The passive application runs in the background, requiring minimal or no input from participants. Data is collected from smartphone "sensors" corresponding to a range of cate-

gories considered putatively relevant to the study, including (i) movement sensors: acceleration, gyration, and steps, and obfuscated relative GPS location; (ii) social characteristics: call duration, a log of SMS communications, contact list, and nearby Bluetooth handshakes; (iii) environmental sensors: ambient light, battery level, magnetic field, and weather conditions; (iv) user interaction with other applications and their phone; and (vi) keystrokes are collected in a subsample. All the collected data is pseudonomised, for example by hashing contacts names and phone numbers, and by using an unknown offset to obfuscate location.

Wearable Sensors
The Fitbit Charge 2 was selected to be worn by participants in RADAR-MDD for the duration of the study, providing metrics derived from the watch accelerometer and photoplethysmography (PPG). These data are processed on the device by vendor algorithms to provide information on heart rate, movement, daytime and sedentary activity, physical exercise, step count, and sleep patterns and efficiency. Data is collected into the RADAR-base platform from the Fitbit Web API, using the 3rd Party Data Integration service.

Active RMT (aRMT) app
Variation in the depression symptoms are measured via the 8-item Patient Health Questionnaire (PHQ8) [13] every 2 weeks throughout the course of follow-up. Variation in self-esteem is measured using the Rosenberg Self-Esteem Scale (RSES) [10]. The RSES is a widely-used 10-item self-reported questionnaire used to quantify self-esteem along a continuum and is administered alongside the PHQ8 every 2 weeks. As with the PHQ8 and RSES, every 2-weeks participants are asked to complete a speech task. This requires participants to read aloud, in a quiet area, some excerpts from "The North Wind and the Sun", which has been shown to be phonetically balanced across all three languages [20]. The excerpts are offered on a random schedule to prevent rehearsal and fluency and preserve prosodic features. In addition to this, participants are asked to respond to the following question: "Can you describe something you are looking forward to this week?". The aRMT app also delivers an Experience Sampling Method (ESM) schedule, designed to collect brief, in-the-moment assessments relating to several domains of interest: mood, stress, sociability, activity and sleep. Participants will receive a series of questions intended to reflect their current state (such as "right now, I feel content"), with 7-point Likert scale answer options (0=Not at all, 7 = Very much). The ESM schedule consists of approximately 44 items, taking up to 3-minutes to complete, delivered 9 random times per day within 90-minute blocks starting from 08.30 and ending at 22.00 for 6 consecutive days every 6 weeks.

THINC-IT app
THINC-it is a third party app used to assess cognitive function both objectively and subjectively, validated for detecting cognitive dysfunction in patients with MDD[14]. It incorporates four game-like digital assays, variants of widely-used cognitive assessments and a 5-item questionnaire assessing perceived deficits in memory, concentration, and attention over the previous week.

In-Hospital Data Collection for Epilepsy seizure detection

Hospital in-patient participants are recruited for the RADAR-EPI substudy of RADAR-CNS study as part of an otherwise typical stay at the Clinical Neurophysiology Department at Kings' College Hospital (KCH), London, UK or the Epilepsy Center at the University Hospital of Freiburg, Germany. Patients are monitored by a video-EEG and seizures are annotated by clinicians as part of the routine clinical assessment of their seizures. This provides ready-made source of

gold-standard labels for use in developing wearable-based seizure detection methods. In parallel, each patient wears 1 to 3 of the study wearable devices; the Empatica E4 wristband, Faros 180, Biovotion VSM1, or an offline IMEC device. The IMEC device records data offline, which is routinely transferred to the RADAR-base storage server. The other devices send data to the RADAR-base platform via a Bluetooth-paired android device through the passive RMT app.

Passive RMT (pRMT) app
The RADAR-base passive app (pRMT) has the capability to quickly integrate data sources (via pRMT plugins) such as wearable devices. The Empatica E4, Faros 180, and Biovotion VSM devices have been integrated for the EPI study. Each device was selected because of its ability to monitor physiologically relevant parameters. Acceleration, EDA, and heart rate. Cardiac features are measured either by PPG in the Empatica and Biovotion devices, or by ECG in the Faros device. The IMEC, although not integrated into the pRMT due to unavailability of a software development kit but has similar sensors to the Faros. Raw data is collected directly over Bluetooth (in comparison to the Ftibit where data is retrieved from the vendor data warehouse).

Study Population

As part of the RADAR-CNS programme RADAR-base is deployed to carry out RADAR-CNS studies at 8 sites across Europe, with the goal of enrolling MS (n=640), MDD (n=500) and EPI (n=200) participants.

Current Status of the MDD and EPI Studies
At present there are 66 enrolled patients in the MDD study at KCH. So far there has been a total of 127 patients enrolled in the EPI study across KCH and Freiburg.

Statistical and Analysis Plan

Preliminary analysis of the MDD dataset will investigate correlations between the PHQ-8 scores and basic aggregated features, obtained from the recording biosensors, which should be representative of behaviours associated with depression. The PHQ-8 questionnaire is taken every 2 weeks, and so the outcome is at a much lower frequency than the raw signals. Simple proxies for sleep, activity, sociability, cognition, and ambulation will be used to classify current depression, where current depression is determined by a PHQ-8 score ≥ 10. The ability of those features both to detect depressive periods between subjects and to monitor the progression of depressive symptoms within individuals will be explored. Rarer clinical relapse will also be reported for the cohort over the 2 year data collection period providing more definitive outcome measure where present. It may be necessary to design features in such a way that they are able to deal with missing data, or else use a model that is able to use missingness informatively [4].

The initial analysis of the epilepsy data requires a different approach. Seizures are typically short and sparse, so the primary challenge is to be able to detect the relatively short periods of ictal activity between the much more common segments of interictal time, while keeping the false positive rate at an acceptable level. Initial focus will be on the offline detection of generalised and focal seizures with a motor component, particularly those with tonic or clonic movements, using a combination of the available signals. The combination of the different signal modalities should improve the estimation accuracy of a single relevant parameter, and also allow the analysis using multiple physiologically-relevant parameters, enabling a panoramic view of the patient's status.

Because a few participants have had a large number of

seizures (n > 15) during their in-patient stay, there is an opportunity to measure the performance of an individualised seizure detection algorithm as successive seizures are added to the model. This will have application to future ambulatory studies, in which it will be important to know how much data is required for an adequately accurate model. Firstly, we will follow a analytical pipeline similar to those in prior seizure detection work, extracting features from the EDA, accelerometer, and heart rate signals, and classifying the ictal period of focal motor seizures using a support vector machine. Although the specificity of previous work has been too low, the larger sample size of the EPI study may help improve accuracy of similar models. Additionally, we will try and determine feature importance with the intention of elucidating performance gain from including additional signal modalities. Subsequently, we will investigate the feasibility of using deep learning techniques which may provide better generalization. Given the relative sparsity of ictal data, it may be necessary to use unsupervised neural networks to extract features, or to use transfer learning from the activity recognition domain.

Results and Discussion

A tonic EDA response during the post-ictal period has been noted elsewhere[8], and often occurs within the RADAR-EPI dataset. An example is given in Figure 1, showing an Empatica E4 recording of acceleration and EDA over a night-time 5-hour period. The convulsive seizure at 05:05 is followed by a large increase in skin conductance, with a peak at 05:10. There are other tonic peaks in the EDA, but they do not coincide with a seizure-like accelerometer trace. Equally, there is not evidence accelerometer traces with repetitive or otherwise confusable characteristics in the inter-ictal period being succeeded by an EDA response. Although not totally consistent across all participants and all seizures, it is a general pattern that illustrates the potential to use multiple modalities for increased specificity.

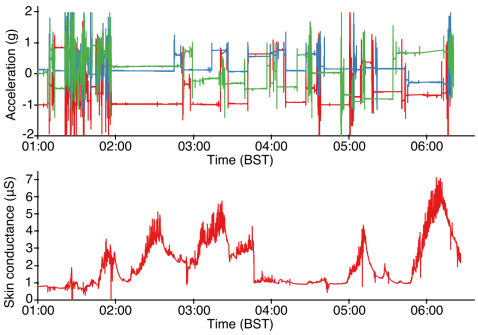


Figure 1: Data stream from a patient wearing an Empatica E4 during a night. The patient had a focal motor seizure at 05:05 (BST), corresponding to a burst of movement in the accelerometer (top), and subsequently followed by a peak in EDA (bottom). Other movements and peaks in EDA during the interictal periods do not follow the same pattern.

The preliminary data from the MDD study shows a range of depressive symptoms, with a mean PHQ-8 score of 10.4 and standard deviation of 6.2 in the 76 PHQ-8 questionnaires so far recorded. Five participants have had a depressive episode, progressing from a PHQ-8 score < 10, no depression, to a score >= 10, current depression, in the following questionnaires. Of those, one returned to a 'no depression' state after a week. There is, therefore, already a small amount of intra-individual variation recorded, although longitudinal effects should become clearer as the follow-up data collection period continues.

1740

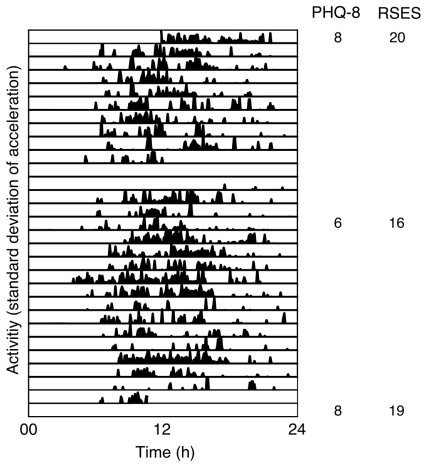


Figure 2: Activity of a participant as measured by the standard deviation of their phone's acceleration with corresponding PHQ-8 and RSES scores collected on the first day and every fortnight thereafter. Each row corresponds to a day. The questionnaire scores suggest the participant is not in a depressive mood (PHQ-8 scores < 10) and has a normal level of self-esteem (RSES scores between 15-25). There is missing data on days 10-12.

Missing values may prove a challenge for the MDD analysis. Firstly, due to technical challenges associated with a project of this magnitude. Secondly, and more commonly, through participant non-adherence and differing levels of engagement with the study applications and their phone in general. Even in patients with high adherence, there are likely to be times during which data is not collected. Figure 2 shows the first month of accelerometer data from a participant, alongside PHQ-8 and RSES responses. Although overall adherence is high for this participant, there is still a 48 hour gap during which no data is available. Disentangling missingness due to technical issues and missingness due to the participant non-adherence, and then directly incorporating that information into a model may be important, because depressive symptoms may affect adherence.

Acknowledgements

This work has received support from the EU/EFPIA Innovative Medicines Initiative Joint Undertaking 2 (RADAR-CNS grant No 115902) www.imi.europa.eu. This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. We would like to acknowledge The Hyve and RADAR-CNS Consortium (http://www.radar-cns.org/partners) for their support. Backend Infrastructure facilities were provided by King's College London Rosalind. The Authors receive funding support from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London.

References

[1] 2018. RADAR-base. https://radar-base.org/. (2018).

[2] M. Alvarez-Jimenez, S. Bendall, R. Lederman, G. Wadley, G. Chinnery, S. Vargas, M. Larkin, E. Killackey, P.D. McGorry, and J.F. Gleeson. 2013. On the HORYZON: Moderated online social therapy for long-term recovery in first episode psychosis. *Schizophrenia Research* 143, 1 (2013), 143 – 149. DOI: http://dx.doi.org/https://doi.org/10.1016/j.schres.2012.10.009

[3] Sándor Beniczky, Isa Conradsen, Oliver Henning, Martin Fabricius, and Peter Wolf. 2018. Automated real-time detection of tonic-clonic seizures using a wearable EMG device. *Neurology* (2018). DOI: http://dx.doi.org/10.1212/WNL.0000000000004893

[4] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8, 1 (2018), 6085.

[5] RADAR-CNS consortium. 2016. RADAR-CNS: Remote Assessment of Disease and Relapse in Central Nervous System Disorders. https://www.radar-cns.org/. (2016).

[6] Hanneke M. de Boer, Marco Mula, and Josemir W Sander. 2008. The global burden and stigma of epilepsy. *Epilepsy and Behavior* 12, 4 (2008), 540 – 546. DOI: http://dx.doi.org/https://doi.org/10.1016/j.yebeh.2007.12.019 Current Views on Epilepsy and Behavior.

[7] Anouk Van de Vel, Kris Cuppens, Bert Bonroy, Milica Milosevic, Katrien Jansen, Sabine Van Huffel, Bart Vanrumste, Patrick Cras, Lieven Lagae, and Berten Ceulemans. 2016. Non-EEG seizure detection systems and potential SUDEP prevention: State of the art: Review and update. *Seizure* (2016).

[8] Onorati Francesco, Regalia Giulia, Caborni Chiara, Migliorini Matteo, Bender Daniel, Poh Ming-Zher, Frazier Cherise, Kovitch Thropp Eliana, Mynatt Elizabeth D., Bidwell Jonathan, Mai Roberto, LaFrance W. Curt, Blum Andrew S., Friedman Daniel, Loddenkemper Tobias, Mohammadpour-Touserkani Fatemeh, Reinsberger Claus, Tognetti Simone, and Picard Rosalind W. Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. *Epilepsia* 58, 11 (????), 1870–1879. DOI: http://dx.doi.org/10.1111/epi.13899

[9] F. FÃijrbass, S. Kampusch, E. Kaniusas, J. Koren, S. Pirker, R. HopfengÃď rtner, H. Stefan, T. Kluge, and C. Baumgartner. 2017. Automatic multimodal detection for long-term seizure documentation in epilepsy. *Clinical Neurophysiology* 128, 8 (2017), 1466 – 1472. DOI: http://dx.doi.org/https://doi.org/10.1016/j.clinph.2017.05.013

[10] Ellen Greenberger, Chuansheng Chen, Julia Dmitrieva, and Susan P. Farruggia. 2003. Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: do they matter? *Personality and Individual Differences* 35, 6 (2003), 1241 – 1254. DOI: http://dx.doi.org/https://doi.org/10.1016/S0191-8869(02)00331-8

[11] B. E. Heldberg, T. Kautz, H. Leutheuser, R. HopfengÃď rtner, B. S. Kasper, and B. M. Eskofier. 2015. Using wearable sensors for semiology-independent seizure detection - towards ambulatory monitoring of epilepsy. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 5593–5596. DOI: http://dx.doi.org/10.1109/EMBC.2015.7319660

[12] Sidney Kennedy. 2016. A Collaborative Investigation of Predictors of Relapse in Major Depressive Disorder: CAN-BIND-1 Extension Study. https://clinicaltrials.gov/ct2/show/NCT02934334. (2016).

[13] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 114, 1 (2009), 163 – 173. DOI: http://dx.doi.org/https://doi.org/10.1016/j.jad.2008.06.026

1742

[14] Roger S McIntyre, Michael W Best, Christopher R Bowie, Nicole E Carmona, Danielle S Cha, Yena Lee, Mehala Subramaniapillai, Rodrigo B Mansur, Harry Barry, Bernhard T Baune, and others. 2017. The THINC-Integrated Tool (THINC-it) Screening Assessment for Cognitive Dysfunction: Validation in Patients With Major Depressive Disorder. (2017).

[15] M. MiloÅąeviÄĞ, A. Van de Vel, B. Bonroy, B. Ceulemans, L. Lagae, B. Vanrumste, and S. V. Huffel. 2016. Automated Detection of Tonic-Clonic Seizures Using 3-D Accelerometry and Surface Electromyography in Pediatric Patients. *IEEE Journal of Biomedical and Health Informatics* 20, 5 (Sept 2016), 1333–1341. DOI: http://dx.doi.org/10.1109/JBHI.2015.2462079

[16] Poh Ming-Zher, Loddenkemper Tobias, Reinsberger Claus, Swenson Nicholas C., Goyal Shubhi, Sabtala Mangwe C., Madsen Joseph R., and Picard Rosalind W. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia* 53, 5 (????), e93–e97. DOI: http://dx.doi.org/10.1111/j.1528-1167.2012.03444.x

[17] John A. Naslund, Lisa A. Marsch, Gregory J. McHugo, and Stephen J. Bartels. 2015. Emerging mHealth and eHealth interventions for serious mental illness: a review of the literature. *Journal of Mental Health* 24, 5 (2015), 321–332. DOI: http://dx.doi.org/10.3109/09638237.2015.1019054 PMID: 26017625.

[18] Janssen Research and Development. 2018. A Prospective, Longitudinal, Observational Study to Evaluate Potential Predictors of Relapse in Subjects With Major Depressive Disorder Who Have Responded to Antidepressant Treatment. https://clinicaltrials.gov/ct2/show/NCT02489305. (2018).

[19] Shivkumar Sabesan and Raman Sankar. 2015. Improving long-term management of epilepsy using a wearable multimodal seizure detection system. *Epilepsy & Behavior* 46 (2015), 56–57.

[20] James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Gregory Ciccarelli, and Daryush D. Mehta. 2014. Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge (AVEC '14)*. ACM, New York, NY, USA, 65–72. DOI: http://dx.doi.org/10.1145/2661806.2661809