Everyday data for COVID-19 from mHealth devices

The PAIDUR framework

Colin Wylie Brown

A thesis submitted in partial fulfilment of the requirements

of Edinburgh Napier University for the award of

Master by Research

School of Computing

Sponsored by

Digital Health and Care Innovation Centre

September 2022

AUTHORSHIP DECLARATION

I, Colin Brown, confirm that this thesis and the work presented in it are my own independent achievement.

The work has not been submitted for any other degree or professional qualification.

Where I have consulted the published work of others this is always clearly attributed.

Where I have quoted from the work of others the source is always given.

With the exception of such quotations this thesis is entirely my own independent work.

I have acknowledged all main sources of help.

I have read and understand the penalties associated with Academic Misconduct.

I also confirm that I have considered consent from all those involved in the work, following the School's ethical guidelines, and that none was required.

Signed

29 October 2021

Colin Wylie Brown

Matriculation : 40454980

DATA PROTECTION DECLARATION

Under the 1998 Data Protection Act, The University cannot disclose your grade to an unauthorised person. However, other students benefit from studying theses that have their grades attached.

The University may make this thesis, with indicative grade, available to others

Colin W Brown

ABSTRACT

This work examines the use of new data from mHealth devices in algorithmic Risk Predictor tools for conditions including the early diagnosis of COVID-19.

The earliest signs of COVID-19 and other viral infection were from FitBit® devices, and data from these and other mHealth devices on three research platforms are analysed. MHealth devices are selected for everyday use for lifestyle. However users in healthcare may lack expertise, and staff can process only the least quantity and highest quality of mHealth data. On-device processing helps by deriving simple outputs such as Heart-Rate-over-Steps, but is not standardised.

Clinicians have long processed simple data in the mind, as was prototyped for frailty with algorithms designed for mental arithmetic. Qualitative accuracy is analysed here as “Healthcare Veracity”, an issue familiar to clinicians and managed pragmatically, usually by time-serialisation. Such clinical quality assurance can also apply to mHealth devices, whose metrological precision and accuracy exceed that of older technologies. Accuracy further improves when each device is paired to its user.

New tools could use machine data directly from local devices, or from remote devices if connected for interoperability. However, new devices require clinicians’ trust, and this has been evaluated in developing the PAIDUR framework:

Precision / Accuracy / Interoperation / Deployment / Use / Reuse

This structures how new mHealth systems can manage socio-technical issues to interoperate their new data across healthcare systems.

Consumers’ inexpensive mHealth devices use the same technologies as branded and healthcare-certified products. Mass deployment is already here, and online platforms can already support multiple device types at scale.

However, the costs of consumer platforms are also to privacy, which funds the business model, with implications for data quality. This suggests how privacy can be actively managed to keep proven healthcare benefits free at the point-of-care.

ACKNOWLEDGEMENTS

Thank you to my supervisors, Dr Alistair Lawson and Dr Peter Barclay for taking this on with many helpful suggestions and appropriate scepticism, and navigating the university Research Degree Framework and Regulations.

Thank you also to my wife and family for general and insightful support on research process, as three are PhD candidates, and two more have research degrees from years ago!

I am also very grateful to
Dr. Ian McNicoll, for initial conversations, support and some mind-expanding citations.
Heidi Kokkalainen MSc, for hands-on advice on archetype development.
Glen Griffiths MSc, for detailed feedback on the engineering and survey sections.
Adrian Smales PhD, for two rich conversations around frailty and resolution.
The late Dr Leo Fogarty FRCGP, for showing how a polymath does detail and depth.
Former colleagues, who must remain anonymous, for their responses at Section 6.1.
and finally, Hugh Kearns, of Flinders University Adelaide for regular tweets on managing one's own research.

TABLE OF CONTENTS

page

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION AND AIMS

1.1 Why

The COVID-19 pandemic has challenged both biomedical science and the delivery of health and social care, and exposed the poor interoperability of health with care data. New data from mobile sources and wearable devices, known as mHealth[1], which are now abundant in range and of global scale, present the challenge of interoperability with data used in current health and care systems.

1.2 What

Diagnosis of healthcare conditions is the essential precondition for safe and effective care. Yet COVID-19 is often difficult to diagnose by conventional clinical methods, especially in old age; and currently no single test is accurate enough on its own to diagnose the multiple variants of COVID-19 at all ages.

In old age, similar issues are found around frailty. This work analyses how different types of data are combined to make diagnoses, how frailty indices are prototype Risk Predictor tools (RP tools), and how new data can improve them.

The role of frailty itself as a risk predictor of COVID-19 is controversial, as to date it has relied on conventional data sources; but new data sources are now widely used that can benefit COVID-19 and also other health and care. These mHealth devices create new content in new types of data that can be added to RP tools to address a range of conditions beyond frailty, which may include the early diagnosis of COVID-19.

The pandemic demands a focus on rapid application, and this work offers a framework of considerations for manufacturers and for health and care organisations when introducing data from mHealth technologies into healthcare systems.

This “PAIDUR” framework addresses how to connect into NHS systems the new data that is acquired increasingly by consumers using their own mHealth devices rather than by Healthcare Workers (HCW).

---

[1] See Glossary for explanation of terms *in this context*.
First use of defined terms is in Initial-Capitals. See Glossary v0.8 - Google Docs for Live version

1.3 How – method overview

The initial research questions were about data gaps for frailty indices, comparing them to other RP tools, and importing data from mHealth devices into social care. OpenEHR data models was selected as an international open-source technological architecture for comprehensive electronic health records (EHR) (Heard, 2010; Koikkalainen, 2019), and are compatible with FHIR®, currently more widely used. Then COVID-19 struck, and rapidly displaced spare capacity for work with no immediate application to this sometimes existential threat. It was decided to change to data sources for COVID-19, for which early work had suggested a role in early diagnosis for data from Fitbit® devices (Heneghan et al., 2019; Natarajan et al., 2020)

This led eventually to these 4 research questions

1. *What are the types of data currently used in Risk Predictor tools?*

The datatypes already in use show the logical types and healthcare content for mHealth technologies.

2. *Which healthcare data can be extended using wearables and mHealth, for early diagnosis of COVID-19?*

New data can be acquired, and current data improved, by these new technologies.

3. *How can Extended Data sources be Automated as Everyday Data?*

Key features of mHealth are automation of continuous asynchronous data collection, which depends on consumer factors such as wearability, battery life and usability.

4. *How to render Everyday Data for generic use in any Health and Care system?*

Thousands of datatypes created by the global market may already be unusable here, due to lack of design for technical interoperability, or for semantic validity in healthcare.

2 BACKGROUND AND CONTEXT

This work originated as the addition of new data sources for frailty or other conditions relevant to social care. This arose from the author's experience as a general practitioner that much of the data used in Social Care was the same as that used in Health Care e.g. medications, conditions, care plans, nursing observations and social demographics, but that it was still managed on paper, using forms customised to each Social Care organisation. In informatics terms it is some two decades behind the electronic health records used in the NHS. This creates administrative work at every transfer of care between Social Care and NHS, and this burden was framed here as due to lack of Interoperability of the Health IT and Social Care IT systems.

This informatics deficit was researched by Johnston et al., (2020) also at Edinburgh Napier University, finding that several datasets could be resolved to a core dataset of fifteen data items, of which no. eight was frailty – yet in six care homes surveyed, four used different measures, and two made no record of this at all. Thus a transfer of frailty data from social to health care settings as unstructured native data would be useless unless interpreted by humans as simple narrative data. This systematic absence of interoperability requires a systematic approach, and this work evaluates its application to mHealth systems.

2.1 Risk Predictor tools: the Frailty Score prototype

If the first task of healthcare is to diagnose, this has long been recognised as more art than science, using the indefinable "clinical experience" to mentally co-process multiple incompatible datatypes that are a mixture of quantitative and qualitative. However as more data becomes quantitative and computable, some quantitative predictors have been developed, described by Dambha-Miller et al., (2020) as "algorithms designed to predict outcomes, aid decision making, support treatment options, manage clinical risk, or improve efficiency. The term clinical scoring system is known interchangeably as clinical decision rule, prediction algorithm, clinical prediction tool, risk score, or scoring tool"; they will be known here as "Risk Predictor tools"

These are used in frailty, long known as a complex diagnosis, to integrate different components to construct a single concept that appears simple to understand in everyday life, but in healthcare practice is hard to quantify. These multiple components are combined as a Frailty Score, to quantify the risk of a frailty-related

outcome such as fall, hospital admission or death. Such a combination is known here as a Risk Predictor (RP) tool: it combines mixed data that may be natively quantitative, or so derived as a quantitative measure from qualitative data.

Most components of frailty tools pre-exist in NHS records, and it was proposed that RP tools could be improved by connecting a limited number of high-value Social Care or mHealth data items using internet technologies.

2.2 New data sources from mHealth

Meanwhile (Smales, n.d.), also at ENU, had established that inactivity, step count and other new parameters could be reliably measured by new mHealth devices[2], and could be algorithmically combined in the (ARMED, n.d.) (Advanced Risk Modelling for Early Detection) RP tool, creating a strong predictor of hospital admission from care homes.

Can these mHealth devices supply new quantitative data to RP tools beyond the simple Frailty Scores, that currently use mixed data from user or HCW observations and traditional devices?

2.3 New data for COVID-19

Then COVID-19 struck, and all healthcare capacity was consumed by the crisis.

It became clear that COVID-19 was also a complex condition with diagnosis complicated by at least 6 syndromes (Sudre et al., 2020) and many features in the elderly that are non-specific, such as fatigue: the most common symptom of all, but wholly non-specific for COVID-19.

Further, it was found that COVID-19 is most infectious in the first days of illness, and for several days before the first symptoms (Mishra et al., 2020). In COVID-19 the value of early detection is very high due to transmission for several days before symptoms.

This “asymptomatic transmission” may both never show symptoms in the source, nor cause them in the target, so this unusual silence of the infection allows unhindered spread. This is much worse with 2021’s delta variant, as detailed at Section 3.4.1.

It poses a massive challenge to the interception of transmission, for example to isolate or treat those infected, instead of population-wide “lockdown” measures with their huge social costs. A mass-scale means of finding cases before symptoms is required.

Could such new devices be used to detect illness more quickly? Early results from California suggested so (Mishra et al., 2020) (Quer et al., 2021).

---

[2] “mHealth” is used here for mobile device(s) for healthcare that connect to remote systems, usually by mobile phone, and to include both body-wearable and point-of-care devices.

2.4 Healthcare scenarios and settings

Feasibility of potential mHealth data sources in COVID-19 is considered here in these six sequential clinical contexts or scenarios in Primary or Community-based Care[3]:

a) C**a**se-finding in the whole population
b) Vulnera**b**le / At Risk e.g. residents of Care Homes, those with co-morbidity
c) **C**ontacts, either self-identified or contact-traced, self-isolating at home for a defined period such as 10d
d) **D**iagnosed new cases, whether by 3 classic symptoms, other features, or tests
e) R**e**covering cases still at home, or returned from hospital, and managed
    - by Primary Care teams, or
    - as "Hospital-at-Home" by hospital-based staff
f) **F**ailing recovery or "Long Covid"[4] cases showing relapses, or few signs of recovery, extending to 12wks in 10% of cases (*Living with COVID-19*, 2020).

These scenarios in Primary Care settings can here be compared to hospital settings:

i. direct care needs are lower, in volume and in complexity
ii. attendance by HCW is infrequent and may be nil
iii. HCW supervision may be remote, by voice- or video-call
iv. local devices can acquire their own data independent of HCW attendance
v. data sharing by mHealth may not need HCW attendance, so is asynchronous
vi. a device is usually used by only one subject, and always so when they have bought it!

An example for scenario e) shows these features: McKinstry et al., (2021) describing pulse oximeter use in a telecare system piloted in NHS Scotland, state that "patients were given a pulse oximeter." This persisting pairing of device to user is important for evaluating its data error, in both precision and accuracy, which support user trust in each device's data – see Section 2.7.4

These metrological concepts of accuracy and precision are analysed further, as new devices may acquire trust by their sole user, but do not necessarily acquire similar trust by a HCW – see Section 6.3

[3] Also known internationally as Domestic or Outpatient Care
[4] Long Covid is a condition of unclear definition, with over 200 clinical features suggested. The 12-wk. entry criterion is controversial, as is overlap with post-viral or other chronic fatigue syndromes

2.5 Health and Care systems and Translational Medicine

It is often said that simple IT solutions to health and care systems are usually wrong. Between invention or discovery of a new data source and its deployment into routine use, lies a complex socio-technical enterprise[5] with multiple points of failure (Prof. Mike Martin, Newcastle) that can be analysed using the framework of Greenhalgh et al., (2017), as due to Non-adoption, Abandonment, Scale, Spread or Sustainability factors - see Section 3.5. This NASSS framework is structured as a dynamic model of criteria for implementing change in socio-technical systems.

These also have many features of Complex Adaptive Systems, characterised by

- having no single leader (*Complex Adaptive Systems Evidence Scan*, 2010)
- actions may produce null or even contrary results due to feedback loops
- change can be better effected by applying several small but repeated "nudges" to multiple parts of the system.

This and other frameworks were explored to develop this PAIDUR framework.

MHealth can also be considered in the strategic context defined by the European Society for Translational Medicine (Cohrs et al., 2015): "an interdisciplinary branch of the biomedical field supported by three main pillars: benchside, bedside and community. The goal of TM is to combine disciplines … to promote enhancements in prevention, diagnosis, and therapies. Accordingly, TM is a highly interdisciplinary field."

This systematic interdisciplinary use of basic scientific research directly in practical health and care use is more goal-driven than "Applied Research" and seeks to bridge the "Valley of Death" between bench and bedside (Seyhan, 2019). Many technologies fall into this as if simply dropped into the complex socio-technical enterprise of health and social care, since as well as design, engineering and manufacture to known quality, they require deployment, initialisation, education, training with standard operating procedures, and maintenance.

---

[5] For Sociotechnical Systems, Prof Mike Martin describes these 4 sub-domains

Engineering	= (health) technology
Informatics	= (clinical) information science
Communications	= conversations in channels (bandwidth, signal:noise)
Socio-political	

2.6 Interoperability: the crisis of Health and Care IT

For reuse in other systems, the types of data obtained from proprietary devices must be so designed. This huge topic is considered here at two levels

2.6.1 Engineering: data can be copied into other systems

This is the technical domain of connecting devices to each other and to current installed HealthIT systems, using basic engineering functions such as Create, Read, Update, Delete actions on data in files, or Get, Post, Put, and Delete actions for data in RESTful web resources, all of which depend on effective IT networking infrastructure. These are outlined for openEHR systems by Sundvall et al., (2013) but are out-of-scope here.

2.6.2 Semantic: data are in a format that can be re-used as if native data.

This is the informatics domain of transforming data while retaining meaning, so requires knowledge of how data are used in healthcare to create actions.

Fortunately there is much overlap in the content of data held in social care systems with that held in healthcare systems i.e. demographics, conditions, observations of health status, medication records, personal and nursing care needs are already held in HealthIT systems. However, unless these have been modelled in the same format in both systems, the interoperability problem arises because different systems must transform the data before reusing it, which work needs time and money.

In Direct Care, there is usually neither, so the all-too-common experience is that data are re-acquired by repeat questioning or testing of the patient (Rooney et al., 2018), since for the HCW this routine healthcare task is easier than the novel IT task of retrieving data effectively lost in other systems.

Because each system models healthcare data to different proprietary standards or platforms, to reuse data requires connections customised for each pair of legacy systems, usually using technologies such as those supported by HL7®: HL7v2 and now FHIR. However, the multiplication of systems increases exponentially such custom connections systemwide (Beale, 2018).

Healthcare data has long been expected by HCW to be generic, and for most its quality is formally guaranteed by certification of the specific device. The supplier or brand of device is relevant to the validity of its data, as evaluated at Section 6.3.

Fortunately much data from new devices is also generic because it is in the standard formats of simple rational numbers expressed in standard units of measure, such as events/sec, weight in kg, blood pressure in mmHg[6]. Proprietary technology is found in the transducer and transformation of its electrical output to these standard units, and the accuracy of these transforms is usually certified - see Section 2.7.4 There has been no formal healthcare practice for the device provenance of data to be recorded as metadata, because there is high trust in devices that are certified, and this may extend informally to devices using the same technology even where uncertified.

But not all data are so simple: much new data is derivative, via on-device processing using non-standard algorithms, and is processed by classifiers into non-standard classes - see Section 3.1.11. These new data and classes include new sleep parameters, derivatives of pulse such as resting heart rate (RHR) and heart rate variability (HRV), activity derivatives such as patterns of inactivity, or sleep components.

These technologies and algorithms are not only proprietary but intentionally kept secret by patents.

The global scale of mHealth devices has created vast quantities of such derived data, and the challenge here is to assess them for safe and effective places in healthcare.

2.6.3 Architecture, terminology and openEHR: the language analogy

It was intended to use openEHR, a standard format with international recognition as a generic "grammar" of clinical models: a fuller account is now at Appendix 10.1

In brief: standard words can be provided by a terminology, and with isolated words a degree of basic communication is possible; but for full meaning to be communicated with language requires a grammar such as for word order, punctuation, phrasing (Beale, 2018). Further, infrastructure is necessary for words and grammar, such as pen & paper, speech, or as here, IT.[7] Essentially, terminology supplies the words, openEHR supplies the grammar, and IT supplies the infrastructure to connect them.

For mHealth, devices provide the data, and openEHR can structure them as if with grammar so healthcare systems can reuse them.

But this semantic interoperability depends on the further principles of the PAIDUR framework developed here.

---

[6] still standard after all these years!

[7] Infrastructure needs to be designed to be neutral: language when spoken can include further information in accent, intonation, etc.

2.7 Types of Data

A major problem for informatics in health and care is that most of its data are unstructured such as narrative text. The essential method of making such data computable is to add structure to text, which can then be machine-processed.

In information science, mathematical logic is the grammar that combines terms and creates meaning from the knowledge model rendered in narrative text. (Coiera, 2015) p15. For qualitative data, mental processing is not always in language[8]: cognition operates in many modes other than language, as sound or images can be understood immediately, as shown by the power of music or film. After we perceive native data, we may immediately find meaning in it by such multimodal mental processing. For example, the meaning of the Red-Amber-Green "traffic light" coding is learned in childhood and is commonplace.

To create meaning from large sets of data, the complex logic of Machine Learning is now used, algorithms from which are now integrated into mHealth devices. The validity of these machine-constructed information-rich models is considered at Section 3.1.12.

These ML rules often require further interpretation of context by "combination" with external qualitative data. This task, known here as co-processing, is critical to safe and effective care, and "sense and safety" checking is embedded – see Section 2.7.4.

The complementary and variable combination of computer with mental co-processing is a frequent theme, illustrated here by Coiera, (2015 p19) for a set of data that may be processed either by human or machine, or in any combination.

[8] Language itself is an enormous knowledge model using words as symbols to model concepts in terms that are mutually understood, by reference to external models of meaning shared with other languages

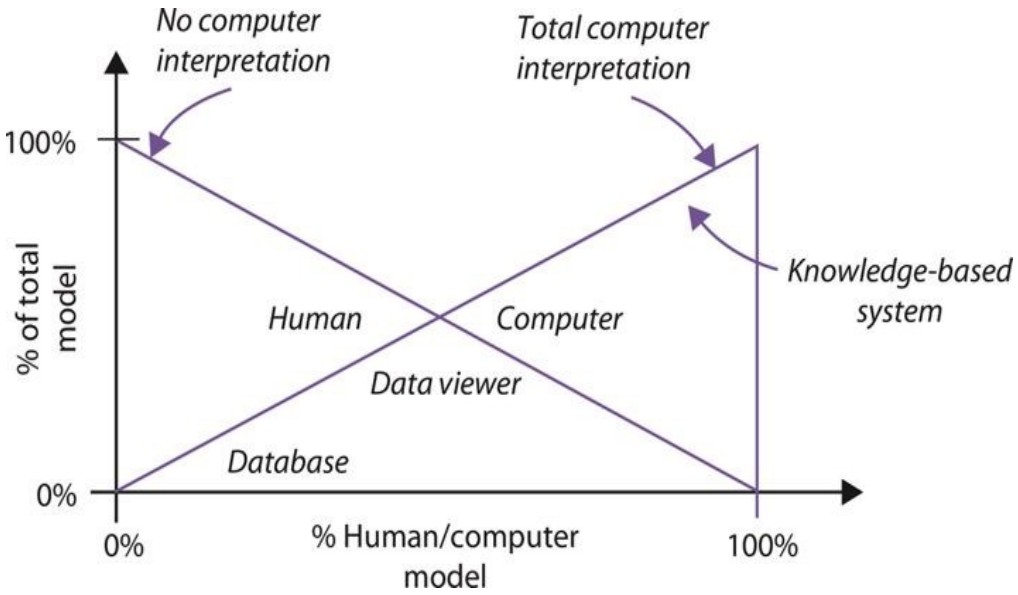


Figure 1: co-processing with computer and brain
Coiera: Guide to Health Informatics 3rd Ed p19

"For example, the computer may organize and consolidate data into a graphical presentation, and the human then examines the processed data to make a final interpretation. The proportion in which models are stored either in the computer or as mental models in the head of a human determines where the interpretation takes place. Computer systems thus form a spectrum, ranging from those that have no ability to assist in the interpretation of data to those that are able to carry out a complete interpretation, within the bounds of a given task"

2.7.1 Digital, discrete, continuous, analogue data – what's the difference?

These terms for types of data need definition, and one criterion is any associated time data. This is easily seen for events in the unit of measure, since all can be rendered as a continuous number by including time data in appropriate units.
For repeating events, the SI unit is Hertz, and for some non-cyclic events there are special units[9]: all of these units are quantifiable.
Thus for simple heart rate, which may be highly irregular, a tally of discrete events is rendered into a continuous number, known in healthcare as beats-per-minute[10] (though baud might also be used)

[9] "baud" for symbol change events, as in radio comms, or "becquerel" for nuclear decay, the ultimate in randomness: both units can render any random data as continuous numbers
[10] the many meaningful patterns are not quantified but described in standardised text terms such as ectopic, regularly irregular etc. Their cardiology meaning persists as text but is not quantified.

The time attributes of events may instead be recorded in metadata: even a singular event may be timed with one date-time stamp. Such timing metadata uses the clock control intrinsic to every microprocessor, and is essential for time-serialisation. Thus any series of events may be transformed to an analogue value measured as events sec^{-1}. As shown for heart rate above, these units make no implication that data are continuous, only that it is a series of discrete events of unknown regularity, and that they can be counted (*Is the Brain Analog or Digital? | NeuroLogica Blog*, n.d.).

Following the analysis by (Maley, 2011), the semantic schema used here is:

Continuous	applies to data	of the datatype "rational numbers"
		- may be transformed to discrete classes
Discrete	applies to data	- includes cardinal and ordinal datatypes
		- multiple classes may be reduced to binary
		- may be transformed to Continuous using s^{-1}
Analogue	applies to devices	process continuous datatype only
Digital	applies to devices	process both continuous and discrete datatypes

However, "Digital" is also a multivalent term, which can mean any of

- quantities rendered in discrete integer values (IEEE, n.d.)
- discrete numeric classes derived from continuous numbers
- a continuous number rendered as discrete digits e.g. by 7-segment displays[11]
- a device or service that can electronically process discrete data
- a process, organisation or culture that is enabled to use the above
- modern and trendy (i.e. a promotional term devoid of technical meaning)

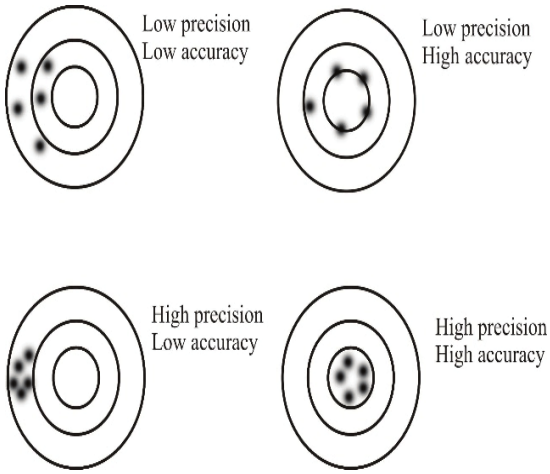
The unqualified term "digital" can thus obfuscate meaning, as shown at Section 6.3.

[11] Numitron® tubes were used since 1930s in valve radios, far from the current use of the term "digital"

2.7.2 Precision, accuracy and resolution: tool performance

These key concepts are analysed at Section 3.1, but are previewed here.

The two components of device error are established in metrology as Precision and Accuracy. These concepts are shown in the classic "target" diagram:



Precision = shots in small group
may not be centred

Accuracy = shots are centred
may not be grouped

Figure 2: target diagram

© Antonije-Onjia ... Representation-of-precision-and-accuracy-as-shots-on-the-target

The meanings in this visual metaphor do not usually require explanation in language[12]. But some of these are expanded here:

2.7.2.1 Precision

This is the certainty due to lack of random errors of a set of measurements. It includes the concepts of *coherence*, *repeatability* (in serial measures) and *reproducibility* (in replica scenarios). Precision can be abstracted to: Is it any good?

For standard numeric data, the precision of test results is easily calculated with basic statistics, to give for instance a confidence interval, means, standard deviation, skew, or preset value comparators such as "normal", commonplace in healthcare practice.

2.7.2.2 Accuracy

This is the certainty due to lack of systematic error, such as bias, non-zero-set or non-linearity. These inaccuracies are usually quantifiable, though the term "bias" is often also used for qualitative data. Accuracy can be abstracted to: Is it any use?

Accuracy also implies *traceability to a reference standard*. This in physical science such as engineering is usually quantitative, but in other science is often qualitative i.e. semantic definitions or other textual statements such as definitions, specifications, citation of state-of-the-art or prior art, regulations or statute.

[12] This suggests this visual model has more meaning than text - as do many diagrams

For qualitative data, accuracy can also be considered in non-language cognitive modes, such as visual pattern matching, which are even less quantifiable. Thus for qualitative data, the term “veracity” is used here for a qualitative “truth” of meaning, such as may be rendered in modes other than language.

Thus accuracy may be termed “trueness” if quantitative as in engineering, while for qualitative accuracy (correctness or truth) “veracity” is used here.

The target diagram shows three further analogies for tool performance:

2.7.2.3 Resolution

This is the most detail it can detect, shown as the smallest visible size of shot.

Both precision and accuracy have maximum values set by the tool’s resolution.

2.7.2.4 Scope

This is the extremes of range it can render, shown as the dimensions of the target.

2.7.2.5 Power

This is the rate of processing a quantity of shots, capacity for which is implied in the integrity of the target. Maxima for both total shots, or their rate, are physical limits.

What are the implications for mHealth devices, which feature high resolution data of high certainty - but for which healthcare veracity is not yet generally accepted?

2.7.3 Error / Certainty and Validity

“Error” is used here as synonym for low precision or accuracy, either alone or together, and “certainty” for its converse. Both types of error are partly quantifiable, and such error can be carried as metadata in a measurement system.

This terminology and knowledge model apply to not only the native data but to their derivatives (BIPM Joint Committee for Guides in Metrology, 2008).

The errors in the data thus also aggregate with each subsequent derivative, due to the principle of Propagation of Error, discussed further at Section 3.1.2.

Certainty of measurement is, however, only one component of Data Quality. In healthcare this has conventionally been described in the CARAT model, such as used by the *Practice Team Information Data Quality Model*, (2012). However, those features for high-quality data may be mapped here to correspond to those more general concepts (on the right) from the paradigm of metrology.

Completeness	precision
Accuracy/Validity	accuracy + precision
Relevance	accuracy (trueness / veracity)
Accessibility	resolution
Timeliness	trueness / veracity

Thus these metrological attributes of resolution, supporting precision and accuracy, both as trueness and veracity, with their quantifiable errors, are the engineering component of this work's evaluation of data quality.

2.7.4 Trust and Data Quality: a taxonomy

In the terms used here, each user develops their own "Confidence" in data from a device with each successful use, and this accumulates with repetition. When each user's own confidence persists, they develop Trust in a device, as an emotion. This depends on the user's ability to match the provenance of data from a specific device, so is promoted by user-device pairing – see Section 3.2.6. Trust in devices old or new can thus be gained; but it is also easily lost, perhaps after a single incident due to bad data. HCW staff will also develop trust, not only from their own accumulated experience of successful use of the device, but also acquired by relationship with each user who trusts their own device. Thus

A new device acquires Trust by accumulating Confidence based on its data quality.

Trust is also acquired more formally by reference to certifications of device quality by external agencies such as Medicines and Healthcare products Regulatory Agency (MHRA) in UK, or Federal Drugs Agency (FDA) in US.

Fortunately, HCW have long experience in managing low-trust devices, as part of their expertise is to routinely check data from low-trust devices for sense and safety:

Sense is commonly checked by time-serialisation, to expose outliers, and by triangulation with other data, part of the expertise known as clinical judgment.

Safety can be assessed by the discipline of clinical safety, a specialisation of risk management.

Both rely on certainty of data, which depends on high resolution devices or tools to create data of high precision (coherence) and high accuracy (trueness or veracity). These are quantitative, except for veracity which is qualitative by this definition.

2.8 Computable data: heuristic use in healthcare

Data with the above metrological attributes are commonly known as ”scientific” and where quantifiable are known here as “computable.” What does this mean in practice?

2.8.1 Semantic classes of data

This is an overview of the broad semantic classes of computable data in the full range of Heath and Care, with fuller version at Appendix 10.1.3

Identification of the person is the essential pre-condition for any processing of Personal Data: mis-identification is catastrophic, and can be seen as 100% inaccuracy.

Condition is a broad generic term with open-ended definition: “A Clinical condition, problem, diagnosis, or other event, situation, issue, or clinical concept that has risen to a level of concern”(HL7_FHIR, n.d.-a) . The definition is intended to include all data used for an individual’s health and care regardless of origin. Since “a level of concern” is not externally referenced for veracity, it can be considered to be so by any party. The condition may be past, current or future; each of these statuses carry probabilities that are further defined by time metadata.

2.8.2 Information Models for Conditions

These are three examples at high-level using their own sub-classes of “Condition”: classic, problem-oriented and SBAR. They share very loose definitions of “Condition”, and that they conclude with a plan.

2.8.2.1 The Operational Paradigm

This suggests a more heuristic paradigm for computable data in healthcare records: the essential function or operational purpose of processing is to *Create a Plan*.

A good plan should use data of high certainty, to minimise further processing by HCW for sense or safety.

Certainty is increased by use of computable data with known errors, and repeated confidence in data creates trust in the device.

Note some data can only be processed by HCW, such as narrative text, and some only by computer e.g. data of high volume or variety, mathematical transforms. (Their co-processing was discussed at Fig.1 at Section 2.7)[13]

[13] Computability is not always required: when the context is direct communication the interaction of 2 trained human minds can achieve high veracity of meaning transfer (Ewan Davis, personal comms)

2.9 Validity of computable datatypes

Condition is subclassified here to show different features of computability and errors.

2.9.1 Direct Data obtained from that identified person

2.9.1.1 Symptoms

These data are self-reported, and are much valued for speed and simplicity, and their accuracy, as healthcare veracity, is generally assured over centuries of use, and specifically by its provenance from that person. However precision can be poor, as a person's quantification of their symptom is limited by poor resolution of language to not more than single-digit, and the translation by HCW from verbal report to computable data are unreliable.

"Patient-reported Outcome Measures" (PROMs) are a recent structuring of the symptom datatype for data capture using preset questionnaires (Sherlaw-Johnson et al., n.d.) for direct data entry by the subject. Their veracity is now established, but as for most symptoms, resolution is poor, so certainty is less. As with other ordinal scales, certainty may be good enough for serialisation and user pairing – see Section 3.2.6.

2.9.1.2 Observations

These are defined as "Measurements and simple assertions made about a patient, device or other subject." (HL7_FHIR, n.d.-b) An observation is a description of an entity, as a response to a query of status, of two general types

- a qualitative value, commonly a simple binary assertion such as present/absent, or an ordinal rank on a scale of qualitative status
- a quantitative value, as a number bound to a Unit of Measure

A sign is an older term for an observation made by someone other than the subject, and can be "elicited" by their action, such as heart rate after doing a set exercise. The distinction between symptoms, signs and observations is less useful for data acquired by devices operated by both subjects and others.

This datatype includes all data that as continuous measures are easily computable, so traditional features such as weight, pulse and temperature. For example, the counting of pulse by direct palpation for only 15 sec is neither precise nor high-resolution, so its derivative (by multiplication by 4 to give beats/minute) is likewise - as the short sample is low resolution so at high risk of missing irregularities.

However, observations of the "simple assertion" type are usually text, so methods to quantify this for computability are discussed at Section 3.2.4.3.

2.9.1.3 Test results

These are observations made with technology, not “simple assertion.”

2.9.1.3.1 Labs

The technology is remote from the subject, so tests require to be conducted on samples (blood sciences, microbiology), or for the subject to travel to the technology (imaging). Off-site transport and result reporting create delay, and the burden of managing transit risks to the sample or its results.

They show a similar range of data types as observations e.g. biochemistry is usually continuous, microbiology is usually text, both coded and freetext, and imaging reports are usually semi-structured text accompanying an image.

However, the more automated the device, the more technology can be embedded within it for ease of operation directly by the user, leading to

2.9.1.3.2 Point of Care Testing

These tests are local to the subject, using new technology to eliminate the above offsite risks. Portable, battery-operated versions of these devices are now suitable for patient use, using new transducers for robustness, low power, integrated signal processing (Bhalla et al., 2016), digital processing, digital display to eliminate user error in reading analogue scales (imprecision), local processing for on-device trend analysis, and some internet connectivity.

These upgrades have been applied to blood pressure devices, pulse oximeters, weight scales, thermometers, glucose monitors etc. so gaining mHealth functionality, so are partly considered here.

2.9.1.3.3 Mobile health “mHealth” Devices

Smartphone inclusion in the Personal Area Network of the sensor device expands device processing, power management and connectivity, and adds user control.

Mobile phone networks are now commonplace, and other technologies for Wireless Wide-Area Networks (WWAN) include

- Mobile phone network connectivity embedded in sensor device
- IoT systems e.g. Google Mesh®, Amazon Sidewalk®, and LoRaWAN networks for location and activity, offering 10x lower power consumption (Fargas & Petersen, 2017)

2.9.2 Derived Data by processing of Direct Data

These new data sources increasingly use on-device processing to add meaning to data too complex when presented native to the user. The data may be rendered as

numbers, or further processed to display classes shown as text, diagrams or symbols, of which the simplest is the time-serialisation of values on a graph. Some examples of graphical displays or class outputs are at Section 10.3 and their derivation is discussed at Section 4.1.11.

Direct Data in healthcare practice includes both qualitative/narrative and quantitative/numeric datatypes, which are variably computable i.e. capable of machine-processing to form derived data, as discussed here.

2.9.3 Group Data inferred for that person by membership of a group

This is collated aggregate data with a probabilistic likelihood of relevance to an individual due to their group membership e.g. men are taller than women. This is a non-exhaustive list of such ecological conditions, of most relevance to Health and Social Care, with their external references for veracity:

1. Demographic
    a. Deprivation defined in geographic areas to collate similar features
    b. Ethnicity defined in racial, cultural or biological groups
2. Housing status
    a. Occupancy defined in simple numbers, or as text
    b. Social / private rented, owned defined by legal criteria
3. Social
    a. Occupation defined by economic criteria

For the whole group when defined in its academic context, the precision of such demographic data may be high in terms of repeatability; and those external references support its accuracy, such as veracity of the construct validity type.

But its accuracy for an individual is very poor: whether this man is taller than that woman cannot be predicted by gender.

Ultimately, there may be *zero veracity for this person*.

This is known as the Ecological Fallacy (Sedgwick, 2011), and is potentially equivalent in gravity to a mis-identification error, per Section 2.8.1 . This is well known as a truism and as a paradox: some of the greatest individuals have defied their deprived origins.

This is problematic for the use of group or ecological data in individual Risk Predictors: the group probability obscures a range of accuracy values for individuals from 0 - 100%.

Such a wholly inaccurate or null result cannot occur with conditions for which data are directly measured [14]. Nor can such a total inaccuracy apply to any other direct data obtained from a correctly-identified person, except in case of unrecognised device failure.

This ecological error has a statistical probability at group level, so is essentially computable as it can be analysed in metrological terms as the Probabilistic Model of Uncertainty, so can be carried in metadata and co-processed. However, for an individual it is not just another data quality issue similar to those for direct data: a potential inaccuracy of 100% greatly exceeds the metrological errors of modern devices, and becomes an unacceptable category error - see Section 5.1.3.4.

2.10 MHealth device development

2.10.1 Device classification glossary

device	an object that acquires and transmits data, including its software
mobile	a device that is self-powered and portable by one person
wearable	a mobile device directly fixed to the body, or worn in clothing
point-of-care	a mobile device that acquires data from a person or their samples
mHealth	the technological systems of mobile device(s) connected for healthcare

mHealth device the combination of wearable sensor and app on smartphone or others.

2.10.2 Overview

MHealth devices feature extensive processing power integrated into portable and wearable devices with transformational possibilities for healthcare as well as lifestyle. This overview is for context - see Section 3.3 for engineering details.

Development of these wearable technologies is an intensely competitive technological industry, which has supplied millions of devices globally. Premium brands offer product design, healthcare approval or certification, and manufacturing quality assurance that the unbranded devices are seen to lack, and so are priced accordingly. Unbranded devices are therefore now at commodity prices e.g. $25 for multi-function smartwatches, and $5 for single-function pulse oximeters. They may precede the

[14] except those that are temporary differential diagnoses and will be cancelled once excluded.

approved devices to market by the many months or years gained by not waiting for formal regulatory approval, which is a powerful business driver to mass-scale adoption.

The devices integrate many components, from transducer to processing to display to the user, and now with connectivity for onward re-use to internet services that are largely owned by device manufacturers.

Can the technology and data be used in healthcare systems too?

Multiple functions are commonly now integrated into the one battery-powered device. As a specific example, to measure activity using an accelerometer, one may select from devices with these levels of integration

I. a transducer with custom firmware and simple real-time display of step count – a common feature set for medically-certified devices (Hodkinson et al., 2019)
II. a device with custom firmware or software to time-serialise data and so identify inactivity data, with audio alarm and graphical output to one customised app – a common feature set of proprietary mHealth devices (CarePredict, n.d.)
III. a generic transducer used with user's choice of generic smartphone and software, sharing to the user's choice of global lifestyle platforms – a common feature set for low-cost lifestyle devices (*Microwear L8 Smartwatch – Specs Review - SmartWatch Specifications*, n.d.)
IV. a proprietary device integrating specific transducer, software, smartphone and global lifestyle platform – a common feature set for each proprietary "ecosystem" such as well-known systems by Samsung® Apple® etc.

So a consumer's choice of device may imply choices from across the whole range of the data journey: if one starts with the parameter to be measured, the choice of transducer within or connected to a device implies choices of operating system, user software and internet platform, with many individual and social consequences.

3 LITERATURE REVIEWS

3.1 Metrology and numbers

The concepts of error discussed here are based on the discussion by (Larrabee & Postek, 1993b) and the "Introduction" to updated definitions in VIM, the International Vocabulary of Metrology by BIPM Joint Committee for Guides in Metrology, (2008) Metrology is "the science of measurement and its application" and applies usually to physical devices generating quantitative data. However, the Introduction to VIM states that *the logical concepts are to be considered generally valid for other data in other fields*. So these logical concepts are here explored for application to data in healthcare, with special reference to quantitative derivatives of qualitative data on which so much healthcare relies, and to new data from new mHealth sources.

3.1.1 International Vocabulary of Metrology, 2008 Revision "VIM"

This third edition (BIPM Joint Committee for Guides in Metrology, 2008) raised fundamental questions about different philosophies of measurement, changing the management of measurement uncertainty from an Error Approach (sometimes called Traditional Approach or True Value Approach) to one based on Uncertainty.[15]

While the best-case measure is always implicit as the goal, the worst-case measure is usually stated as an upper limit of the absolute value of the total error, and is named "Uncertainty." The change in focus is from defining the best-case "true value" to defining the worst-case uncertainty.

In the traditional Error Approach, the measurement process produces an estimate of the true value that is as close as possible to that single true value. Random or systematic errors are defined in terms of the deviation from the true value, as "precision" or "accuracy" errors respectively, and must be treated differently since imprecision is random while inaccuracy is systematic. Further, "No rule can be derived on how they combine to form the total error of any given measurement result, usually taken as the estimate." (para 6, p viii).

In the Uncertainty Approach, however, the objective of measurement is not to determine a true value, but to permit assignment of an interval of reasonable values, by defining the measure's uncertainty. Further "even the most refined measurement cannot reduce the interval to a single value because of the finite amount of detail in

[15] This analysis also assumes that there are no catastrophic technical failures or mistakes, that would be wholly unpredictable and not amenable to statistical processing.

the definition of a measurand [16]. *The definitional uncertainty, therefore, sets a minimum limit to any measurement uncertainty."*[17]

Para 2.27 defines Definitional Uncertainty as "the component of measurement uncertainty resulting from the finite amount of detail in the definition of a measurand." The International Electrotechnical Commission (IEC), despite working with hard-engineered and precisely-fabricated electronic devices, also recognises here the significance of uncertainties in definition.

Para 2.11 defines True Quantity Value and in note 1 describes the two approaches:

"In the Error Approach to describing measurement, a true quantity value is considered unique *and, in practice, unknowable"*

"The Uncertainty Approach is to recognize that, owing to the inherently incomplete amount of detail in the definition of a quantity, there is not a single … but rather a set of true quantity values consistent with the definition. *However, this set of values is in principle and in practice unknowable"*

Thus in both approaches, the impossibility of a True Quantity Value is explicit[18]. These reference values are key to the "traceability" concept of Accuracy for physical measurement data in the form of {Number}{UnitofMeasure} (Larrabee & Postek, 1993b). They also describe Traceability as "Shifting the accuracy problem to someone else" (p679). This certainly applies to healthcare, which lacks the certainty of reference values in physical science.

These uncertainties set a worst-case value for any measurement. In this Uncertainty Approach, the concept of true value survives only to describe the purpose of measurement, and even "the adjective "true" is considered to be redundant."

The "error" approach to variation implied that there is a "true" value within this uncertainty, whereas the "uncertainty" approach makes no such implication, relying only on the statistical meaning of the quoted error. Both meanings of "error" are implicit in uncertainty values, usually rendered as {value} +/- n%.

However, the valid goals of accuracy remain: to quantify and minimise the systematic uncertainty around a measure, albeit that the pursuit of infinite accuracy is futile.

---

[16] "Measurand" means "that which is to be measured"

[17] This seems to apply across the physical universe, from the sub-atomic, per Heisenberg, to the cosmological enigmas of black holes, per Hawking

[18] However, True Quantity Values are recognised in the special cases of True Constants of Nature,

3.1.2 Propagation of Error/Uncertainty

That a device carries through to the output all errors from the input might be considered common sense, or a basic logical principle. It is commonly known as GIGO: Garbage In > Garbage Out. But it can be more logically proven by *reductio ad ridiculum:* it would be ridiculous for multiple errors to vanish by cancelling each other out. It is further formalised in the Principles of Propagation of Error or of Uncertainty (NIST, n.d.; *NIST TN 1297: Appendix A. Law of Propagation of Uncertainty | NIST*, n.d.) Because these uncertainties accumulate by mathematics that are far from simple arithmetic, the term "aggregation" is used here, since no better term is offered by VIM, which barely discusses this issue.

However, (Larrabee & Postek, 1993a) write "The concept of uncertainty is concerned with *quantifying some appropriate sum* of the random and systematic errors (i.e. *an appropriate combination* of imprecision and inaccuracy)" (my italics) and go on to detail as an example that "the imprecision and inaccuracy can be added *in quadrature* instead of algebraically."

Further, they revive the term "compounding of imprecision" which was used by traditional gauge-makers for the propagation of measurement uncertainty in their engineering craft.

This is critical for all co-processing of multiple data into user-friendly derivatives, such as by mHealth devices with on-device processing into classes, or for RP tools that co-process multiple data types including classes.

3.1.3 Precision

This is defined at para 2.15 as "closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions."

This work uses "coherence" for "closeness of agreement"

Note 1 describes how "precision is usually expressed numerically by measures of imprecision, such as standard deviation, variance, or coefficient of error under the specified conditions of measurement" which may be repeatability, intermediate (lesser) precision, or reproducibility conditions of measurement, per Note 2.

In the classic "target" diagram discussed earlier, at fig 2 Section 2.7.2, note that precision refers only to the mutual coherence of measures, whereas accuracy refers to their closeness to the reference standard, implied here as the centre of the target.

3.1.4 Resolution

This is defined at para 4.14 as “the smallest change in a quantity being measured that causes a perceptible change in the corresponding indication.”

3.1.5 Issues for healthcare

This updated VIM is helpful for the evaluation of uncertainty in healthcare, addressing:

1 Definitional Uncertainty, defined as “the component of measurement uncertainty resulting from the finite amount of detail in the definition of a measurand” is a massive problem for qualitative data in healthcare, where the data to be captured are often defined only in text, and there is no quantitative tool available.

2 True Quantity Value is an obsolete concept even in physical science, where it *had* seemed possible to quantify an absolute truth for quantitative data by reference to absolute values. For qualitative data in healthcare, there is no merit in referencing to quasi-absolute values that are only human constructs.

Wherever numerical data can be found, it could be aggregated in a probabilistic measure of error that is quantitative. However, in healthcare such combination is logically fraught, as will be discussed at Section 3.2.4. We shall also see that in healthcare, accuracy, whether as trueness or as veracity, is not a new problem, nor is it unduly problematic for mHealth devices, which create data of relatively high certainty.

3.1.6 A glossary of uncertainty

This table shows some terms for related concepts of data quality

Table 1: glossary of uncertainty with **Preferred** Terms

Domain	**Precision**	**Accuracy**	**Both**
Datatype	usually quantitative	often qualitative	
Variation	random	systematic	**error, uncertainty** (quant, qual)
General	coherent, reproducible, repeatable	truth, utility	**error, uncertainty** (quant, qual)
Statistics	variability: (standard) error	bias, linearity	
Metrology	reliability	trueness, calibration, zero-set, traceability	**resolution,** tolerance (of device) granularity; **significant digits** (of data)
Healthcare	quality	**veracity**	**certainty,** validity

These concepts are mutually independent, so data can be said to be either accurate, precise, both, or neither.

For the most certainty, both concepts are needed: precision alone may yet be wrong, however precisely; and accuracy is needed for truth, trueness or veracity of the data.

“(Un)certainty” is a more global term: where it is quantifiable “error” is preferred.

3.1.7 Validity of data

Most quantitative data are in rational numbers, but cardinal and ordinal numbers are also found also known as counting numbers, they are attributes of an item only due to its membership of a countable set. So simple arithmetic such as summing is valid only for cardinal numbers in finite sets, and ordinal "numbers" should not be processed simply, but require statistical processing (Pepe, n.d.) – see Section 10.1.2.

3.1.8 The data journey: managing error/uncertainty

At each stage of a data journey from transducer in a device to healthcare user in an ecosystem, quantitative error can be defined, and considered as metadata i.e. data about the "signal" data. Error can arise in the source, usually a transducer, or in the processing that converts one number to another, or to a class. Thus data from transducers, usually solid-state analogue devices with continuous electrical output, is either rendered as a rational number, or transformed by classification to a discrete binary class such as a high/low. This processing is often by Digital Signal Processor (DSP) integrated onto a device – see Section 4.3. The transformed data are usually presented to a user in decimal number format, but may be presented as a graphical display, with colour, audio prompts or alarms, or as classes, by their text names. A user interface can control this interaction, such as choice of display type, or to time-serialise with data stored on-device or remotely.

Data can also be directly reused in other devices, for which those human interface transforms may be irrelevant. But the classification may also be imperfect, so while they may be valid enough for the intended context of the device, once passed to another process, any error will also propagate, and aggregate with others. Thus for further processing, the unprocessed data with its native error should be preferred to part-processed derivatives such as classes. However, this is found to be seldom done in RP tools, such as eFI and NEWS2 – see Section 3.2.3.

For continuous data as rational numbers, the error metadata are easily calculated with basic statistics. For instance

- a Confidence Interval is a range of true values with 5 – 95% probability
- several types of mean, standard deviation or skew can be derived
- a minimum, maximum or threshold values may be set.

For ordinal data however, other statistical methods are required, such as those in (Morris et al., 2010), because the intervals between ranks may not be comparable, or the meaning of each rank may not be consistent, which are features of uncertainty in qualitative data.
Since ordinal data are not easily calculable, these statistical methods are beyond the simple “mental arithmetic” that can be used directly by users in the absence of computers.
Further, this *semblance* of calculability for ordinal numbers is hazardous: in cognitive terms Jacob & Nieder, (2008) show that neural processing of cardinal and ordinal numbers is indistinguishable, so there is a risk that simple “mental arithmetic” can be mistakenly applied to ordinal numbers despite their logical differences - as in several examples at Section 3.2.1.

Thus, when RP tools use ordinal numbers, they are not as simply calculable as other components as rational numbers.

To mitigate this risk of human error, it is advocated that ordinal scales never display rank by numbers, but use other well-known serial symbols, such as Roman numerals or an alphabet.

3.1.9 The meaning of certainty in healthcare

It may be argued that precision of measurement is seldom required in healthcare practice, which often creates actions from a simple discrete class derived from any measurement, such as its time-serialised derivatives “increasing” or “decreasing.” The healthcare meanings of precision and accuracy of data are discussed here:

1 a marathon runner recording his training times for a big-city marathon found they were worsening by the odd minute, and even on race day was a few minutes worse again. Just two months later he was dead: these single-digit worsening of timings had been the first sign of disease.

In this anecdote, the accuracy (as trueness) of a marathon time is not of intrinsic healthcare validity, such as for diagnosis. But we will see how user-device pairing improves its precision, and serialising shows a trend, so it acquires meaning – see Sections 3.2.5 and 3.2.6.

The mean documented all-comers marathon time is 4hrs 32 mins (*What Are the Best Marathon Times?*, n.d.) , so a 1 minute change is a resolution better than 0.4%. This is better resolution than most healthcare measures, and this is critical for the serialisation to be valid.

2 Complex Adaptive Systems are often found in the biomedical science of healthcare, and by definition are normally self-stabilising - until unexpectedly they aren’t. This status change can be rapid and occur between the intervals of scheduled monitoring. However it is also often presaged by apparently minor perturbations, for which self-correction is usual – until it doesn’t (Scheffer, 2010)

These minor changes may therefore be significant in biosocial CAS in healthcare, so important to identify as a change from noise to signal, at whatever measurement resolution can be achieved.

3 Finally, co-processing of data aggregates the uncertainty of each of its components, so the uncertainty (precision and accuracy) of each component should be minimised, and represented as metadata in further processing.

3.1.10 Classifiers in Quantitative data

Consider the observation “weight”: a value for weight acquires meaning by relationship to other weight values, and so might be named, or “classified” as underweight / normal / obese /morbid obese. This set of four classes of weight is the output of a four-way classification rule that allocates one class, with its text name, to a measurement according to four ranges of value.

Why should a continuous numeric value for the continuous variable “weight” be so classified? What is the added value of a four-way text classification with much less resolution than the native continuous weight measurement?

The purpose of a text name for a class is to reference semantic context, from which the mind infers meaning.

Further, derived classes can co-exist simultaneously, so a single datum can be classified in several ways, with each class adding meaning.
Thus a single weight may be classified *at the same time* as normal, decreasing, above one target, below another target, stable, approximate etc*.,* with meaning increasing as each class is derived.

Commonly the number of classes is only two, so such a classifier is described as binary, which in this example might be described as “More than” or “Less than” any other weight.

When the mind processes data, its capacity to process classes seems limited, and seems to default to two-way classifiers creating binary classes, which the mind then uses as logical switches. Thus in many healthcare scenarios it is a common classifier function to compare at least two measures as an increasing or decreasing time series, because this may suffice to decide on action.
So the essential purpose of classification is to transform a simple set of continuous numbers into text-named classes with added meanings, to derive actions.

3.1.11 Validity of Classes

3.1.11.1 Boundary issues

Classes may be further combined, such as “Overweight” combining the highest two of previous example’s four classes. However, by reducing the number of classes from four to two, the meaning of each class is halved.

Further, for two measures across the boundary of the doubled classes, double the resolution is needed to maintain certainty: a large step change in class can be too counter-intuitive for the mind, especially if the mind is also aware, via co-processing, that the native data were continuous – see Section 3.2.4.

3.1.11.2 Classifier issues

A perfect binary classifier would classify all cases correctly, but there is inherent error in the classification process, which is represented as:

Sensitivity: the proportion of true cases that are detected
- the undetected are “false negative.”

Specificity: the proportion of “cases” that are not valid
- the over-detected are “false positive.”

The standard framework for quantifying the errors of a classifier is the Area-under-Curve (AUC) value of the Receiver Operating Characteristics[19].

This diagram from Pepe, (n.d. pp 85, 93, 96) displays the transform from continuous to binary, which if perfect would show red and green divided by one straight line

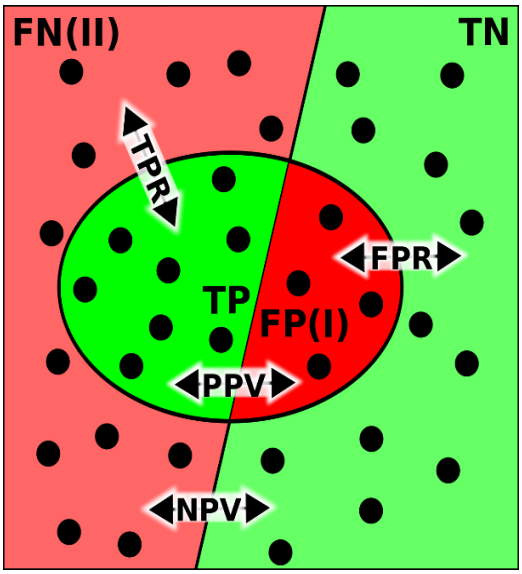


Figure 3: classification errors
Pepe: https://doi.org/10.1198/jasa.2005.s19

“The left and right halves contain instances that in fact have, and do not have, the condition.
The oval contains instances that are classified (predicted) as positive (having the condition).
Green and red respectively contain instances that are correctly (true), and wrongly (false), classified.
TP=True Positive; TN=True Negative; FP=False Positive (type I error); FN=False Negative (type II error); TPR=True Positive Rate; FPR=False Positive Rate; PPV=Positive Predictive Value; NPV=Negative Predictive Value

[19] these arcane terms originated during the 1940s for radar receiver calibration to detect aircraft.

The term “accuracy” is often used for the proportion of correct predictions i.e. both true positives and true negatives, both shown in green above. However, it is (mis)used here, as it conflates precision with systematic accuracy, so here “error” will be used for the proportion of *incorrect* predictions.

Further, only the values within this dataset are used, so this model lacks the traceability to a reference standard implicit in metrological accuracy. This absence can be critical when each classification error is unable to be assured by sense-checking against external reference, so it may remain as severe as a category error (Ryle, 1984).

Further, “precision” also has other statistical meanings in this domain that differ from that in metrology.

The error of the data is acceptable until in a specific context the derived action appears unacceptable, when confidence in the data is lost, per Section 2.7.4.

How such uncertainty is managed is discussed at Section 5.1.3.7 - it is usually repeated[20].

The more datapoints can be serialised, the more precise the trend detection. This feature is used in mHealth devices to detect trends that were undetectable by conventional healthcare, and so create meaningful new signs as conditions – see Section 3.4.2.

3.1.11.3 Classifiers in mHealth

The principle current method in mHealth for conversion of continuous to discrete is by Digital Signal Processing, by quantisation in an analogue-digital converter, or ADC, such as that integrated in the device[21].

For healthcare purposes electronic A-D conversion has high resolution and certainty. However, the veracity may be problematic, especially as values approach the boundary where classes change, as in Fig 3 above.

If a classifier such as a DSP creates discrete data from continuous, and such processing is always lossy (less than 100% accurate) what are its purposes?

- to derive from numeric data classes shown as text, so

[20] or disputed, with re-interpretation of its meaning and/or appeal to higher authorities …

[21] This use of ADC is well known to the public in audio, where consumer marketing uses bitrates of 48 / 96 / 128 bit/s etc. as a quality measure, while the imprecision of quantisation, or its noise, is specially termed “jitter”, and found only in the engineering discourse of “digital sound.”

- to apply classical logic, with the apparent simplicity to the mind of processing IS/NOT status, AND/OR, NAND/NOR logic, and IF-THEN operators
- to enable switching functions, that support actions
- to create multiple classes simultaneously from the same data, creating meaning

Thus classification assists humans to infer meaning from large datasets.

There are also engineering benefits

- to reduce the scale of large datasets substantially, which for mHealth devices may reduce power for on-device processing enough for on-device battery
- to define batches for quality assurance in production engineering, by batch selection according to performance and stress testing, as part of a Statistical Control Process (Shewhart, 2012) – see Section 6.8.

3.1.12 Creating meaning

The ML techniques of data mining are examined at Section 10.1.4: of these basic techniques , rule induction is further examined here because it is independent of prior knowledge to form a decision tree, and so *creates new knowledge.* This is the goal of ML when applied to healthcare practice; it requires computable data on a mass scale, such as shown here from mHealth devices.

Binary logic is often known as classical logic, and sometimes as Boolean[22]. An example of binary classification is Shannon's analysis of the identification of radio signal from background random noise into Presence / Absence (Shannon, n.d.) chapter 2.

The ratio of signal to noise is highly relevant to biological data processing, and is analogous to the clinical process of detecting meaningful data as if signal among irrelevant clinical noise.

So here again, *meaning is created* by binary classification into two classes, labelled "signal" or "noise," of the data that were originally continuous when measured as baud.

Thus, as so much bioscience data are of a continuous nature rendered by rational numbers, when it is co-processed with other data for a Risk Predictor tool, to preserve certainty we should

- process continuous data in its native continuous datatype of rational numbers

[22] though Boole's work on probability was to develop it as a continuous arithmetic function, as was that of Bayes and Venn.

- use classifiers only for those processing tasks where
    - humans must process or interpret
    - device power cannot process native data volumes
    - rule induction can be applied to create new knowledge
- retain the native continuous data as metadata for subsequent reuse e.g. co-processing into RP tools
- co-process the error/uncertainty of quantitative/qualitative data as metadata
- co-process data of different types using the different logic they each require.

3.1.12.1 Veracity as Qualitative Accuracy

Narrative data can be understood variably by users according to their linguistic expertise, and to the semantic (lexical and grammatical) quality with which the narrative was composed. Such analyses are beyond this scope, but it may be suggested that certainty of transfer of full meaning is impossible to calculate.

Numerical data have some metrological issues as above, but the meaning of the measurement itself is unambiguous, being defined by its units and its name, which both calibrate and define its context. When the clinician supplies their mental understanding of context to co-process it with the data, they add information to their knowledge model, to create “healthcare veracity.”

Derived data are created by formal algorithm from continuous data, so their error should propagate from its component continuous data through its derivations into classes. However, as the purpose of co-processing derived data may be for another healthcare context, healthcare veracity may need to be established again.

3.1.12.2 Validity: co-processing Precision and Accuracy

Precision is usually expressed as quantitative data, so can be mathematically co-processed, as can quantitative accuracy.

However, when accuracy in healthcare is qualitative, as veracity, it can only be rendered as those discrete classes that can be defined by preset text, and often ultimately into only two, such as present/absent. For example, fatigue, mood or pain are not reliably quantifiable even as text -see Section 3.2.3 - so may often be classified in binary e.g. present/absent or increasing/decreasing. Such time-serialisation to form a trend is much used in healthcare practice.

Accuracy improves when data are “scored” by the same person using the same tool, which improves internal validity by eliminating inter-user and inter-device errors.

A key feature of mental co-processing is that the several processes not only appear to be simultaneous but are co-ordinated by the mind, so that the output of “co-processing” is more valid across several processing threads. This co-ordinating function depends on time-stamping, as discussed by Yen et al., (2016) in a HCW context where “multi-tasking” is considered to be very rapid task-switching. Further, it is not easy, imposing significant cognitive drag, experienced as mental fatigue.
Further, co-processing the veracity of a classifier trend needs another judgment in the mind of whether the certainty of the classification supports the trend.

There is clear benefit in transferring this co-processing from human to machine.

We have seen at Section 2.7.1 that intrinsic time data can be added to any data, even if singular, by timestamping as metadata, or by converting to frequency in units of measure. For instance, while considering SpO_2 data in direct care, the HCW will include their timing with other clinical factors, and their confidence in the data provenance from any one device, or even their trust in that device type, per Section 2.7.4.
Yet every time its data are successfully co-processed, confidence and trust will accumulate, though hard to quantify.

There is no neat solution to the difficulty in understanding co-processing, which is essentially because only the human mind can co-process different datatypes.

But there are mitigations for co-processing data in healthcare when it is low quality. In this, HCW have expertise, assisted by years of training to support “clinical judgment.” The next section shows how this clinical expertise is applied to data of dubious certainty, in examples of RP tools.

3.2 Risk Predictors: frailty and other prototypes

The concept of frailty is well understood in public discourse, yet in health and care "definitions of frailty abound, defining frailty as synonymous with disability, comorbidity, or advanced old age" (Fried et al., 2001). Clinicians have long struggled to define it in biomedical terms: as Perna et al., (2017) write "The main characteristic of frailty is a decrease of the reserves in multiple organ systems. The distinction between age and frailty appears to be so blurred that it is widely held that everyone becomes frail when they grow old … physicians have long used the term "frailty" to characterize the weakest and most vulnerable subset of older adults. However, 'frail' does not mean comorbidity or disability …"

"Vulnerability" here links to the concept of instability of status, with risk of sudden deterioration due to loss of biological reserves or failure of homeostatic mechanisms. The feature of sudden, even catastrophic, status change is found in other health & care domains, such as critical care (Flaatten & Clegg, 2018). Scheffer, (2010) shows that in Complex Adaptive Systems, minor fluctuations are a characteristic feature as they approach a rapid non-linear change of status.

3.2.1 What is a "Frailty Index" (FI)?

A Frailty Index is a set of data used to compute a Frailty Score, which is a number of predictive value specific to an outcome in the specific health and care context of frailty.

All such clinical indices are cited from here now as Risk Predictor tools.

As frailty is a continuous rather than a binary variable (Hubbard et al., 2020) there are many models used to record various frailty components and to quantify it, in two general paradigms(Chao et al., 2018):

3.2.1.1 the Burden or Cumulative Deficit (CD) model

These models, such as by K. Rockwood et al., (2007) are large datasets extending up to 70 items, originally intended for use after a live update by Geriatric Clinical Assessment (Turner & Clegg, 2014). This is good practice in Direct Care, but cannot be done for Indirect Care, or Secondary Use of existing data.

In the original Rockwood CD model, "each (deficit) has an equal weight in mathematical modelling of the frailty index." (Clegg et al., 2013)

This was recognised even then as clinically implausible, but was needed for ease of live mental calculation by a clinician.

A Clinical Deficit may be any type of Condition (HL7_FHIR, n.d.-b) so the term "deficit" is misleading. However, where data are absent, as often, it is due to interoperability failure i.e. although data may already exist somewhere in healthcare systems, the location is unknown, so it is inaccessible from direct care per Section 2.6.2.
While "evidence of absence" may be highly valid, "absence of evidence" has zero validity, and is another instance of category error, as summarised at Section 5.1.3.4.

3.2.1.2 the Biologic Syndrome or Phenotypic (PT) model.

These models are small datasets in which data points are specifically constructed for the purpose and must be assessed and entered by the clinician - see Section 10.1.5.
For example, the original FRAIL index per (Fried et al., 2001) uses only five items with binary values, so is of low resolution. Further, only weight is commonly already available: the others must be created, adding to HCW workload.
Analysis of the comparative review of 8 such PT type of frailty indices by Theou et al., (2013) shows that these RP tools are intrinsically lossy of information: in the Groningen scale, a binary value for "Fitness" is by self-report of frequency of four modest activities, but these activities are not individually recorded, losing resolution.

CD models are increasingly feasible as electronic records become richer in the routine recording of "Clinical Deficits" as conditions. For example, VisionHealth software also includes an automatic calculator based on the Clegg 36-item CD model (Frailty in Elderly People, 2013) (Vision, n.d.) (E-frailty-index-brief-guide.pdf, n.d.)
Since it is also used in by EMIS® and TPP® software by GPs across the UK, it is considered to be valid (have semantic veracity) across UK Primary Care.
DHI's work on frailty includes the RP tool known as Advanced Risk Modelling for Early Detection (ARMED, n.d.) which can be considered a CD type of RP tool as it is updated automatically with serialised mHealth data; it is now validated for use in the social care setting of care homes (Smales, n.d.).
The widely-used Dalhousie/CHSA Clinical Frailty Scale (Kenneth Rockwood et al., 2005) is a combination model: a 7-point scale uses clinical judgment to assess up to 70 deficits. This scale's data are simpler than that for other CD models, and has been validated as reliable for mortality.
However, it has been found to be not reliable as a Risk Predictor of mortality in COVID-19 (Miles et al., 2020).

3.2.2 How is computable data used in Risk Predictor tools in health and care?

"Risk Predictor tool" is a new generic term for scales, scores and indices such as these frailty indices. They originated as simple arithmetic constructs of different types of data, in which the clinician uses "mental arithmetic" to predict risk in live, often urgent, direct care settings. Simplicity of use is a major constraint, as these were developed before any widespread use of machine processing.
The predictive value is not always obvious in direct care, where it is intended to add value to the direct use of its component data, which would use the RP tool just as a simple checklist.

Yet the veracity in direct healthcare of "a simple checklist" is very high, where it can function not only as a simple tally, but to sequence data in priority. There is even higher value in a canonical checklist, being a comprehensive list of all data options. (*The Checklist Manifesto | Atul Gawande*, n.d.) The added value of RP tools may be higher on aggregate use on a group or population level.

Sharma et al., (2021) describe the poor interoperability between devices creating data from clinical observations, and EHRs, in the highly mechanised Critical Care domain. There, HCWs function as "human interoperability bridges"[23] to work around the lack of technical interoperability for data presented to them visually by many device screens and the EHR.

Outside these highly mechanised domains, combining disparate data into RP tools renders the derived data more usable by HCW in more general health and care.
Here, mHealth devices also may require their users to function as a "human interoperability bridge" to bridge the gaps between devices and other displays.
It is still rare for machines to combine this data quantitatively, so the expertise of the clinician is needed for this.

[23] Such as "human screen-scraping" by viewing and collating multiple displays at once

3.2.3 Risk Predictor tools

Twenty examples of three main types are shown in Table 2

3.2.3.1 Scale

A set of numerical measures of one item, from which one is selected as a score. When used for text descriptions, this ranks them for healthcare significance in a semantic progression, so marked by a ranking symbol, usually a number.

Conversion of serial text descriptions into rank order is widely used in mental health, to quantify the selection of a qualitative statement. This ranking number may be one of the items input for calculation in a complex RP tool, as below.

3.2.3.2 Score

			Datatypes
1	simple:	one number selected from a simple scale, as above	ordinal
2	complex:	a number calculated from any combinations of	
	•	a simple tally of items	cardinal
	•	values from a continuous observation, measurement or test result	rational
	•	numbers selected from other simple scales, or their derivatives	
		○ numbers from scales (see “index”)	rational
		○ key values e.g. max/min, rate of change, specific values	rational
	•	classes derived from these continuous numbers e.g. first quartile	discrete
	•	numbers to weight these e.g. by frequency, duration, veracity	rational

3.2.3.3 Index

a score whose calculation includes fixed denominator values for range, expressed as a percentage (bleeding, glycaemic, therapeutic) or as a number (body mass, eFrailty).

However, please note these semantic confusions:

Both “Scale” and ”Index” terms are also used to describe the whole tool that calculates a Risk Predictor score: this semantic mismatch is shown by * in Table 2 below.

“Score” is also used as a verb for the selection of a number on a scale, score or index!

Table 2: Risk Predictor tools: Twenty Scales, Scores and Indices

These are in alphabetic order, except for *frailty types shown together*

Name	Nominal Type	True Type	No. items, or number	Max range	Data type
Apgar	Score	=	5	9	N+T
Bleeding	Index	=	Num	%	N
Body Mass	Index	=	Num	biology	N
Cephalic	Index	=	Num	%	N
Epworth *	Scale	Score	8	24	T
Erythrocyte	Index	Set	3 x units	n/a	N
*Edmonton Frailty **	Scale	Score	11	17	T
*Dalhousie / Clinical Frailty **	Scale	Score	9	9	T
*Falls Risk Assessment Tool **	Tool	Score	5	5	T
*FRAIL **	Scale	Score	5	25	T
Clegg eFrailty Index	Index	=	36	36	N
General Anxiety Depression GAD7	Score	=	7	21	T
Glycaemic	Index	=	Num	%	N
Glasgow Coma Scale *	Scale	Score	3	15	T
Index of Relative Need 2	Index	=	19	19	T
MM Personality Inventory 2	Score	=	567	567	T
Natl. Early Warning Score 2	Score	=	6	20	N+T
Palliative Performance Scale v2	Scale	=	11	10x10	T
Patient Health Questionnaire 9	Score	=	9	27	T
Therapeutic	Index	=	Num	%	N

Legend:

Nominal type	common type name
True type	correct type name if varies
*	mismatch of above
No of items	resolution
Max range	scope
Datatype	Numeric N or Text T

3.2.4 Metrological discussion of examples

3.2.4.1 Precision

This is the data quality feature of minimum random error.

3.2.4.2 Accuracy

This is the data quality feature of minimum systematic error.

This is the minimum error around the “true value” when expressed quantitatively, or if in qualitative terms, deviation from “the truth,” for which “veracity” is used here.

Maximum values for both precision and accuracy are set by the measuring tool’s resolution. In Table 2 above, it is indicated by the “No. items” value. For example

- a FRAT score of 3/5 is to one part in five, since each of 5 items is reduced to binary classes of presence/absence – a low resolution of one digit on a small value range of 5.
- an MMPI2 score of 273/567 implies a resolution to three digits on a large value range. This is ambitious for a mental health measure.

However, resolution is only one component of certainty: the MMPI can show wide inter-observer error, an inaccuracy that counters the certainty implied by high resolution.

The precision of RP tools is reduced by inaccurate processing of mixed datatypes: thirteen of the twenty RPs listed here (shown as T) use text statements rendered in order with numerical symbols which are simply added together as if rational numbers. This creates errors that are essentially unpredictable, as if random, so are considered here as issues of precision.

3.2.4.3 Veracity in narrative data

The meaning of narrative data can be variably understood according to the expertise of sender and receiver, and to the lexical and grammatical accuracy with which narrative text is composed; but the science of linguistics is out-of-scope here.

However, note that the thirteen RP tools that use narrative statements risk all the semantic ambiguities intrinsic to the use of text. For example, in the GAD, one item has the simple text “trouble relaxing” which can be interpreted very variably.

For numerical data (data type N) the meaning of the number is not in question, being defined and quantified by its units and its context, and formally by the name of its unit, which in the basic sciences is a permanently traceable external reference.

Inter-observer error is a systematic error which is a major factor in healthcare veracity – see ”fatigue” at Section 3.2.4.

Further, per Section 3.1.2 these errors will propagate (NIST, n.d.) in data derived from continuous data by algorithm, which is explicitly reproducible. So also should be the propagated error, persisting between users. RP tools with these and other error types are discussed here.

A **Body Mass Index** is a “simple” calculation on height and weight: BMI = kg/m^2. This derived number is meaningless until compared to age-sex standardised values. In practice, the mental calculus is not so simple, requiring squaring then division of 3- or 4-digit numbers. Most users avoid this, by directly looking up the height and weight printed on a grid diagram showing ranges of the index with precalculated classes (normal, obese etc) shown by colour, so rendering a 3D model on paper.

BMI Chart

WEIGHT lbs	100	105	110	115	120	125	130	135	140	145	150	155	160	165	170	175	180	185	190	195	200	205	210	215
kgs	45.5	47.7	50.0	52.3	54.5	56.8	59.1	61.4	63.6	65.9	68.2	70.5	72.7	75.0	77.3	79.5	81.8	84.1	86.4	88.6	90.9	93.2	95.5	97.7
HEIGHT in/cm	Underweight					Healthy					Overweight					Obese				Extremely obese				
5'0" - 152.4	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
5'1" - 154.9	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	36	37	38	39	40
5'2" - 157.4	18	19	20	21	22	22	23	24	25	26	27	28	29	30	31	32	33	33	34	35	36	37	38	39
5'3" - 160.0	17	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	32	32	33	34	35	36	37	38
5'4" - 162.5	17	18	18	19	20	21	22	23	24	24	25	26	27	28	29	30	31	31	32	33	34	35	36	37
5'5" - 165.1	16	17	18	19	20	20	21	22	23	24	25	25	26	27	28	29	30	30	31	32	33	34	35	35
5'6" - 167.6	16	17	17	18	19	20	21	21	22	23	24	25	25	26	27	28	29	29	30	31	32	33	34	34
5'7" - 170.1	15	16	17	18	18	19	20	21	22	22	23	24	25	25	26	27	28	29	29	30	31	32	33	33
5'8" - 172.7	15	16	16	17	18	19	19	20	21	22	22	23	24	25	25	26	27	28	28	29	30	31	32	32
5'9" - 175.2	14	15	16	17	17	18	19	20	20	21	22	22	23	24	25	25	26	27	28	28	29	30	31	31
5'10" - 177.8	14	15	15	16	17	18	18	19	20	20	21	22	23	23	24	25	25	26	27	28	28	29	30	30
5'11" - 180.3	14	14	15	16	16	17	18	18	19	20	21	21	22	23	23	24	25	25	26	27	28	28	29	30
6'0" - 182.8	13	14	14	15	16	17	17	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27	28	29
6'1" - 185.4	13	13	14	15	15	16	17	17	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27	28
6'2" - 187.9	12	13	14	14	15	16	16	17	18	18	19	19	20	21	21	22	23	23	24	25	25	26	27	27
6'3" - 190.5	12	13	13	14	15	15	16	16	17	18	18	19	20	20	21	21	22	23	23	24	25	25	26	26
6'4" - 193.0	12	12	13	14	14	15	15	16	17	17	18	18	19	20	20	21	22	22	23	23	24	25	25	26

Figure 4: BMI printed chart
www.MyFitnessPal.com

Reference to this is quicker than re-entering the data into a calculator!

Immediate calculation by machine is now commonplace, but the simplistic nature of the index has not been updated to use this, and its healthcare veracity is increasingly questioned, for example in comparison to girth for abdominal obesity.

B The "**Erythrocyte Index**" is not a calculated number at all, being only a compound plural noun for a set of three separate test results that clinicians interpret together "in the mind." This is an extreme example of the concept of co-processing: despite using quantitative data, it uses only "clinical experience" to evaluate the numbers, but no agreed algorithm that could be machine-processed.

3.2.4.4 Resolution, precision, and accuracy in

C In the QMUL **FRAT** "**tool**" (Nandy et al., 2004), only 5 items are given a binary value for the presence or absence of a "deficit" (the term for condition in this context). It includes one that was originally continuous: polypharmacy is the presence (binary = 1) of >4 drugs, but not a continuous tally of their number, which would preserve resolution. The reason for setting this classifier boundary at 4 is not discussed. Further, the polypharmacy definition in Clegg's eFI score (Clegg et al., 2013) is set differently at 5 drugs, from a preset list of drugs, and also not discussed.

D In Clegg's **eFI score** the full range is divided into 4 quartiles described as fit <12, mild 12 -23, moderate 24-35 or severe >36% of max score. These four derivatives of the scale are often used in place of the native scale value, and are intended to align with the same-named categories of the Rockwood CFS (Clegg et al., 2016)

Table 3: eFI classes from continuous data

Rockwood CFS name	V fit, well, managing, vulnerable	Mild	Moderate	Severe	Very severe, terminal
Clegg eFI	<12	12-23	24-35	>36	-

However, for the implicit benefit of simpler logical processing of the data, there is no published validation of this claim of equivalence, and these groupings of values incur a loss of resolution, and introduce non-linear or step changes to a continuous measure. However, if machine co-processing could reuse the native continuous data of the sub-score, the resolution of this original data could persist.

E In the NEWS2 score "A score is allocated to each parameter as they are measured, with the magnitude of the score reflecting how extreme is the parameter from the norm. The score is then aggregated - and uplifted by two points for people requiring supplemental oxygen to maintain their recommended oxygen saturation."

(*National Early Warning Score (NEWS) 2 | RCP London*, n.d.)

Seven items (one is pre-selected from two scale options for SpO_2) are ordinal numbers allocated to continuous variables. Three of these (respiration, pulse and temperature) are very non-linear, and so are allocated non-linearly to grouped values to represent their clinical meaning. These ordinal numbers are then summed according to logic in Guidance Notes e.g. selection of one of two SpO_2 scales. However, again this summation processes ordinal numbers as if they are rational.

Physiological parameter	Score 3	2	1	0	1	2	3
Respiration rate (per minute)	≤8		9–11	12–20		21–24	≥25
SpO_2 Scale 1 (%)	≤91	92–93	94–95	≥96			
SpO_2 Scale 2 (%)	≤83	84–85	86–87	88–92 ≥93 on air	93–94 on oxygen	95–96 on oxygen	≥97 on oxygen
Air or oxygen?		Oxygen		Air			
Systolic blood pressure (mmHg)	≤90	91–100	101–110	111–219			≥220
Pulse (per minute)	≤40		41–50	51–90	91–110	111–130	≥131
Consciousness				Alert			CVPU
Temperature (°C)	≤35.0		35.1–36.0	36.1–38.0	38.1–39.0	≥39.1	

Figure 5: NEWS2 tool
https://www.rcplondon.ac.uk/file/9434/download

That a single point change on one scale is allocated equivalent clinical validity to a single point change on another is not supported in logic, a failure of design. That NEWS2 is apparently valid enough for safe clinical use depends on the practical evaluation of the whole instrument, which includes repeated guidance that users must use judgment and often obtain "clinical supervision" i.e. the clinical judgment of others to check each datum.

A binary value, such as for use of oxygen over air, is a lower resolution than the numerical % of inspired O_2 that might be used as a continuous value; but is used for consistency with the simple arithmetic criterion for calculation of the score (*Using the National Early Warning Score (NEWS2) in Primary Care. London: RCGP 2020*, n.d.) No machine is required for calculation, as intended for the emergency context for which NEWS2 was designed.

A further example of inaccurate processing is found in the modification by (Liao et al., 2020): age is added as a binary value of 3 points if age >65y, but not as a continuous variable. Yet in COVID-19, age is a strong (so semantically valid) and remarkably exponential (so precise) predictor of mortality, (Goldstein & Lee, 2020), so easily machine-calculable. But it is unsuitable for simple addition as a single value: clinicians must be alert to a score incrementing by 3 on a 65th birthday!

Bramley shows how the paper basis of the tool may be developed into a dynamic "digital" display (Bramley, n.d.) This is a front end to enter and display the underlying data calculations.[24] However, it still uses simple additions of inconsistent datatypes that are class derivatives of native continuous data that should be directly computable.

For this, as for many complex RP tools, it is not always clear how the construct adds value to monitoring of its component parts, for which it might serve more simply as a checklist, per Section 3.2.2.

The added value has been described in terms of reduced training needs for staff who require specific actions to be attached to changes or thresholds of the aggregated score, that might not apply to the same changes to one component. However, it is not explained how each specific action is better initiated by this one aggregate RP, rather than by its own specific aggregation of the relevant components. For example, increase of O_2 might be triggered by using only SpO_2 and pulse more accurately than by the NEWS2 score aggregating four further observations.

Further, multiple RP tools may be available in direct care, introducing a further problem for the HCW of selection between them. Realistically, only processing by machine can render multiple outcomes for the clinician to evaluate.

[24] NEWS2 has also now been rendered in openEHR , so is now available under open-source licence to any supplier to include in their own EHR system
https://twitter.com/bjornna/status/1451592860764147712

F The Palliative Performance Scale v2 shows an unusual underlying logic model that processes ordered text statements according to a specific logic: the sequence of these text statements as laid out on the formal published chart – see (*Palliative Performance Scale (PPSv2) Version 2*, 2001).

PPS Level	Ambulation	Activity & Evidence of Disease	Self-Care	Intake	Conscious Level
100%	Full	Normal activity & work No evidence of disease	Full	Normal	Full
90%	Full	Normal activity & work Some evidence of disease	Full	Normal	Full
80%	Full	Normal activity *with* Effort Some evidence of disease	Full	Normal or reduced	Full
70%	Reduced	Unable Normal Job/Work Significant disease	Full	Normal or reduced	Full
60%	Reduced	Unable hobby/house work Significant disease	Occasional assistance necessary	Normal or reduced	Full or Confusion
50%	Mainly Sit/Lie	Unable to do any work Extensive disease	Considerable assistance required	Normal or reduced	Full or Confusion
40%	Mainly in Bed	Unable to do most activity Extensive disease	Mainly assistance	Normal or reduced	Full or Drowsy +/- Confusion
30%	Totally Bed Bound	Unable to do any activity Extensive disease	Total Care	Normal or reduced	Full or Drowsy +/- Confusion
20%	Totally Bed Bound	Unable to do any activity Extensive disease	Total Care	Minimal to sips	Full or Drowsy +/- Confusion
10%	Totally Bed Bound	Unable to do any activity Extensive disease	Total Care	Mouth care only	Drowsy or Coma +/- Confusion
0%	Death	-	-	-	-

Figure 6: PPSv2 tool
Palliative Performance Scale (PPSv2) Version 2, 2001)

This navigation of this chart with “leftwards precedence”, then up/down, to select a PPS level, is described in Guidance Notes. Such guidance notes are used for many RP tools to improve veracity by expanding on the brief text labels with further contextual definition. They may also be used to instruct a user in how to calculate a score by “mental arithmetic” that is not-so-simple.

Guidance notes are a best effort to reduce inter-observer error in their use. But they are unable to prevent misuse: (Lau et al., 2009) describe the use by clinicians of averages of the ordinal numbers of PPSv2, which appeared to them easier than reaching a clinical agreement on a value despite wide inter-observer variation.

G The FRAIL scale is a simple scale, an ordinal number formed from 5 narrative items scored and again simply added together for a FRAIL score, even though the numbers are ordinal symbols.

The item F(atigue) is a rank scale of 1-5 for frequency from this scale of semantically progressive text descriptions: none / a little / some / most / all of the time.

The semantic progression is in terms of one-time parameter: frequency of repetition.

It is also qualitative, despite the clear opportunity to quantify a frequency.
There are other options for time-related data that might be measured, such as regularity (with subtypes), duration, or absolute date/time. Is it a snapshot status at the time of enquiry, or some ill-defined integral of the symptom, such as over the calendar day so far, or over 24hrs-to-date, or since last report, or symptom onset? Any of these timestamp data could be carried in metadata, for computability.

Further there are semantic options that might address intensity, or quality. However the latter is notoriously difficult even for this most commonplace symptom: does "fatigue" mean drowsiness, or weakness? (Asher, 1959)
The rating for Fatigue is thus interpreted by each individual to quantify their own serial readings, and it is assumed that uncertainty for each individual is internally consistent e.g. if they interpret fatigue to mean weakness or drowsiness (veracity), or if an individual's data are coherent in time, repetition, or replication (precision)

These RP tools show no systematic method beyond guidance notes to harmonise the interpretations between individual users, so inter-observer error may be high. This major source of imprecision is usually addressed in published validation of the RP tool.

3.2.5 Serialisation for healthcare certainty

Recent work on frailty has tracked serial scores for each individual and found this usage to be a more reliable predictor than either single scores or between-person scores (Stolz et al., 2020); this precision applies to four different RP tools.
In this study, data consisted of "*self-reported, mostly dichotomous health items* and the few ordinal/metric[25] items covered multiple physiological systems, and included chronic diseases, limitations in basic and instrumental activities of daily living, mobility restrictions, cognitive functioning, sensory impairment, self-rated health, somatic symptoms, depressive symptoms, BMI- deficit, and low physical activity."

This shows improved clinical certainty of a tool with inter-subject and -observer error, when some of these variables are controlled for by re-use on the same subject (if not by the same observer) – that is, by pairing the tool to each user.

[25] Sic. "Ordinal/metric" implies conflation of ordinal ranks with rational-number measures

3.2.6 Pairing of single device with single user

It is not only RP tools, but mHealth devices that can now be paired to the user. Where data are liable to inter-user and inter-device variations, due to imprecision of the device, user-device pairing will eliminate these, and so support time-serialisation.

Thus, users using only their own device eliminate inter-user, inter-device and misidentification errors, all essentially random, so precision is much improved. Further, in metrological terms, not only quantitative precision but also qualitative veracity improve with pairing of a device to each user who develops understanding and confidence with each successful use, per Section 2.7.4.

Fortunately, new mHealth devices are now economical for each user to acquire individually. The feasibility of this also depends on the care setting context:

- always, for an individual's own readings of their own mHealth devices
- usual, in domestic care settings e.g. household readings of each shared device
- seldom, in group care settings such as hospital with many users of many devices

The healthcare scenario of use for a device is defined here in six scenarios of mHealth in COVID-19 per Section 2.4.

3.2.7 Uses of Risk Predictor tools for COVID-19

A Risk Predictor tool is a computation of the clinical features that are found by clinicians to form a Syndrome, which means a set of conditions (usually symptoms + signs + tests) that commonly occur together. For adults these have been grouped into six by Sudre et al., (2020), and many further syndromes in children and in old age are now recognised.

The various syndromes of COVID-19 disease remain hard to identify and so the full diagnosis is often late. Yet this disease requires the earliest possible diagnosis, as transmission is worst from 2-3d before the onset of symptoms - see Section 3.4.1.

Its protean features in the early stages are a huge diagnostic challenge.

Other COVID-19 RP tools are underway by many global researchers, though out-of-scope here. For example, (Fiorentino et al., 2021) and 15 colleagues' are developing the RECAP (Remote COVID Assessment in Primary Care) RP tool, though it includes only one mHealth data source – for SpO_2 from pulse oximeter.

However, it also transforms continuous data into, at most, four classes, and usually two: for example days since diagnosis into (binary) < 7d, age into (binary) >65yrs and BMI into (binary) >35. This reduces the resolution of the tool.

3.2.8 Healthcare benefits of computable Risk Predictor tools

Now that more data are available electronically in routine healthcare, live systems can apply preset weightings to each datum, and reuse the data.

Different RP tools apply in different contexts, and each patient journey may cross several contexts of which many overlap e.g. A&E, Critical Care, Care Homes, planned reviews can each reuse some of the same data. Yet all this efficient reuse of data depends on its quality. So live computation can also

i. for direct care, select relevant RP tools live for each context, and
- select data entry items and render intelligently for healthcare usability, and
- reduce repetition of data collection from patients
- reduce burden of acquiring and entering new data in this pressured context

ii. for indirect care, process other RP tools that include data acquired offline.

Since most RP tools were deliberately simplified for human processing in direct care, they support human processing by “mental arithmetic” so

- rational numbers for this are limited to one or two digits
- simple arithmetic is limited to simple operators, usually addition.
- cardinal and ordinal numbers are limited to small sets.

RP tools are thus currently tools of low resolution and limited precision and veracity, and their added value over structured checklists of their component data is not clear. However, they can improve with machine processing of the native data, usually continuous and with resolution to several digits.

3.2.9 Informatics benefits of electronic Risk Predictor tools

i. manage above data entry options by evaluating quality of data already present
ii. configure RP tool to context, by matching selection of data to the context
iii. support any number of RP tools simultaneously
iv. acquire data including mHealth from connected new systems if data are interoperable. This is the purpose of RQ4.

3.3 MHealth device technologies

3.3.1 Design overview

The essential engineering components are a sensor, on-device processor, software to control and output as human-readable derivatives, connection to share to internet, and a power source. These components are considered here for the two most significant new technologies.

3.3.1.1 Motion detection

a. accelerometry, for linear motion in the mm range, and
b. gyroscopy for angular motion, both by piezoelectric sensor (InvenSense, n.d.)

Several further technologies are not yet in general use, including

c. micro-radar motion detection at sub-millimetre resolution of chest movement[26] to monitor sleep – see Section 3.4.7.
d. pressure sensors, such as in a bed mat (Nokia, n.d.)

All sensors are solid-state, commonly integrated with their own microprocessors into both wearables and smartphones, and now well established. As solid-state sensors they draw very little power, and are robust in everyday use.

Both also detect absence of motion, or inactivity.

3.3.1.2 PhotoPlethysmoGraphy

PPG technology is used for several parameters. This uses one or more visible or infrared LEDs to illuminate vascular blood flow, an optical sensor to detect pulsatile reflectance[27] from all vessels, and integrated digital components (Valencell, n.d.) to deliver:

- Opto-mechanical coupling
  - tight or stable coupling may not be tolerable – see "Wearability"
  - correct wavelengths for different body locations
  - geometry of sensor placement " " "
- convert optical receiver's output from Analogue to Digital (ADC), to enable
- Digital Signal Processing (DSP) using algorithms in firmware to
  - discard motion and other artefacts for S/N ratio (precision)
  - manage DC offset drift due to positional instability (accuracy)

[26] in Google-brand smartphones (Google, 2021), and since Mar 2021 in the Nest Hub for bedside use
[27] Except early oximeters that transilluminate where feasible e.g. on fingertip or earlobe.

- calibrate individually-paired device for skin tone error between users (accuracy)
- identify step and pulse rates, that share trends and may "crossover" in the same numerical range of 90 – 160/min (precision)
- process numeric data into classes, to add meaning (veracity)
- present data to user with good user experience

All algorithms are proprietary to the device manufacturer, though several share the same sensor and may share algorithms by the sensor manufacturer, such as (*Valencell | Virtual Patent Marking*, n.d.) PPG technology is used for

a. Simple Heart Rate, by processing mean pulse intervals (RR interval "RRi")
   - Resting HR is derived by algorithmic co-processing with motion and time
b. Heart Rate Variability, by processing variability of pulses (HRV, RRi)
   - variability analysis requires very high sensor resolution
c. Oxygen levels ($Sp0_2$), by spectrum in reflected (skin) or transmitted light (finger, ear) using the differential chrominance from dual LEDs, one visible red and one infra-red, and dynamic differential for arterial and venous vessels
d. Blood pressure, by processing contour of pulse waveform
e. Respiratory rate, by processing the pulse waveform data that shows respiratory sinus arrythmia "RSA"[28]
f. Sleep, by algorithmic co-processing with several data inputs to identify awake, light, REM or deep sleep phases, including
   a. time
   b. motion detection by accelerometry, gyroscope
   c. PPG data for cardiac status (HRV), BP, respiratory rate or SpO_2.

These new PPG-based technologies are opening up applications for everyday use, with a range of sensors offering not only RRi analysis, but also BP measurement and SpO_2. However, the strongest predictor of illness or frailty is inactivity, as detected by motion detectors.

MHealth technologies for HRV, SpO_2, BP and Respiratory Rate will not be considered further here, but are outlined with others at Section 10.2.

[28] initially known as continuous wavelet-transform technology when discovered in Edinburgh only 7yr ago by (Addison et al., 2014)

3.3.2 Wearability: location, stability and comfort

PPG consumer devices are most commonly worn on the wrist, as pioneered by FitBit in 2011[29]. However, a problem for precision of PPG data is instability of the device on the skin: a firm fit is uncomfortable so unsuitable for continuous automatic monitoring. The issues are summarised in this diagram

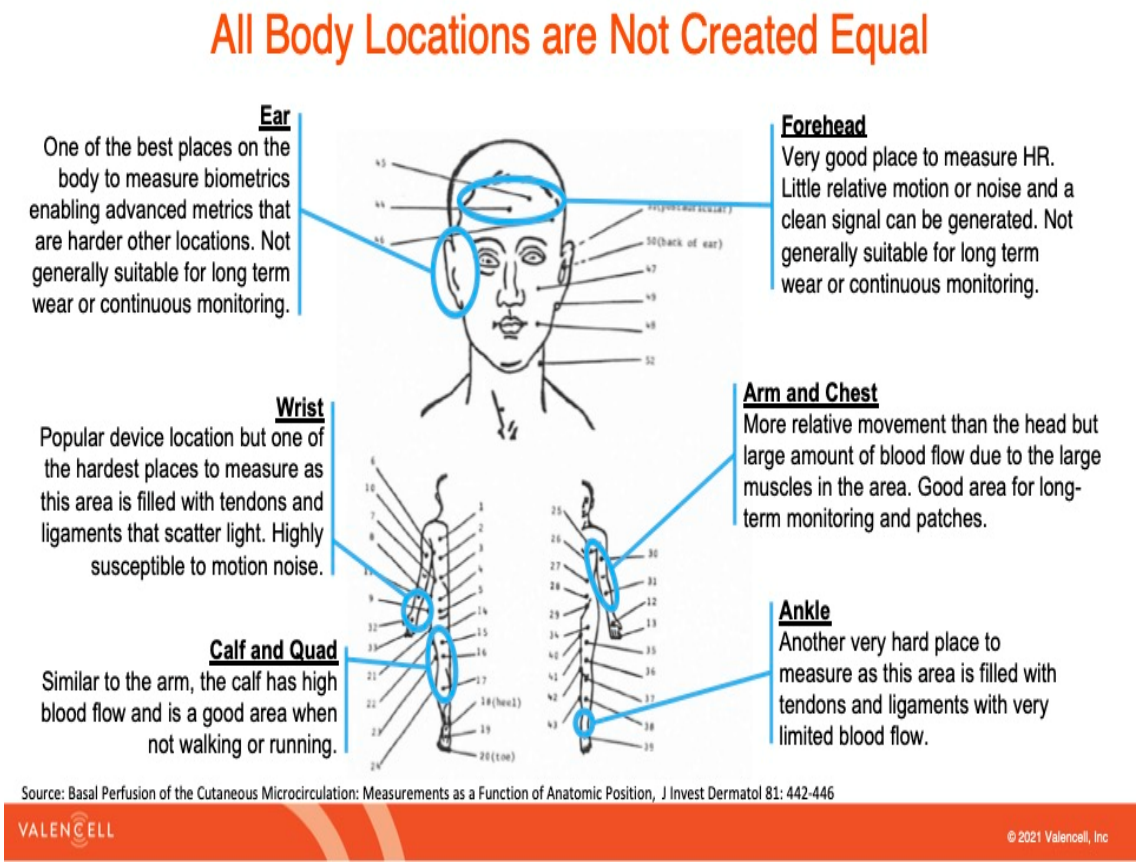


Figure 7: body locations for wearables
Optical Heart Rate Monitoring Technology: What You Need to Know (valencell.com)

This shows that the wrist is the worst place to try to fix a PPG sensor, and the best may be on the upper arm, thus requiring a dedicated device. HRV analysis from the wrist is impossible during daily activity; even RR intervals are inaccurate due to the heavy filtering needed to get just basic HR data.

In-ear fit is optimal for stabilisation of the device, but not tolerated well unless the user chooses a device combining it with earphones, or a hearing aid that is already worn continuously. Further, ear wax impaction is not a trivial problem, so it is not attractive for 24-hr use.

[29] FitBit's first device in 2009 was accelerometer only, worn in clothing

3.3.3 Design and manufacturing issues

A specialist manufacturer of PPG, Valencell, publish a series of White Papers that address these issues

- Design / Intellectual Property
    - all use cases (healthcare added to sports, lifestyle)
    - optical emitters/photodetectors, accelerometers, algorithms
        - Active Signal Characterisation" [30] <1:1000 of reflected light represents circulatory signal – it is swamped by noise.
    - PPG for HR and SpO_2 since early 1980s now miniaturised to finger / earclip devices
        - Sporadic motion artefact removal for professional devices
    - Continuous-motion-tolerant PPG is step change in last 5-10y, to enable being wearable everyday.
- Validation testing for standard and boundary cases: sunlight, skin tone, sweat
    - Skin tone a substantial problem, remarkably only reported in 2020 (Sjoding et al., 2020): "Black (hospital) patients had nearly three times the frequency of occult hypoxemia that was not detected by pulse oximetry as White patients." This is especially alarming as COVID-19 features "silent hypoxia" for unknown reasons, yet many rely on pulse oximeter alerts for such critical actions as calling an ambulance.
- Manufacturing mass production, automated testing, quality control.

Manufacture at mass scale is subject to Quality Assurance issues such as manufacturing tolerances and batch selection. McMahon, (2013) describes in detail the calibration of oximeter output vs SpO_2 in an R-curve that is flashed into the firmware of that model, with the highest specification applied to the most expensive FDA-certified devices for professional use in hospital settings, where absolute accuracy is required to use interchangeably for multiple subjects.

- Marketing see Section 6.4.1 for issues around Suunto® questionnaire

Why Accuracy Matters: Why Accuracy Matters in Biometric Wearables (valencell.com) addresses this (though uses "Accuracy" where this work would suggest Precision)
For example, Samsung introduced in 2015 SpO_2 on a smartwatch using dual-LED PPG, but withdrew it after 3 mo. due to failure to achieve FDA approval as a Medical Device,

[30] similar to Active Noise Cancellation for headphones, but 2 orders of magnitude more severe.

as was a version integrated on their premium Galaxy phone. It is reported that a new FDA-approved version will soon be released, and that Apple Watch 6 will also feature FDA-approved SpO_2 measurement. The value of FDA certification is to the user and HCW perception of accuracy, as evaluated at Section 6.3.

Meanwhile other **branded** watches offer non-approved SpO_2 levels [31] and many other **unbranded** and non-approved products are available: for example the L8 smartwatch from an unnamed supplier in Shenzhen uses the nRF52832 chipset from the widely-used Nordic range, whose specifications and SDK are openly published (*NRF5 SDK - Nordicsemi.Com*, n.d.). It uses dual-LED PPG to measure Blood Pressure and SpO_2 as well as Heart Rate, which it also processes into a Resting Heart Rate - see Section 10.3.

Note while there is a standard BLE protocol for sending HRV data, many manufacturers use proprietary e.g. FitBit, Garmin, iThlete, apparently for better performance – but locking the user into proprietary software.

Sensor technology is fast moving, and Valencell report that for HRV their PPG on a dedicated chest or upper-arm device is now as good as contact technology. Their note from 2017 shows some of the issues now addressed:

"Valencell pointed out that the previous sensor in the Rhythm+ is from 2012 (!), whereas the sensor in the Rhythm24 is now their latest gen. They noted that "since then we've upgraded the MCU for lower power consumption, upgraded the LEDs in the sensor, made countless improvements to the signal processing algorithms, and made the firmware field upgradeable for future enhancements to sensor"

Note that this PPG device dedicated to HRV has a low battery life around 24hr.

[31] Examples are Garmin Venu®, Huawei Watch GT2e®

3.3.4 Battery power: duration and endurance

The size of battery directly determines the duration of use of the device, the size of which is thus also determined. So small devices, such as in-ear PPG sensors, or finger-ring devices such as Oura® may not deliver duration, such as for five days required for monitoring contact cases in scenario c: per Section 2.4.

Manufacturer claims of duration may be valid when new, but it is a well-known feature of rechargeable lithium batteries that a claimed lifespan of around 500 charge cycles depends on careful user maintenance of charge between 10 and 70% of the maximum. Thus the Oura® ring manufacturer claims a duration of 4-7d (Oura, n.d.), but users soon report much less. Such user experiences are common to all lithium battery-powered devices. This low endurance reduces lifespan of any device with a non-replaceable battery.

Power consumption is reduced for sensor devices without display that rely on the PAN connection, usually to smartphone, for display as well as operation.

It is then a matter for user choice whether to rely on the substantial power capacity of a current-model 6” smartphone, for which re-charge every other day is already acceptable to the user, or to increase sensor device power consumption, hence recharge frequency, by using its own display. Further consumer factors are ease and speed of recharge e.g. by wireless or thick cable, as thoroughly discussed on internet user fora.

Power management is a significant burden for a person who has not bought their own device for lifestyle purposes, but had it supplied for healthcare purposes when they may already be ill, as in scenarios d – f. It may require HCW attention, such as more frequent attendance, and so becomes a significant factor to evaluate for deployment.

3.4 The early diagnosis of COVID-19

3.4.1 Context

In COVID-19 the value of early detection is very high due to transmission for several days before symptoms. Johansson et al., (2021) have shown "the generation interval of SARS-CoV-2 was shorter than the serial interval … the epidemic was growing faster than would be expected if transmission were limited to the period of illness during which individuals were symptomatic. By the time a second generation of individuals was developing symptoms, a third generation was already being infected" The effect is that up to 59% of proven cases were infected by those who *were at that time asymptomatic.*

This unusual silent infectiousness before symptoms, that in "asymptomatic transmission" may neither show symptoms in the source nor cause them in the target, enables very rapid spread. Since summer 2021 this is much worse with the delta variant: 74% of all transmissions occur before symptom onset, and that pre-symptomatic window lasts for 4d, with peak infectiousness at 2d *before* symptoms, compared to the original virus strain's peak infectiousness only 0.2d before symptoms (Kang et al., 2021).

Mobile devices are well established at mass scale since the original Fitbit, and now >100m US citizens have such a band or smartwatch i.e. 1 in 5 US adults (McCarthy, 2019). As of 2019, more than 100 million global consumers owned Huami® wearable devices (Zhu et al., 2020), Huami® being only one of several major Chinese manufacturers.

These new parameters for early detection have been investigated:

3.4.2 Resting Heart Rate and other new cardiac signs

Radin et al., (2020) had already found that elevation of Resting Heart Rate (RHR) is a reliable initial sign of viral infection, as did Zhu et al., (2020) on 1.3m subjects in China and Europe.

Resting Heart Rate is a processed derivative of basic heart rate, which per Section 2.7.1 is a continuous variable of events/min, commonly counted by HCW for short periods. Machines now commonly monitor this continuously, and this new clinical sign depends on continuous monitoring of pulse with on-device co-processing with activity, which is only feasible with wearable devices. The personal stability of RHR had also eluded clinical discovery until the algorithmic processing of automated data from thousands of

mHealth devices. Further, the mean rise in RHR discussed below is only 7 bpm, so even with this new knowledge of what to look for, RHR is not detectable in healthcare practice: a mHealth device wearable for >24-hr. is essential.

Fortunately, the need for user-device pairing, as at Section 3.2.6, coincides with dramatic falls in price to consumer levels, so that most users can purchase their own device, or healthcare services can afford to supply each user. However, there remains a social gradient in device usage, quantified by Quer et al., (2021) to show lowest usage in those with lowest annual earnings, or educational attainment. An age gradient is found for those >50; however there is no racial gradient independent of those factors.

Unfortunately, RHR has several definitions, some in terms of a 24-hour data collection, so early RHR versions can only be determined once daily, which is unhelpful for clinical data.

Thus “Basal Heart Rate” is defined as the minimum median value of sinus-rhythm HR within 3-min windows during the day (Yuda et al., 2018). It is a little greater than “Minimum Heart Rate”, that is found around 04.00 with normal circadian lifestyle (Hayano et al., 2017); BHR is more robust than Min HR to variable circadian activities and sleep times.

Other definitions enable measures at any time, such as those by Logan et al., (2000) who compared five definitions of RHR “following different but commonly used protocols: four are derivatives (mean of lowest heart rate, plus all heart rates within three beats / means of lowest 5 / of lowest 10 / of lowest 50) and their own live measurement (for 5 min after a 10-min rest.) They found significant differences across population but none for individual rank orders. Thus “physical activity could be over or underestimated by about 30–35% simply as a result of the method of defining the RHR." Unfortunately, there was no statistical analysis of the correlations between the absolute values under the five RHR definitions.

The inter-subject error of RHR is also so high, ranging over 100% from 40 to 100, that it must be individually calibrated for that individual using that device. Both inter-subject and inter-device errors can be controlled for by custom calibration of each device if paired to each user, to eliminate those systematic errors and to maintain accuracy. The personal choice of device thus fixes the choice of RHR definition, which determines that device’s fixed algorithm to create RHR from HR readings.

RHR is now determined by processing of HR data in multiple periods of rest over 24 hrs. Rest is identified by co-processing inactivity data from an accelerometer, and with pattern recognition of stable low HR, and optional other factors such as mobility or sleep data. Proprietary algorithms are used, such as that by FitBit, selected by Radin et al., (2020) that is proven to match at *any* time an earlier finding that the most reliable 24 hour RHR is obtained when waking and still supine (Heneghan et al., 2019).
Further potential research questions are discussed at Section 5.2.1.1.
The algorithms used to determine this value are not in the public domain, so there is uncertainty if RHR is comparable between different devices; even those by the same manufacturer do not necessarily use the same sensor or algorithm.

3.4.2.1 DETECT: "Wearable sensor data and self-reported symptoms for COVID-19 detection"

This research program from California's Scripps Research Translational Institute was pre-published in outline at https://detectstudy.org/faqs.html and has now been reported formally by Quer et al., (2021).
This shows that a combination of step counts and sleep duration with rise in RHR is the most reliable sign to precede COVID-19 infection that presents with the 3 classic symptoms. However, it is not yet known if it is valid for infections with other symptoms, or for the asymptomatic, for whom it may detect infection of which they would never become aware.
Detection of signs for those without symptoms is not new to healthcare, being implicit in "case-finding" strategies, such as for blood pressure. Its extension to COVID-19 at the time of highest transmission would prompt anyone showing this new sign alone to intensify infection control, and/or to use new treatments.
Inter-user errors proved a significant issue when investigating the initial promise of RHR alone: only 30.3% of users with COVID-19 showed an RHR increase of over two SD, and RHR alone had insufficient sensitivity on AUC metrics as a sole risk predictor. It also confirmed the expectation that early RHR rise in COVID-19 would be consistent with the evidence from other viral illnesses i.e. that RHR rise precedes temperature rise.[32]

[32] See www.youtube.com/watch?v=Sc5ULh7jzpo&feature=youtu.be at 35mins. Dr Radin at 29m shows the normalisation of her own data to a RHR of 62, and during "Influenza-like Illness" it rose to 77.

The integration platform is either Google Fit or Apple Healthkit, so the DETECT program is device agnostic for the major brands, and some value-branded devices that also integrate with these two ecosystems. Thus inter-device error, such as use of different RHR definitions as well as the better-known differences in algorithms or sensors, has not yet been disclosed, not least because over 30,000 participants yielded only 54 cases to study. However, it appears that inter-device errors are sufficiently small that pooling their data is not invalidated; it remains to be reported if inter-device errors account for some of the low RHR increases found.

In a DETECT-based sub-study published in July 2021 (Radin et al., 2021) 37,000 participants had yielded 234 cases, and confirmed that in the pre-symptomatic (days -10 to 0 in diagram below) raised RHR alone is neither large nor early enough to be valid *as a sole risk predictor* in the whole population.

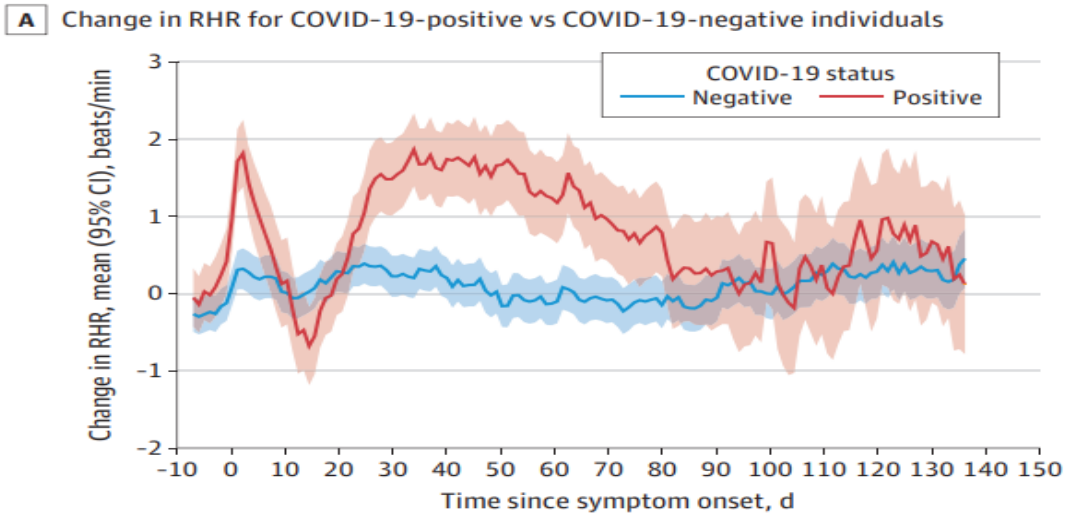


Figure 8: RHR early changes
Radin et al., (2021) p2 of full paper

Two new findings are

- initial transient *low* RHR around 9 - 15d. *after* symptom onset, as above, and
- prolonged relatively high RHR that did not return to baseline, on average, until 79d. after symptom onset. Step count and sleep quantity returned to baseline sooner than RHR at 32 and 24 days, respectively. For the worst-affected 14% of cases, this applied at a RHR level of 5 bpm over baseline for over 4 months.

This long-duration sign may be relevant to monitoring recovery (scenario e - per Section 2.4.) and to the diagnosis of Long Covid (scenario f.) It is said to be the first-ever study on long-duration wearable sensor data, for which consumer-oriented features of the wearable device are desirable – as evaluated here.

For this sub-study, inter-device error was minimised by use of Fitbit devices only to ensure consistency for the daily RHR calculation. Thus no further data are yet presented to clarify the issues above, though the authors predict that inter-user error may be explained better with larger sample sizes.

3.4.2.2 Stanford: "Pre-symptomatic detection of COVID-19 from smartwatch data"

"Longitudinal Baseline Profiling" had been studied at Stanford University School of Medicine for four years: "By profiling healthy people over time, we can identify early signatures of disease, keeping people healthy instead of treating them when they get sick." [33]

These potential new signs were all evaluated at night on FitBit® devices

- Respiratory Rate during deep sleep
- Heart Rate during non-REM sleep (another variant of RHR)
- HR Variability by RMSSD derivative (only available on premium devices)
- RRi entropy – a measure of its randomness, associated with autonomic health

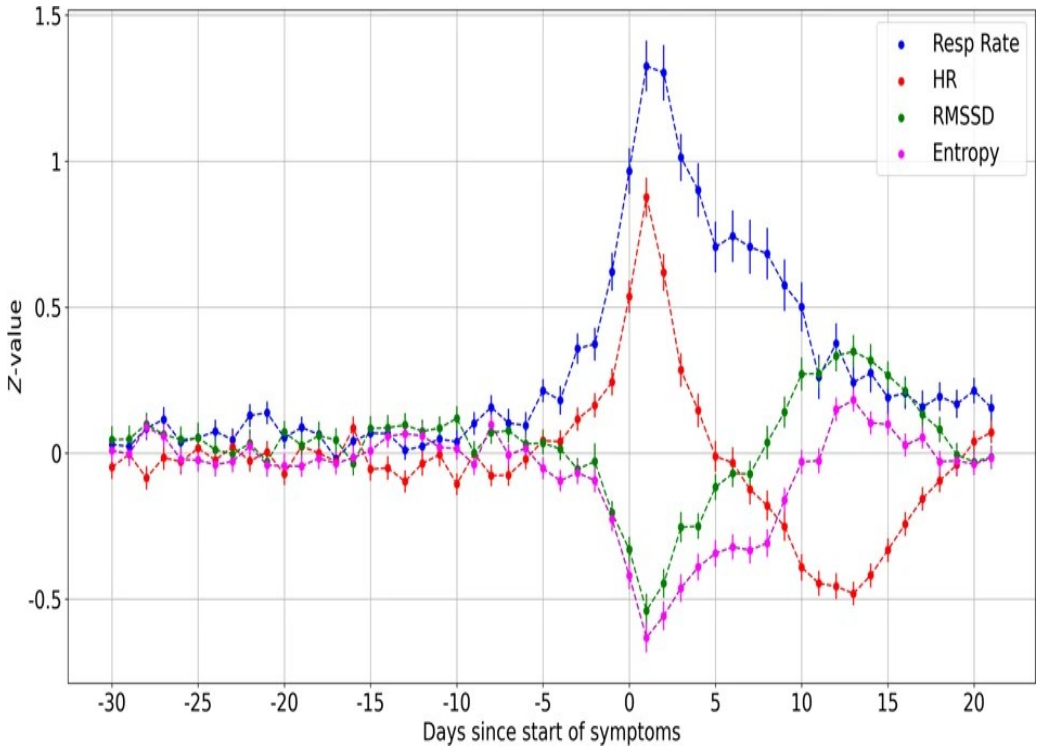


Figure 9: respiration, heart and HRV early signs
Natarajan et al., (2020 Fig2)

The first three were found to detect nearly 50% of COVID-19 cases, as above, but only one day before symptoms, and with a 70% specificity.

[33] See Wearables and Early Detection of COVID-19 Using a Smartwatch - DOM Grand Rounds - 30 Sept 2020 - YouTube start at 16min

In 2020 (Mishra et al., 2020) suggested that early warning of viral infection might be detected even 8 days in advance of COVID-19, as for this person

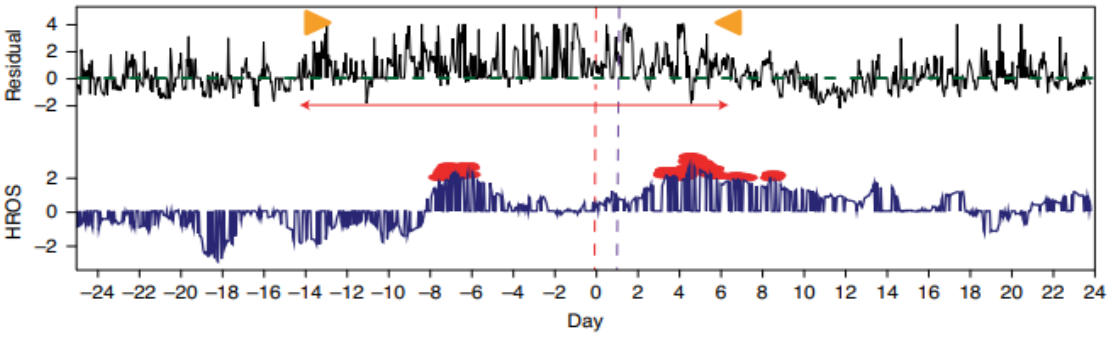


Figure 10: sign 8 days in advance
Mishra et al., (2020)

This was the first use of activity, as step count, to combine with Resting HR shown here as HROS = Heart Rate over Steps. Note that RHR here has yet another definition: the heart rate at a given time point where the step count for the previous 12 min was 0.

It is noted that COVID-19 shows an early rise in RHR, and this is earlier than for other viral illnesses. Mean elevation is only 7bpm, which is not reliably detectable clinically: a device is essential. Retrospective analysis gives 85% detection of COVID-19 using HR, steps and sleep data; and real-time detection currently gives 63% detection before or at symptom onset.

Data collection is now by MyPHD app, available for iOS and Android, and recruiting continuously, though currently for FitBit devices only. The platform is scalable to global scale and use of distributed computing across the world.

Elevated RHR data as native data would give frequent false positives, so it is processed into two phases of alert: yellow shows the first period of elevated RHR > 1hr, and blue for a severe rise in RHR. This is a visual display of a 3-way algorithmic classifier that was processed centrally, then returned to the user in the app. However, it only applied to under 2/3 of cases, and there is no discussion of integrating this onto the device.

Further, some “noise” factors can never be so processed e.g. the sensitivity is set to give a false positive rate in the healthy at 1 yellow alarm every 2mo., commonly associated with holidays as these are (in cardiac terms) stressful, such as raised HR due to long flights also causing reduced SpO_2.

The RHR data are normalised to each user’s sliding 28-day data, either offline (RHR-Diff method) for retrospective study, or online (CuSum method) to enable real-time alarms. These different methods complicate comparison of RHR alone between these landmark studies.

Of 5262 subjects, only 32 were cases, but 6 of them (18%) showed no changes in any of these 3 parameters – that is, there is a false negative rate of 18% due to inter-subject error that is unexplained. This is disappointing, and similar to that for RHR alone in the DETECT study.

Further, 17% of subjects have unstable baselines so were not suitable. It is not stated if further algorithmic analyses are underway to clarify this, such as using alternative definitions of RHR (of which several are discussed here) or if this is an inter-device error.

Some engineering data are reported by this Stanford research: it supports several models of FitBit device, and 8 other makers[34] of device; the integration platform is Apple Healthkit except for FitBit devices that used the FitBit API. Native heart rate, steps and sleep data were collected in JSON format: HR at 15s resolution, steps at 1 min resolution, and sleep data as 4 sleep stages (wake, light, deep and REM). Thus steps are already rendered as steps/min, which is standard; but the sleep data's classification into 4 standard phases is by algorithms that are not specified, though this is usual for consumer devices.

Code for the algorithms is available at GitHub – mwgrassgreen/WearableDetection (RHR-Diff and CuSum) and GitHub – gireeshkbogu/AnomalyDetect: Detects anomalous resting heart rate from smartwatch data. (HROS-AD), which is described as applicable to "any smartwatch data like FitBit, Apple, Garmin and Empatica."

Device resolution is discussed on the video: it is considered that FitBit devices have the highest resolution that is accessible, and that Apple technology has higher resolution, though accessible only via Apple software – that is, the resolution of timing data for others such as MyPHD using Apple Healthkit is limited [35].

Note that resolution limit here is not that inherent in the technology, but an intentional limit for sharing, as a business decision.

[34] Apple, Garmin, Oura Ring, Bio Strap, Masimo SpO_2, Empatica, Motiv Ring, and unbranded.

[35] See Wearables and Early Detection of COVID-19 Using a Smartwatch - DOM Grand Rounds - 30 Sept 2020 - YouTube, at 52 min.

3.4.2.3 Corona-Datenspende app for Robert Koch Institute (RKI)

Germany's major public health and research organisation had recruited, to Aug 2021, 537,000 users, known as "data donors" for this public platform.

As with the researchers above, a combination of RHR and activity has been found the best Risk Predictor, verified when aggregated on a population scale.

It also calculates each user's baseline cumulatively over their entire use.

Devices supported are FitBit, Garmin, Polar and Withings/Nokia, as well as all devices connected to GoogleFit or to Apple Health – so this should include many other consumer devices that can acquire HR and activity data.

RKI data donors use a pseudonym for full control of reuse of their data via the app, including full retrospective data deletion by the user. RKI does not return to the user their own risk prediction, for data protection reasons.

RKI published in March 2021 this table of which of the various devices' data could be processed by their webservices:

ROBERT KOCH INSTITUT

	Fitbit	Garmin	Polar	Withings	Google Fit	Apple Health	Samsung Health	Oura	Amazfit
Activity	✓	✓	✓	✓	✓			✓	✓
Activity detail (running, cycling, sports etc.)	✓	✓	✓	✓	✓				✓
Quiet	✓	✓	✓					✓	
Steps	✓	✓	✓	✓	✓	✓	✓	✓	✓
Calorie consumption	✓	✓	✓	✓				✓	✓
Covered track	✓	✓	✓	✓	✓			✓	✓
Climbed stairs	✓	✓		✓					
Sleep	✓	✓	✓	✓	✓	✓	✓	✓	✓
Vital signs									
Pulse	✓	✓	✓	✓	✓	✓	✓	✓	✓
Body temperature			✓	✓					

Figure 11: reuse numeric data from multiple devices per FAQ Corona Data Donation

However, according to user reviews of Corona-Datenspende at Google Play, connectivity is unreliable, for example excluding any Samsung devices.

Samsung and Apple devices must cancel data sharing with their own ecosystems to send it to RKI.

There is not yet published any data on inter-device variations.

Algorithm development is described, though not their definition of RHR.

National aggregate data are published on GitHub – corona-datenspende; the code is published under MIT licence for easy reuse. The cloud-based processing normalises for climate factors that affect activity, at small-district-scale.

3.4.2.4 Data interoperability

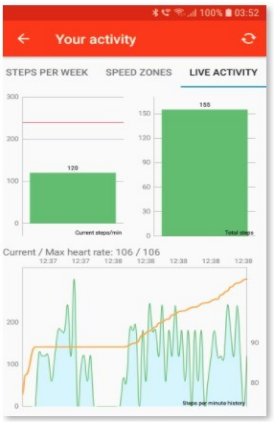
FitnessSyncer® is one of several internet-based services that accept data from over 50 sources, many more than do DETECT Stanford or RKI, and route it intelligently, to tasks such as alerts, further devices or services like emails, usually overnight in batches. Configuration by the user includes authentications and permissions according to source /target policies, so some will share data automatically. Where automation is not possible, it displays the native data, or converts to standard numeric ranges with UoM, so it is easy to co-process by the "human interoperability bridge" per Section 3.2.2. Graphical "dashboard" views are also rendered, with copious online advice.

A more extensive interoperability bridge between devices is open-source, and available at www.gadgetbridge.org , with code at codeberg.org. The title page states "Gadgetbridge is an Android (4.4+) application which will allow you to use your Pebble®, Mi Band®, Amazfit® Bip and Hplus device (and more) without the vendor's closed source application and without the need to create an account and transmit any of your data to the vendor's servers."

Interoperability of RHR data: both these platforms currently show only HR and activity data as step count. These are the basic data co-processed by both Stanford's and RKI's systems for early warning, so it is simple to derive Resting HR, with similar semantic validity. Addition of sleep duration may improve it, but this data is non-standard.

Gadgetbridge

A free and cloudless replacement for your gadget vendors' closed source Android applications.



Feature Matrix

	Pebble OG	Pebble Time/2	Mi Band	Mi Band 2	Mi Band 3	Mi Band 4/5	Amazfit Bip	Amazfit Cor
Calls Notification	YES	YES	YES	YES	YES	YES	YES	YES
Reject Calls	YES	YES	NO	NO	YES	YES	YES	YES
Accept Calls	NO(2)	NO(2)	NO	NO	NO	NO	NO	NO
Generic Notification	YES	YES	YES	YES	YES	YES	YES	YES
Dismiss Notifications on Phone	YES	YES	NO	NO	NO	NO	NO	NO
Predefined Replies	YES	YES	NO	NO	NO	NO	NO	NO
Voice Replies	N/A	NO(3)	N/A	N/A	N/A	N/A	N/A	N/A
Calendar Sync	YES	YES	NO	NO	NO	NO	NO(3)	NO
Configure alarms from Gadgetbridge	NO	NO	YES	YES	YES	YES(1)	YES	YES
Smart alarms	NO(1)	YES	YES	NO	NO	NO	NO	NO
Weather	NO(1)	YES	NO	NO	YES	YES	YES	YES
Activity Tracking	NO(1)	YES	YES	YES	YES	YES	YES	YES
GPS tracks import	NO	NO	NO	NO	NO	NO	YES	NO
Sleep Tracking	NO(1)	YES	YES	YES	YES	YES	YES	YES
HR Tracking	N/A	YES	YES	YES	YES	YES	YES	YES
Realtime Activity Tracking	NO	NO	YES	YES	YES	YES	YES	YES
Music Control	YES	YES	NO	NO	NO	YES	NO	YES
Watchapp/face Installation	YES	YES	NO	NO	NO	YES	YES	YES
Firmware Installation	YES	YES	YES	YES	YES	YES	YES	YES
Taking Screenshots	YES	YES	NO	NO	NO	NO	NO	NO
Support Android Companion Apps	YES	YES	NO	NO	NO	NO	NO	NO

Figure 12: vendor-neutral reuse of data per Gadgetbridge/FEATURES.md

3.4.2.5 *Veracity of RHR*

MHealth data such as RHR shows low error due to the high resolution of the devices. However this new clinical sign is insufficiently sensitive alone, and the Stanford authors do not discuss formally its potential place as one of several factors in combination as a RP tool, which we have established at Section 3.2 as the commonplace healthcare practice of diagnosis using many datapoints as in a RP tool.

For example, COVID-19 diagnosis does not rely on any single symptom alone, using three of them only in combination as a pre-condition to get a PCR test, or as one of several data to be co-processed as input. Could RHR be included?

Dr Snyder of Stanford remarks in discussion[36] that "RHR is certainly superior to temperature" as a sign of COVID-19. Temperature is generally accepted as a valuable sign – but not in isolation.

3.4.2.6 *Heart Rate Variability*

This is the variation in pulse rate, often known as R-R interval (RRi) in reference to the start of each heartbeat that is known as an R-wave on an ECG. It has long been used in cardiology research e.g. (Guidelines et al., 1996) and known as an early indicator of infection, but for COVID-19 this has only recently been confirmed to also apply, by bulk processing of all FitBit data from 1180 people with COVID-19 (Natarajan et al., 2020) However, in this paper there is no discussion of the relative merits of HRV compared to RHR: this is still a live area of research, as in the Stanford research at Section 3.4.2.2

3.4.3 Oxygen levels (SpO_2)

These are of high value in early detection of progression of COVID-19 infection to pneumonia, of which many elderly report so few of the usual symptoms, including lack of raised respiratory rate, that they elude diagnosis (Jouffroy et al., 2020) but rapidly succumb at home.

For elderly people, the common presenting symptoms in adults – cough and fever – are seldom present, but 19 other symptoms are listed (*Atypical Covid-19 Presentations in Older People – the Need for Continued Vigilance | British Geriatrics Society*, n.d.). However none of these are specific to COVID-19. Of these only two – falls and weakness – are recognised as components of frailty, and so open to early detection by frailty monitoring systems such as Armed™.

---

[36] Wearables and Early Detection of COVID-19 Using a Smartwatch - DOM Grand Rounds - 30 Sept 2020 - YouTube at 58:45

Further, unexpectedly, the Rockwood Clinical Frailty Scale has recently been found to be *not* a reliable predictor of COVID-19 mortality (Miles et al., 2020)
The accuracy in prediction of COVID-19 of established Risk Predictor tools is subject to continuing research by others and out of scope here.

Pulse oximeters now have mass availability at consumer prices, but must be worn on a finger, or ear, and operated to a schedule, so are intrusive to the user.
Semi-automatic operation is now more feasible with oximetry by PPG built into smartwatches from major brands and some other less known suppliers who promote them as just another lifestyle device, and also in formats for finger (Oura® ring) and legs, that are acceptable for continuous wear, requiring no user action, and supply continuous data during everyday life – that becomes "Everyday Data."
It would thus seem a clear candidate for NHS supply to at-risk COVID-19 cases in scenario c per Section 2.4. Finger oximeters are already supplied by NHS, so this fast-moving field is out-of-scope here.

3.4.4 Continuous glucose monitor (CGM)

This new technology has wide uptake in the everyday lives of people with insulin-dependent diabetes, confirming its numeric precision with healthcare veracity, continuous data streaming, and wearability.
(Wang et al., 2021) have now shown that a rise in glucose is a powerful early sign of rapid progression of COVID-19. It is also now clear that COVID-19 specifically attacks the islet cells of the pancreas, aggravating type 1 diabetes (Wu et al., 2021).
Rising glucose is well known as a concurrent sign of infection, but it is not yet clear if it is also an early sign, which would be easily detectable for those already using CGM.
This now-simple and widespread Point-of-Care type of mHealth technology is easily adaptable and would thus seem a clear candidate for NHS supply for home monitoring in scenarios d - f per Section 2.4.

3.4.5 CRP monitor

CRP is a biomarker of inflammation that is a strong indicator of deterioration, often due to sepsis, and now in COVID-19 seen as a proxy for IL6, the major marker of inflammation in the 2nd phase. It is often used on its own to prompt hospital admission. This test is recently developed as mobile Point-of-Care technology.

It uses a drop of blood by finger-prick. This was the same technology as earlier blood glucose monitoring, so sampling is widely known to be feasible by the subject alone, with usability assisted by semi-automated lancet devices.

The sensor is in a portable opto-electronic device, but is not yet designed for patient use, so is currently operated by HCW, for example in a Hospital-at-Home team (Healthcare Improvement Scotland, 2020), or as currently implemented in Medway (personal comm)

It would thus also seem a clear candidate for development to solo use and NHS supply to COVID-19 cases in scenarios d – f.

3.4.6 Inactivity

This is a new sign derived from continuous activity monitoring by wearable devices. The initial framework was set by Polar® devices, for which calibration data are published (Polar Inc., 2021). The overall and peak activity levels are of interest to athletes, but are not useful for falls, frailty or illness: for this, a standard *inactivity* period of 1hr has been found a valid sign (ARMED, n.d.) Accelerometer data are processed on-device to alert for inactivity at 55mins, alerts the user, and if it continues a further 5 mins, an "Inactivity" counter is incremented. These parameters are currently fixed in firmware, as is the sensitivity of the wrist-worn accelerometer sensor. (Smales, A, personal comm)

The device and its algorithms are proprietary, but other manufacturers have standardised on the 1hr inactivity metric, which has face validity to HCW and users.

3.4.6.1 Precision

Wrist-worn devices are very liable to motion artefact – see Section 3.3.2 - but this should be minimised when the data of interest is inactivity itself.

3.4.6.2 Accuracy / veracity

What is the relation between shorter periods of inactivity and a longer timespan, or patterns within the inactivity e.g. circadian, or association with mealtimes?

Is the relation between inactivity and harm linear? Should simple inactivity be distinguished from a nap, now known to have positive health restorative effect (Naska et al., 2007) as well as cognitive benefits? How should inactivity be distinguished from being sedentary[37] (Tremblay, 2012). These semantic issues are critical to the

[37] Cycling is classified as a sedentary activity, due to seated posture. Bradley Wiggins' view is unknown!

healthcare veracity of the various measures that can be derived by ML from activity monitoring, which creates derivatives that are too complex for humans to derive, but may lack meaning.

In terms of traceability, it is not biologically plausible that the optimum correlate with harm is a quantity from a continuous measure sampled between two time-point markers set by humans[38] at 1/24 of a day. It lacks "bio-logic", nor any external reference for accuracy, whereas a 24-hour period has biological plausibility in the omnipresence of circadian rhythms. The use of hour-long periods for observations derives from HCW management, or simple time-stamping with clockwork clocks. Yet 24hr or longer durations are simple for machines, that can also process the vast quantities of data so produced into human-readable derivatives at *any* interval.

3.4.6.3 Precision

Is the device resolution consistent e.g. in detecting a threshold of inactivity?

Is this adjustable?

Other parameters of inactivity are used by many other devices as part of their algorithms to detect sleep, and now to detect fatigue due to early viral illness.

These may be the first of many new signs that have eluded clinical detection because that is impossible, and depend wholly on the continuous and "everyday" automatic measurement with asynchronous machine processing of mHealth devices.

The role of accelerometer devices in early detection of COVID-19 is still being established in the work by Radin et al. but there does appear to be a role for on-device derivatives of activity data as part of a RP tool for scenarios a – c -see 5.2.1.1.

3.4.7 Sleep

Sleep detection technology in mHealth is fast-moving with two major use classes here

- direct analysis for signs of viral illness
- indirect use via algorithms for that individual and device, to normalise RHR and HRV, and manage inactivity, to co-process for the above

Sleep is detected by algorithmic processing of several data streams, such as from sensors for movement (accelerometers), position (gyroscopes), heart rate and

[38] The Sumerian allocation of 24 hours per day appears to be for mathematical not biological reasons

variability (PPG), BP, or SpO_2 (PPG oximetry), and/or dedicated devices such as at Section 3.3.1

It can also identify and quantify the time spent in deep or REM sleep phases.

These algorithms are proprietary, and rapidly developing, but remain far from standardised, so calibration of each device to its user is essential.

3.4.8 New technologies

Three new technologies still under development are candidates for evaluation at 6.6.3

- Ulisses project (CO_2 Sensors in Smart Phones: A ULISSES Webinar, n.d.) will add CO_2 sensors to smartphones, giving each user a live indicator of the risk of shared air that is a major risk for infection.
- Cough phonics (Laguarta et al., 2020) are analysable on-phone to show COVID-19 as an early sign.
- Paper dot smell tests for C19 (Basu et al., 2020) are inexpensive detectors for loss of smell, with evidence of being an early sign – but need manual operation and data entry to smartphone app, so not at all automated.

All “devices” that rely on the subject’s manual data entry are far from automatic, so score low on Deployment. They may benefit from

3.4.9 Voice agent input

such as by Alexa or Google, or local-only custom software, to add consumer-level useability. See (*Create a Question and Answer Bot with Amazon Lex and Amazon Alexa | AWS Machine Learning Blog*, n.d.) (*Alexa Can Help Make Strides in Telecare | HeraldScotland*, n.d.)

3.4.10 Other parameters

3.4.10.1 Location tracking using GPS

This is already deployed on a global scale, and is built into all recent smartphones, requiring no further app to be installed, so collating social-scale mobility data. Bluetooth-only systems detect mutual proximity alone, so rely on users installing an app; however the Personal Area Network range coincides with the airborne infectiousness range of Sars-CoV-2 virus. It was first used in 2011 by *FluPhone Project: Computer Laboratory*, (n.d.) for fully voluntary research. It is now used in China, Taiwan, S Korea and some states of Australia with varying levels of data reuse by official contact tracing systems. (*Corona Apps: This Is How Countries Are Dealing with Privacy (in Times of Covid-19) | VPNoverview.Com*, n.d.) and in voluntary systems[39]. Its positional accuracy is less than for systems using Bluetooth, and its more comprehensive usage should better detect potential super-spreading social events. But it has been excluded from the Exposure Notification Systems API of Apple and Google (*Exposure Notifications Frequently Asked Questions Preliminary-Subject to Modification and Extension*, 2020) due to major privacy concerns. However, it remains a technically feasible option to improve the scale of social mobility tracking, if privacy concerns were addressed. This could be by explicit consent, or full user control of data, as demonstrated by RKI, and appears to be as implemented in Iceland, where the app stores data only locally, and is wholly user-operated.

The risk appetite of users varies: those in the Clinically Extremely Vulnerable (CEV) category after 18 months of confinement to their home may tolerate this means to enable safer exposure to society despite their other risk factors.

3.4.10.2 COVID-19 Point-of-Care tests

These are now widely available at consumer prices, or inexpensive enough to be distributed free-of-charge by health systems. However, they have false positive and negative rates like all tests, which has led to much public education on the accuracy of tests, and their validity in different contexts. For example, the presence of one of three other classic features of COVID-19 was required to access PCR tests, showing that even with their high precision and accuracy, they were not thought valid enough *in isolation*.

[39] in Bahrain, Brazil, Bulgaria, Chile, Colombia, Czechia, Spain, Iceland, India, Indonesia, Israel, Kuwait, Macedonia, Mexico, Norway, Poland, Qatar, Russia, Singapore and Thailand.

3.5 Framework development

The framework required here was developed by following the "hermeneutic analysis" by Greenhalgh, A'Court, et al., (2017) of several different frameworks to evaluate "Telehealth" systems for chronic heart failure, which had variable success.

That paper addresses several tensions between the direct care needs of people when ill, and the business processes around supply, configuration and maintenance of technical remote monitoring systems, usually by teams of strangers managed by remote organisations and "based on a modernist vision of efficient, rational, technology-mediated and guideline-driven ("cold") care." Other tensions are between textbook, single-diagnosis guidelines and multi-morbidity, relevant to interpreting data; and between levels of support in domestic settings, that in reality can be obstructive or counter-productive, especially for contact and airborne infection control of COVID-19 that requires isolation to standards that cannot be achieved in many domestic settings.

This led initially to a simple list-of-criteria type of framework to re-evaluate telehealth as the ARCHIE framework (Greenhalgh et al., 2015) which describes standards for services to serve as a counterpoint to technocratic evaluations that rely on data: they should be

Anchored in what matters to patients

Realistic about the natural history of illness

Co-creative – evolving and adapting solutions with users

Human – supported through interpersonal relationships and social networks

Integrated, through attention to mutual awareness and knowledge sharing; and

Evaluated, to drive system learning.

This hermeneutic approach would clarify, create insight and so add meaning to this socio-technical domain, as per Section 2.5. The telehealth paper also used a simple, if lengthy, list of 37 factors "compiled from eleven various sources." While allocated into six classes – patient, staff, technical, team/service, governance and financial/business, this may be too many, as several of these 37 factors overlap, and few projects will need all these factors.

These researchers (Greenhalgh, Wherton, et al., 2017) then reviewed 28 frameworks and suggested they fall into six classes: theory-informed reviews, logical models, lists-of-criteria, static models, dynamic models, and one "individual adoption" framework. They then developed their NASSS framework as a dynamic model for how (Non)adoption/Abandonment, Spread/Scale-up, and Sustainability factors apply.

Here, the concept "Framework" implies some causation between its listed elements that may be logical, or physical, as a mechanical framework is coupled by gravity. So of the six classes identified here, while a one-dimensional list-of-criteria might show new connections, a theory-informed review, or logical or dynamic models should create a more logical framework.

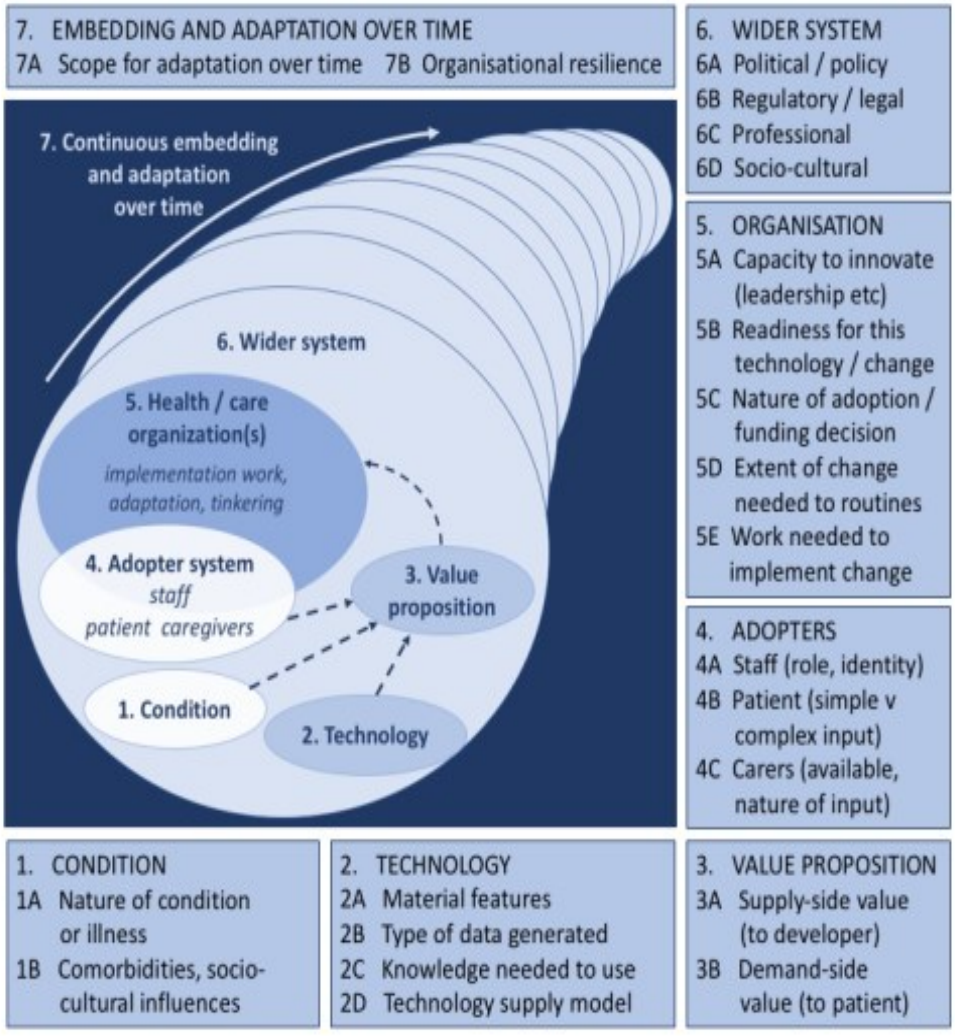


Figure 13: NASSS framework
Greenhalgh, T., et al (2017). Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *Journal of Medical Internet Research*, *19*(11), e367. https://doi.org/10.2196/jmir.8775

The logic of PAIDUR is simple: the flow of data from devices to population.
Further, a framework where too many elements cannot be supplied with content from the topic is excessive for its purpose.

The ambitious scope of NASSS appeared to suit such a complex socio-technical system of large scope as mHealth. Since interventions in these systems often fail to deliver the real-world benefits expected by their technical designers, they are also examples of another analogy discussed at Section 2.5: the Valley of Death between Bench and Bedside described by Seyhan, (2019) when discussing Translational Research. Both paradigms signpost analogous dangers for technical mHealth systems in the social world.

This work takes a heuristic approach, so it is

- secondary research in hermeneutic review style
- adds accounts of mHealth technologies appropriate to healthcare scenarios
- applies metrological engineering principles to their data quality
- analyses real-world healthcare usage for further data quality procedures
- infers components for the framework from these, and
- describes the interactions of the components
- application to COVID-19 is embedded in the framework development.

It is not

- a literature-led top-down synthesis, such as TEHAI
- systematic, but selective for a solo author
- a catalogue of mHealth devices or how they all work

4 METHODS

4.1 Overall approach

The work has been shaped by COVID-19, which continues to produce new fast-tracked research[40]. It explores and constructs links between what is found in the mHealth domain - an interpretivist approach, rather than a positivist "mining" of the domain to discover its hidden objective truths[41]. This approach applies also to literature review. These issues were discussed with the supervision team and the sponsors.

The scope of the project is shown in the axiom of translational research "from bench to bedside", which frames it as the journey taken by a new scientific idea (Cohrs et al., 2015); the idea that is applied here is the journey of mHealth data. Here, "bench" is considered as the start of the data journey from a transducer, and "bedside" is considered as deployment beyond the individual to a population.
Because the whole data journey can now be integrated into one device, and for a population into one ecosystem, a whole-system approach is required: while each data journey is one-way and so might be divided into sequential sections, these sections are not separately accessible, as discussed at Section 2.10.2.

At intermediate stage, the plan was to

- scope data content for eight COVID-19, frailty and related RP tools for standard clinical, wearables and mHealth, and Social Care contexts
- select those data items of most value in addressing the COVID-19 emergency
- develop 4 research questions on
    - computability of data in Risk Predictors
    - how they can support COVID-19 management
    - snapshot the recent State of the Art in mHealth technologies
    - research the informatics of interoperability and reuse
- develop a heuristic framework to source and manage new data from wearable and mHealth devices for reuse with other data in RP tools.
- evaluate online HCW confidence in data, and trust in current devices

[40] such as a new RP tool developed this year for COVID-19 (Fiorentino et al., 2021).
[41] per Kvale's classic epistemology of research (*Issues of Validity in Qualitative Research. - PsycNET*, n.d.)

4.2 Content analysis: the meaning of words, phrases and sentences

Key terms and language were developed by building a custom glossary on-the-fly with contextual descriptions of terms, to ensure consistency- see Glossary v0.8 - Google Docs for live version, and some in-situ glossaries. 5 further examples are

1. Scales, scores and indices are ***measuring instruments***, and so might be analysed as physical measuring tools: this was discovered quite late in the project. Clarification of earlier draft texts standardised on "Risk Predictor tools" as a term that references their origin as tools, so based in metrology.
2. ***User-device pairing*** was also recognised late on as a common method of increasing certainty. This is commonplace in devices retailed for individual use, but rare in those supplied for professional use, as discussed at Section 2.4 contrasting domestic and professional scenarios. Pairing is an intrinsic part of each user initialising each new device for personal use, but has not been evaluated as part of metrological accuracy, the engineering basis of certainty.
3. ***"Computability"*** is widely used in informatics, but required definition here, which led to Research Question 1. However co-processing different types of data was not easy to define. This shows in the ugly neologism "mung" with its vague and ambiguous definitions related to "data wrangling" – see (Mung (Computer Term) - Wikipedia, n.d.) This difficulty in defining "co-process", the term preferred here, was discussed at Section 3.1.12.
4. ***Types of numbers*** per mathematical set theory was explored as a basic cause of low resolution in RP tools that ignore correct logic in co-processing. Section 3.2 evaluates how quantitative and qualitative data are combined in RP tools with critical analysis based on these mathematical and metrological concepts.
5. The term "Share" implies one act of copying permitted by a data controller with immediate effect. However, "Re-use" better describes the indefinite further data copying now permitted, often automatically by remote machines configured long ago, to new business associates unknown to the original data controller when permitting a "share" - so is preferred here. This also affects the framework's acronym development described at Section 5.5.1.
6. ***Users*** are described by this taxonomy: "users" are both individuals located worldwide, and collectively global, and are "consumers" of devices for lifestyle. "Users" includes when they engage with healthcare when well, and are also known as "patients" when ill, or "subjects" in scientific discourse.

4.3 Thematic analysis: four research questions to drive research

4.3.1 RQ1 What are the types of data currently used in Risk Predictor tools?

Data used in health and care are first reviewed for data quality, for both quantitative and qualitative data, and the metrological, engineering or qualitative uncertainty arising from conventional and new data sources. To be meaningful to the mind, whether of consumer or HCW, native transducer data need processing by machine, and the concept of "computability" determines the processing methods applied to conventional and to new mHealth data to create meaningful classes. This is answered in two sections:

4.3.1.1 What are the logical types of data used in Health and Care? - see Section 3.1

A common "Operational Paradigm" is derived from three information models that lead to Actions, including sharing to HCW for their reuse. The data may be either direct (native) or derivative (processed). Analysis of the computability of data, its logical structure supporting algorithmic processing, leads to analysis of types of number and of text, and heuristic consideration of their use in healthcare, in terms of data quality incl. semantics, and co-processing with other healthcare data.

4.3.1.2 How is healthcare data used in Risk Predictor tools? - see Section 3.2

Healthcare data used in RP tools, traditionally scales scores and indices, are critically reviewed in metrological terms for validity.

Traditional devices have metrological flaws, yet can be used safely and effectively enough for healthcare, by quality assurance procedures.

The output "action" is the major purpose of current RP tools, by deriving classes, usually binary such as present/absent that are easy for the mind to process.

4.3.2 RQ2 Which healthcare data can be extended using wearables and mHealth, for early diagnosis of COVID-19? – see Section 3.4.

The healthcare data processed for frailty, and other conditions, are analysed also as content for use for COVID-19. Early diagnosis is critically reviewed to select data for further analysis: resting heart rate, activity and sleep. Extending these leads to RQ3.

4.3.3 RQ3 How can Extended Data be automated as Everyday data? - see Section 3.3

Automation is a key feature: "Everyday" automation is achieved if frequent and automatic enough to become a daily background service as if infrastructure. This is a generic function that is shared by many others, a concept promoted for mHealth by Seshadri et al., (2020). The "Everyday" concept for frequency + usability identified Asynchronicity, that decouples data acquisition, on-device processing and presentation of data to user or HCW for action.[42] Can HCW time be liberated by machine-powered data processing being decoupling from HCW work?

A more detailed analysis of mHealth devices' operational issues includes wearability and power management.

4.3.3.1 Device features to acquire data for early diagnosis of COVID-19

The development from traditional manual to semi-automated Point of Care and automated mHealth devices shares new key technologies such as photoplethysmography (PPG) and motion transducers – see Section 3.3.

The automation of these data sources is essentially driven by smartphone development, connecting the data from local device to internet.

4.3.3.2 Heuristic and Technical issues of automation and data acquisition

Automation depends on intrinsic transducer performance, and extrinsic factors such as usability: while everyday operation may be automatic, each user must still acquire and initialise the device by pairing it with themselves, one-off user actions that cannot be automated. This is no problem in the world of lifestyle healthcare with fit-enough users choosing their own device, but not so easy for the same people when they are ill.

[42] This is part of the business process revolution in use of time for data management, driven by COVID-19, now known as Digital Asynchronous Consulting, for which HealthIT Systems are being re-designed.

4.3.3.3 Information Governance

Validity is considered as part of Data Quality, of which timeliness is the key new feature supported by automation of timing metadata to manage asynchronicity.
A citizen's own data should be no less shareable than is NHS data about that citizen.
This is detailed further in the Co-PHR Guidance (*Apperta Foundation*, n.d.)
Data sharing by the user to an internet platform is usually required as part of initialisation, but the further re-use of that data by unknown international business associates is intentionally obscured to most users.
User control of re-use is not feasible within current business models, but can be crudely managed as shown by RKI at Section 3.4.2.3.

4.3.3.4 Selection framework for Devices to supply *Everyday Data*

Review of mfr. literature incl. White Papers, user reviews incl. personal use.
This led to the metrological approach (***P****recision*, ***A****ccuracy* and resolution) and analysis of co-existence of quantitative and qualitative data.

- ***I****nteroperability*: can the data retain semantic validity in new contexts?
- ***D****eploy*, initialise (connect, automate, schedule)
- ***U****sability,* now driven for mHealth by the consumer market
- ***R****euse:* automated so also decoupled from either user or HCW action

Thus the bolded initials of italicised items create the heuristic framework "PAIDUR."
Asynchronous data acquisition was also identified as a new feature for everyday use.

4.3.4 RQ4 How to render Everyday Data for generic use in any system

4.3.4.1 Data integration and ReUse - Data QA, transforms

A brief overview of Interoperability - how systems should talk to each other - describes how openEHR archetypes are generic clinical models that incorporate disparate data types into a common model, so it is generically available for reuse in other systems.
PPSv2 archetype development was undertaken to explore some of the issues e.g. that co-processing logic cannot be incorporated in the archetype, because the archetype is only a structure for data, without code. To process the data requires code, which needs software outwith the archetype – so out-of-scope here.

4.4 Discourse analysis: communication and meaning in social context

Due to COVID-19 effect on university capacity and functioning, opportunities for dialogue on this work were sparse. However, from those that resonated in retrospect, these comments by Subject Matter Experts were immediate and brief, suggesting System One or intuitive thinking (Kahneman, 2015) that is here valuable exactly because of their expertise. These System One responses are evaluated at Section 6.1.

4.5 Literature Review method

This was initially scoping, then narrative-led. Due to its multi-disciplinary and fast-moving nature, led by international business not by academia, systematic review was not considered to be timely enough due to the time-lag in academic publication. Some key papers served as sources for reverse snowballing (Wohlin, 2014). Due to the rapid changes in technology, "grey" literature sources such as manufacturers' White Papers and promotional material were used. These may not persist as original publications, but may be captured on web archives such as Wayback Machine, and their relevant content is captured here.

This may be considered as a hermeneutic method, described by (Boell & Cecez-Kecmanovic, 2014) as "continuous engagement with and gradual development of a body of literature during which increased understanding and insights are developed":

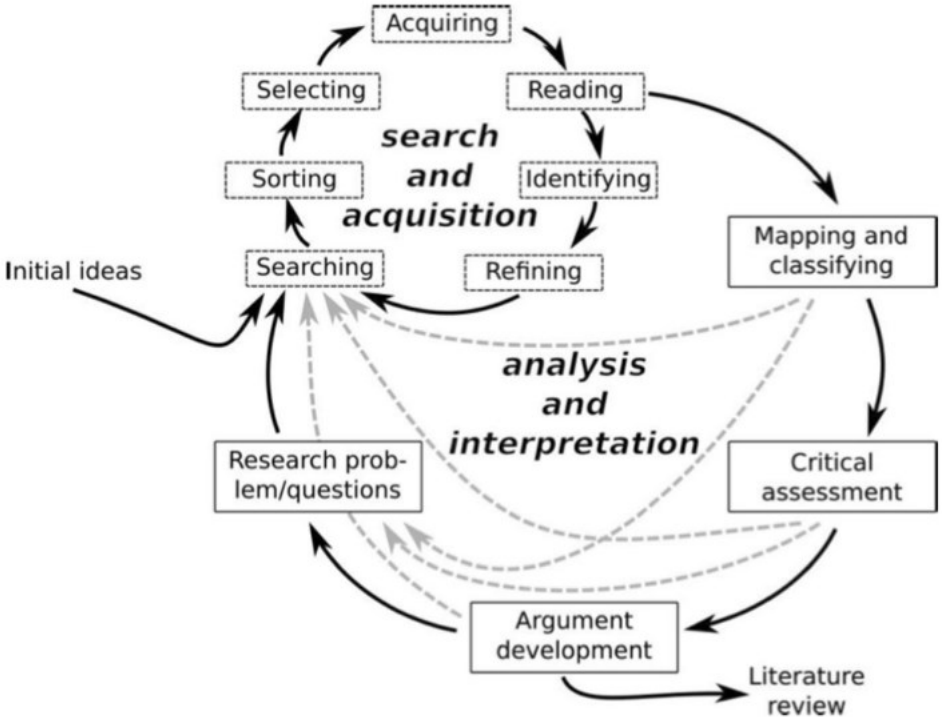


Figure 14: the hermeneutic method
Boell & Cecez-Kecmanovic, 2014

Such a hermeneutic analysis was cited by Greenhalgh, A'Court, et al., (2017) for their evaluation of the different frameworks for "Telehealth" systems deployed for chronic heart failure, which are complex socio-technical systems, and discussed at Section 3.5. Reverse snow-balling of this citation led to Boell et al's theoretical framework, which further cites Kuhn and Wittgenstein[43].

This group's NASSS study (v.i.) used secondary research (hermeneutic systematic review) to identify key domains, then empirical case studies of six technologies to investigate these domains and their connections.

The literature reviewed is a snapshot at end-2020 for project management reasons, one key paper was updated: the DETECT study update on 7.7.21 (Radin et al., 2021).

4.5.1 Key papers

1 IVM Metrology introduction:

BIPM Joint Committee for Guides in Metrology. (2008). JCGM 200 : 2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM). http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf

2 DETECT study:

Quer, G., Radin, J. M., et al (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. Nature Medicine, 27(1), 73–77. https://doi.org/10.1038/s41591-020-1123-x

3 NASSS Framework:

Greenhalgh, T., Wherton, J et al (2017). Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *Journal of Medical Internet Research*, *19*(11), e367. https://doi.org/10.2196/jmir.8775

Stylistic guidance was also obtained from the Academic Phrasebank by Morley John, (n.d.) and from "Fowler's Dict. Mod. English Usage," (2015) Fowler's Dictionary of Modern English Usage - Oxford Reference

---

[43] which will not be further cited here, as this academic provenance should suffice.

This "traceability" of the academic provenance is analogous to that here described for Accuracy in metrology: there is heuristic judgment to be made on how far to pursue this, just as physical data in healthcare are implicitly grounded in the SI metre in Paris, and the value of g – but never explicitly.

5 DISCUSSION AND SYNTHESIS

5.1 RQ1 What are the types of data currently used in Risk Predictor tools?

This summarises the uncertainty, utility and meaning of types of data in healthcare records in metrological terms, how they are managed heuristically in healthcare, and identifies how logical datatypes are “computable” when co-processed.

5.1.1 Logical: algorithmic

Data are structured, including metadata: quantitative data about errors, such as for physical measurements, provenance or timing, or derivatives from native data, such as classes, or risk predictor scores. Algorithmic processing depends on data structure.

5.1.2 Heuristic: scalable

The scale of data is within the processing power of systems which range from the human mind, which is quantitatively feeble but qualitatively powerful, to massive machine- or cloud-computing power, which is the converse. In healthcare it ranges from “mental arithmetic” for simple data, as in frailty indices, to machine-powered processing on mHealth devices, for which the power of “cloud” processing is available via smartphones and the internet, as in DETECT, Stanford and RKI programs.

For healthcare, the data content depends on data quality, including

5.1.3 Uncertainty of data

This for quantitative data is commonly quantifiable, but for qualitative data less so, though ordered text statements can be time-serialised.

Uncertainty can itself be processed as numeric metadata, as both precision, where error is due to random factors, and accuracy, where due to systematic factors

For qualitative data, the term for accuracy introduced here is “veracity.”

Both are limited by the tool’s resolution, the smallest change the tool can detect.

For whole numbers, errors should be rare, such as mis-counting due to transducer error. However, processing errors due to poor design have also been discovered here:

5.1.3.1 Ordinal processing

Statistical processing applies to ordinals even when they confusingly use whole numbers as symbols, not simple arithmetic. In summary *these proposals* are made:

Ordinal scales should not display rank by numbers, but instead use other known serial symbols, such as Roman numerals or an alphabet.

5.1.3.2 Co-processing:

Different data types can be processed together. RP tools that aggregate different number types should *process each according to their specific mathematical logic.*

5.1.3.3 Propagation of uncertainty

Co-processing data aggregates the uncertainty of each component, which propagate with each reuse of the data – see Section 3.1.2. As probabilities are expressed as %, they should be co-processed with their measurement uncertainty also expressed as %. *Uncertainty metadata should be available via APIs to be persisted when reused.*

5.1.3.4 Category errors

Probability can apply to all errors, so can also apply to other datatypes than measurement data, such as group data. In the cases of missing data, misidentification and ecological fallacy, the data may be wholly wrong for an individual i.e. 100% inaccuracy[44]: these are major or catastrophic category errors.

This may partly explain the poor performance of RP tools at individual level.

All healthcare data are also liable to misidentification error.

How systems should manage these major category errors is an information governance issue and out-of-scope here.

5.1.3.5 Machine processing

for RP tools can support a massive scale of data at speed. It can process continuous values using ADC, so does not require data to be input as binary, nor pre-classified.

It can support data granularity at high resolution that may seem to exceed direct care needs, but is valuable because

- uncertainty aggregates with co-processing, and
- complex adaptive systems can rapidly destabilise on small perturbations, that DSP controlled by algorithms can detect

5.1.3.6 Error, probability and clinical judgment

Where quantitative errors can be rendered as probabilities, processing by machine is required, especially for aggregate derivatives such as RP tools .

[44] as defined here as "veracity,"

However, when data are qualitative, this co-processing of different conditions with the supporting data, all with uncertainty, is in the mind, and difficult for even the most experienced clinician. This expertise here is known as “clinical judgment.”

5.1.3.7 Uncertainty management in healthcare

Direct data that can be rendered as numeric, with standard units of measure, can be easily used by consumer or HCW, who currently use data of lower-resolution and less certainty. For those devices, healthcare practice has developed and trained in standard procedures to minimise device error.

The major procedures in direct care are

- repetition for precision
- triangulation with other data for accuracy
- serialisation for accuracy

and for planned use

- scheduled recalibration for accuracy
- device standardisation by external agencies for both, and for trust

All of these mitigations may also apply to new devices.

However, only mHealth devices[45] can add

- user-device pairing for accuracy and veracity
- on-device processing creates meaning, such as by
  - intelligent displays, such as graphs to time-serialise dynamic data
  - classes, for pre-validated derivatives such as actions
  - RP tools, to combine multiple data to derive actions

Such classification adds further meaning for both consumer and HCW user and so improves decision-making for actions.

[45] “mHealth devices” are the combination of wearable sensor and app on smartphone or others per glossary at section 2.10.1

5.1.4 Classification of data

The semantic purpose of classification with a text label is the same as that for specifying its units: to reference context, from which the mind infers veracity. Multiple classes can be applied simultaneously, increasing meaning.
The logical purpose of classification is to downscale processing to the power of the mind, so ML processing uses two major methods

- decision-tree processing for known parameters
- rule induction to create new classes, that may be new clinical signs.

Numeric data may be of high quality, using high resolution for precision and accuracy. But once transformed by classification into text terms, this data quality is in jeopardy due to the classifier, and the semantic validity of the text terms.
New data such as from mHealth needs to establish healthcare veracity that is similar – but not better.

Since classification always loses resolution, so
for re-use, native numeric data should be preferred to derived data such as classes

5.1.4.1 Reuse of data

MHealth devices can process on-device their native continuous data into classes for consumer usability on the device. However they may not expose this in an API for others to reuse, which is a major obstacle to simple reuse of native data.

5.1.5 The operational paradigm

The operational requirement for a tool may not be for resolution per se, but that its serialisation is sufficiently valid to create action(s), the hallmark of the operational paradigm. A trend of only two datapoints may be valid when there is also sufficient confidence in the data, or trust in the device, to use intuitive or fast thinking to derive a trend. This mental function as a binary classifier can infer action.

5.1.5.1 Confidence and trust

User confidence in data from their own device accumulates with personal use, and for the device type, user and HCW trust accumulate with its shared professional use.

5.1.6 A consumer user overview

This approach is well explained in this consumer view, from (Anon, 2021) on Amazon® titled "good for the price, has some flaws though" reviewing a "smart scale" to measure body fat %:

"I am not really interested in numbers themselves; I just like to see them shifting if that makes sense. For example: I don't care if these scales say my body fat it's 31% … and the Fitbit Aria says 21% - although I would say just by looking at the mirror that the Aria is closer to the truth - what I really want to see is that number changing over time (which they have). We mustn't forget this device is not medical grade, therefore we must take the info for what they are, the baseline might be more or less precise but what really matters to me is the numbers shifting and you being able to track them."

Interpretating this account, it explicitly discounts absolute accuracy, for which the two devices give unacceptably different results, so they are triangulated with a third data source that is wholly qualitative : the impression from body image in the mirror "device" that is described as "the truth."

Trend analysis requires only enough precision, as minimum random errors in each device-user pair, to expose a trend in the data.

Good-enough precision requires only good-enough device design and manufacturing quality (since pairing has eliminated inter-user and inter-device errors.)

Confidence is expressed that each device has good-enough precision, while accepting their different accuracies. If accuracy (of the baseline) were more important[46], this would be expected from a "medical-grade" device; but it doesn't matter in this scenario that these devices are explicitly not so. Confidence in the (serialised) data is expressed, and trust in the device seems enough to motivate writing this positive review then publishing it.

"What really matters" (healthcare veracity) is whether the device promotes action to change behaviour. We will never know, but there is no reason to suspect it any less likely than simple scales, that are long accepted as useful; and it seems to acquire this user's trust, a necessary pre-condition.

[46] When absolute accuracy is required e.g. blood glucose, SpO_2, medical device certification suffices.

5.2 RQ2 Which healthcare data can be extended using wearables and mHealth, for early diagnosis of COVID-19?

The utility of many healthcare data for this has been analysed, of which three are:

5.2.1 Heart Rate

This derivative of conventional heart rate wholly depends on automated data acquisition from wearable mHealth devices. Three research programs have concluded that when combined with activity, it has high precision and accuracy.
While the precision and accuracy of RHR alone is insufficient to diagnose COVID-19 early, this is a high bar to meet, seldom met by conventional tests in isolation. The principle of joint use of several tests to predict a risk is common healthcare practice, and its formalisation into RP tools has been critically analysed. However its application in a RP tool for COVID-19 such as RECAP has not yet been reported.

5.2.1.1 Future clinical development of RHR

RHR's healthcare veracity, beyond the specific scenarios here, is still being established: for example, it is yet unclear if the identification of repeated 3-min measures of heart rate when resting is best determined by heart rate or by activity data, or if other time windows, or derivatives of these such as rates of rise or fall of HR, will also have clinical meaning. This meaning may initially be qualitative as veracity, but mHealth data are potentially quantitative, as accuracy and precision. For example, the Stanford researchers speculate that HR after a standardised minor stress e.g. 30sec walk may be more sensitive than 24-hour RHR tracking [47].

The use of RHR to monitor recovery or diagnose Long Covid was discussed at Section 3.4.2.1: variation between devices, such as use of different RHR definitions, has not yet been reported, partly due to low numbers of cases analysed. However, it has not invalidated the pooled data, and it remains to be seen if inter-device errors account for some of the low RHR responses found.

[47] Wearables and Early Detection of COVID-19 Using a Smartwatch - DOM Grand Rounds - 30 Sept 2020 - YouTube – at 1:01:00

5.2.2 Activity

This is the focus of lifestyle uses of motion-sensing technology, for sports and general health.

Inactivity is its converse, and also co-processed as above to identify “Resting” for RHR: it is *reducing* step count that is found in both Scripps and Stanford research as a precursor sign of infection, though again not found to be valid enough in isolation.

However, it is surely not hard to improve on “activity” expressed as serial text statements such as those in the PPSv2.

However, the frailty-related work on counting 5 min. durations of inactivity suggested that further derivatives at the inactivity end of the scale may be discovered.

5.2.3 Sleep

This requires on-device co-processing of such diverse data sources, depending on device, as activity, HR, sound, temperature and respiration rate, this latter determined by sinus arrythmia of the pulse waveform, and now even by micro-radar.

Of these, only HR and temperature may be considered simple observations rendered as numbers with standard UoM; all others are class derivatives that are not standardised, so their certainty depends on proprietary design.

5.3 RQ3 How can Extended Data be automated as Everyday data?

Two principal features of mHealth address this question:

5.3.1 Asynchronous data acquisition

MHealth data collection continuously over >24hrs is clearly of higher utility, but creates alarming problems of scale, and in the medium term of device wearability and of battery power management, with interruptions to recharge.
But its continuous nature implies liberation from the schedules of user or HCW.

5.3.2 On-device processing

Fortunately this is now feasible for

- managing data quantity and frequency of output to user/HCW
- processing native data into derivatives such as classes
- process into graphical or other non-text displays of higher value to user.

For example, data can be serialised on-device for immediate trend analysis. However, on-device processing can worsen power consumption, a key user performance feature.

24-hour data collection and on-device processing liberate the HCW from having to process native data themselves (a common fear expressed by HCW in discussion): it's now on-device, and available at times of their choosing.
Further, on-device processing supports the economic feasibility of pairing to one user: it can auto-calibrate the device to the normal parameters of one paired user, as here used for RHR, sleep, or activity, but potentially for any mHealth parameters.

This is transformational for the recognition of conditions that previously had to rely on a user's alignment with population-based parameters – that is, data from group memberships, with all the inaccuracy risks exposed in the ecological fallacy at Section 2.9.3. Only initialisation, connectivity, user-device pairing, and maintenance of the device are irreducible tasks for the user.

In summary: *automated everyday data is as transformational as the recording of music: it liberates data acquisition from synchronous encounters with HCW, and bulk processing by the device of the data can also be independent of user or HCW time.*

5.4 RQ4 How to render Everyday Data for generic use in any system

5.4.1.1 Data co-processing

Interoperabilty of data depends on the data type

- Numeric observation-type data in rational numbers with standard UoM can simply be shared online, and can be processed by many technologies, including the classic manual exchange of .csv files

  Roles of GadgetBridge and FitnessSyncer® are described, presenting this data for human use as displayed – a "human interoperability bridge" workaround for lack of interoperability, depending on the data being intelligible to the user.
- Class output is only re-usable if the same classifier is used, in the same configuration. This is shown for these two major classifier outputs
  - Sleep data: the classification of sleep quality into Deep, REM or Light is proprietary to each device and the algorithms used may be not shared even within one manufacturer's devices. So we see that in the sharing web portals the only metric that is shared is total duration – and since this is automatically sensed too, with the same uncertainty as above, the inter-device variation causes imprecision
  - Inactivity is a classifier output from activity as measured by accelerometers – see Section 3.4.6 - which is also not standardised, so not yet interoperable.

Can academia or the NHS set standards for this data and procure compliant systems?

5.4.1.2 Privacy

The controller of this data should have the same ability to access and re-use as traditional user-created healthcare data.

Thus the global industry is once again accompanied by a volunteer open-source community, supported by some academic researchers, showing how data can be reused by human interoperability bridging independent of proprietary platforms.

5.4.1.3 Deployment and Business factors

The commercial viability of new mHealth devices may require different features from those for healthcare. This is illustrated in the Suunto® consumer survey evaluated at Section 6.4.1. The general issues of product design and manufacture in an unregulated

global industry are out-of-scope, but design is seen to be led not by engineering but by marketing, and servicing not by repair but by product replacement.

5.4.2 The perfect pathognomonic test

It is so rare for a sign, observation or test to be *solely sufficient* for a diagnosis that it attracts the obscure description "pathognomonic." Such a sign has neither false positives nor negatives. It may surprise that even PCR tests for COVID-19 are not pathognomonic i.e. they may be *necessary* for diagnosis of COVID-19 in its many syndromes, but *insufficient* in isolation.

This is because all diagnosis relies on co-processing of other data to refine it, here as one of six syndromes, which is established here as normal clinical practice.

MHealth devices should not be expected to supply pathognomonic results, but should similarly be used with other data.

5.5 PAIDUR: a framework for evaluating data from new devices for healthcare

The global industry of mHealth devices is now well established and offers user control of data that was not previously available. It innovates fast, driven by consumer features such as usability and battery life.

However, it is independent of the formal regulated healthcare industry.

Frameworks have been critically analysed at Section 3.5 and a new framework "PAIDUR" is proposed, using 20 criteria in 6 domains, with some overlap or interaction as shown.

Both industries can use this PAIDUR framework to evaluate consumer mHealth data for healthcare use, such as in RP tools.

The development of PAIDUR and examples of usage are evaluated at 6.6.

PAIDUR: a framework for evaluating new devices and data for healthcare

	Domain		*Criteria*
1.	**P**recision		
	a.	design and QA i)	uncertainty metadata quantified
	b.	resolution	supports precision and accuracy
2.	**A**ccuracy		meaningful for healthcare
	a.	design and QA ii)	certified by FDA, MHRA, ISO, CE mark etc
	b.	personal device	one-per-user pairing to calibrate
3.	**I**nteroperate		data are computable across systems
	a.	values	numeric values are standard
	b.	data models	standardised structured numeric, or text
	c.	access	to native data and derived classes e.g. by API
	d.	uncertainty/error	metadata access e.g. by API, so can propagate
4.	**D**eploy		by or to users incl. healthcare workers
	a.	trust	data quality and meaning acceptable to HCW
	b.	price	net cost of ecosystem incl. one-per-user
	c.	initialise	HCW assist, or by patient/consumer solo
	d.	connect	use of smartphone, other PAN and Internet
5.	**U**se		usability of system of user+ app+device(s)
	a.	automate	how complete is this, need for HCW assist
	b.	UI, UX	user interface/ experience where not automated
	c.	reliable	failure rates, configure, maintain by user or HCW
	d.	power	duration of use per charge, per scenario
	e.	wear, fix	per scenario e.g. up to 5d
6.	**R**euse		data sharing with one or multiple systems
	a.	user	user control of sharing, and of multiple reuse
	b.	NHS	is healthcare-essential function Free of Charge
	c.	business	premium features chargeable to user or to NHS

5.5.1 Acronym development

It was noted that frameworks become known by their acronyms, so its new acronym should avoid inappropriate terms in other languages, or names such as widely-used registered internet domains. The six domains names are sequenced by the data journey, and chosen for a pronounceable word. If "Reuse" replaces Share, as discussed at Section 4.2, this forms "PAIDUR": this appears to be not widely used, nor meaningful in 41 languages[48], so this is selected in Sept 2021 as the latest framework title.

5.5.2 How PAIDUR creates Trust

This work explores the use of new data from new devices, so addresses the quality of their data. For mHealth devices in personal use, each user's confidence can accumulate with each successful use.

Trust can be seen as the time-integral of a series of confidence-forming processes for each user of each device. When many users develop trust and share that emotion, it can be said that the device class is trusted.
This trust can then propagate between users.

Trust will persist unless reduced, most commonly by unsuccessful use.
Without trust, there is little utility in the data, so all depends on trust developing in both users and HCW.

Table 8 at Section 6.6 shows where the Research Questions underpins its development, so the PAIDUR framework promotes trust in a new mHealth device that starts with none.

[48] On www.translatero.com

6 EVALUATIONS OR CRITICAL EXAMINATIONS

6.1 Initial feedback responses

These were immediate and brief comments on brief outlines of the work as it evolved, by Subject Matter Experts, suggesting System One or intuitive thinking (Kahneman, 2015) that is here valuable precisely because of their expertise.

1 What do you mean by Computable?

This drove Research Question 1 that answers at length.

2 Prefer Reuse to Share

This word replacement was on reflection a key issue: that permission to "share" is usually intended to mean once, with a specified target - not indefinite repeat dissemination to unknown others, per Section 3.2.5.

3 Clinical Frailty Scale recited with intonation: "a bit frail, quite frail, definitely frail, quite frail, more frail than that, very frail, really really frail, dying anyway "

This exposes the non-verbal meanings of spoken narrative data, lost as text.

4 Emergency Medicine Dr: "I don't want my ER full of people attending with abnormal HRV!"

This applies to many other "inappropriate" presentations to EM depts, such as a single high BP. This common NHS experience is not a new issue for mHealth – for which on-device processing may derive advice or actions that manage this risk earlier.

5 GP: "so long as it's not extra work"

Asynchronous acquisition and on-device processing reduce HCW work.

Improved diagnosis and care reduce HCW work.

6 GP: "how would you manage an isolated HR abnormality detected by FitBit?"

As any other single data item: acquire more context data.

#4 was expressed with some anger, and it was noted that emotion was expressed in several of these comments. That areas of confusion were often attended by anxiety or anger is an expression of frustration. This led to the distinction drawn here between

- the cognition of confidence, that is generated with each successful use of a type of data. Some is clearly quantified, as in “confidence limits”; or may be quantifiable as uncertainty
- the emotion of trust, that is essentially qualitative. This becomes attached to the device if it is known as the common factor in multiple data processing events that each generate confidence.

The heuristic value of this distinction is that recognising an emotion can lead to a more rational reconsideration, such as that emotions may have become attached to the wrong object, as well established since Sigmund Freud.

If the HCW feels anger towards a new device, because on other occasions its use has produced poor data, this is problematic on many levels, including that the data on this occasion may be valid. It may also introduce conflict with the user whose positive trust has developed over long use of seemingly valid data.

6.2 Four Frameworks for managing new data

6.2.1 List-of-Criteria models

The first frameworks drafts were List-of-Criteria in table format, derived from each mindmap that was developed for each research question to iteratively scope and define its content – see Section 10.4.3

RQ2 Extending data v1 this is an early draft of one list-of-criteria framework

Table 4: extending data v0.8

Data /Device	**Source**			**Acquire**	**Trans ducer**	**Pro cess**	**Re Use**	**Data Qlty**	**Device Qlty**
	User alone	**HCW input**	**wrble**	**Min/max**				**0-9**	**0-9**
PROMs	User	-	-	User	User	User	CSV	8	9
Temp	User	R,O	A	User/auto	SS	Auto	Auto	8	8
Glu sensor	User	R,O	A	User/auto	Bio	Auto	Auto	9	9
SpO_2 finger	User	R,O	-	User	PPG	User	Num	9	9
SpO_2 wrist	User	O	A	Auto	PPG	Auto	Auto	7	8
Sleep	User	R,O	A	User/auto	PPG	Auto	Auto	4	6
HR	User	R,O	A	User/auto	PPG	Auto	Auto	9	9
HRV	User	R,O	A	Auto	PPG	Auto	Auto	7	7
RHR	User	R,O	A	Auto	PPG	Auto	Auto	6	7
Inactivity	User	O	A	Auto	XLR	Auto	Auto	5	9
Resp	User	R,O	A	User/auto	PPG	Auto	Auto	8	8
Housing	User	O	-	User/auto	User	User	Num	9	9
Location	*User*	*O*	*A*	*Auto*	*GPS*	*Auto*	*Auto*	*9*	*9*
Phonics	*User*	R,O	-	*User*	*User*	*User*	*CSV*	*5*	*5*
Smell	*User*	-	-	*User*	*User*	*User*	*CSV*	*5*	*4*
C19 test	*User*	R,O	-	*User*	*Bio*	*Auto*	*Num*	*8*	*8*

12 data sources were initially considered, listed in 3 types

1. Input from the **U**ser alone, as for traditional symptoms, or some PoC devices
2. Input by HCW **R**equired, as for signs, and other PoC devices, or **O**ptional per context, such as to assist user input whenever HCW is available
3. Data acquired **A**utomatically from wearable mHealth device. This is considered for degree of automation (RQ3) in best-case scenarios, or fallback to user.

However, 4 further *potential new* device technologies – see Section 3.4.8 - were added to explore each criterion.
Source ranged from "user" i.e. a value formed in the user's mind (a symptom) or a PROM score, to solid-state sensors for temperature, motion or PPG, bio sensors for glucose or COVID-19 RNA, or GPS or beacon data for location.

Processing, and ReUse factors were found very difficult to evaluate. It became clear that auto-processing on-device is a pre-condition for automatic onward reuse: for example, a finger oximeter must be user-applied for each reading, so cannot be fully automated, whereas a wrist- or band oximeter can be worn continuously. However, these are less reliable, accurate or precise, and more expensive.
On the cost issue, however, finger oximeters have long been inexpensive enough to be bought by consumers, so accuracy improves from user-device pairing, such as calibrating for skin tone. Continuous-worn oximetry is still limited to intermittent data acquisition, but is now much less expensive and included in $25 smartwatches[49].

Similarly, for RQ3 on automating data, the same 12 current and 4 new devices were considered, and rated for likely supplier and agent of installation, configuration and maintenance, and initial activation.

[49] L8 smartwatch + Fundo software – see 10.7.1

RQ3 Data automation

Table 5: data automation v1

Item	Supply	Instl, cfg, mntn	Activate	Retain	Everyday rating	Data Qlty	Device Qlty	Qlty total
PROMs	NHS	User	User	n/a	•••	••••	•••••	12
Temp	User	User	User/auto	•••	••/••••	••••	••••	13
Glu	User	User	User/auto	••••	•••••	•••••	•••••	19
SpO_2 fngr	NHS	NHS	User	•••	••	•••••	•••••	15
SpO_2 bnd	User	User	User/auto	••••	••/••••	•••	••••	13
Sleep	User	User	User/auto	••••	••/••••	••	•••	11
HR	User	User	User/auto	••••	••/••••	•••••	•••••	16
HRV	User	User	User	••••	••	•••	•••	12
RHR	User	User	User/auto	••••	••/••••	•••	•••	12
Inactivity	User	User	Auto	••••	•••••	•••	•••••	17
Resp	User	User/NHS	User/auto	••••	••/••••	••••	••••	14
Housing	User/NHS	User	Week	•••••	••••	•••••	•••••	19
Location	*User*	*User*	*Auto*	•••••	•••••	•••••	•••••	*20*
Phonics	*user*	*User/NHS*	*5d*	*?*	*?*	•••••	•••••	*?*
Smell	*user*	*NHS*	*5d*	*n/a*	*?*	•••	•••••	*?*
C19 test	*User/NHS*	*User*	*5d*	*n/a*	•••	••••	•••	*?*

An overall rating for data quality, and for device quality, was made, initially on a 10-point scale, then reduced to 5-point, rendered as blobs for easier visual evaluation.
Sensor retention (fixation/stability) is also rated, as a key part of automation.
Device safety proved too difficult to evaluate, as "Clinical Safety Cases" have not generally been made for mHealth devices: this CSC for pulse oximetry is only for the edge-case hazard scenario of inappropriate placement, not for the device itself. (*Patient Safety Improvement.Nhs.Uk/Resources/Patient-Safety-Alerts*, n.d.)
However, it is suggested that PAIDUR may assist in the scoping of hazards for CSC development.
User criteria were rated for the worst-case low-skill user, considering the range of clinical scenarios, as at 2.4, in which the same device may be used; it includes those in a domestic care setting who may become very ill at times when supervision or assistance from carers is unscheduled and intermittent.

Design features are valid independent of specific device' use of current technologies that may not rate well in other parts of the framework. For instance, for continuous $Sp0_2$ data, ring- or band-wearing design for continuous sensor retention/stability is valid, although due to battery performance and cost issues the current Oura® ring device may not be a candidate to deliver this.

While wearability is required for Everyday use, for only five day use, as in scenarios c and d, users may tolerate a lower grade of wearability, or more frequent interruptions for recharging.

Overall quality values were noted to be very weak estimates, due (to use the metrology paradigm of this work) to propagation of uncertainty when processing data of very low resolution. But they may suffice to order the technologies: the most highly rated for automation seem to be

- continuous glucose monitoring (for known cases)
- location tracking by GPS, for case-finding (in a consenting population.)

This is already deployed on a global scale, but is not part of the Exposure Notification Systems API of Apple and Google due to privacy concerns – per Section 3.4.9.1.

This led to consideration of adding a privacy criterion - but to reduce this to a 5-blob scale seemed absurd.

So it was soon clear that Lists-of-Criteria had served some purpose in identifying issues from some devices that should apply to others, but need much more work.

6.2.2 Dynamic Model: the NASSS Framework

The development of the NASSS Framework was considered at Section 3.5

This framework is a Dynamic Model type, using 7 domains with 22 sub-domains, organised with cross-referencing as per dotted lines, and intended to be iterative.

It was initially explored as a flat list-of-criteria for 3 known and 1 new technology:

Table 6: NASSS criteria applied to 4 devices

1 Condition	**RHR**	**Activity**	**Geolocation**	**Smell (new)**
Nature	COVID-19	=	=	=
Co-morbidities	++	=	=	=
2 Technology				
Material	sensor device	sensor device	Smartphone	New kit
Data type	Continuous	Continuous	Continuous	Scheduled
Knowledge	+	+	+	+
Tech supply	User / NHS	User/NHS	User	NHS
IP owner?	Tech/ User	Tech/User	User	user
3 Value proposition				
Developer	£	£	0	£
User	££	££	=	£
4 Adopter system				
Staff	+	+	=	=
User	Passive	Passive	Passive	Active
Carers	?	?	=	?
5 Healthcare org.				
Innovation	+	+	0	+
Change appetite	+	+	0	+
Funding	+	+	0	=
Routines	+	+	0	+
Management Work	0	0	0	0
6 Wider system				
Political/policy	=	=	=	=
Regulatory/legal	MHRA	MHRA	0	0
Professional	?	?	?	?
Socio-cultural	?	?	?	?
7 Embedding over time				
Adaption	?	?	?	?
Resilience	?	?	?	?

Legend: =, ?, 0, n/a: same as current / no change + : more than current

#1 Conditions is not required here as COVID-19 is specified for all devices, but would apply if a technology-first approach is considered for multiple conditions.

#2 Technology would require expansion to add the technical considerations here, and was the only class covering the 2nd Research Question for data extension.

#3 Value propositions are incalculable in this project; however at a more abstract level business drivers apply to both supplier and user, and are considered.

#4 Adopter system changes are required of staff, user and HCW (including carers.)

#5 Healthcare organisation change management, such as Policies and Procedures for routine healthcare use of new data is further work.
However, it is expected that, due to asynchronous on-device processing to create class derivatives meaningful for healthcare action, only the most high-value derived data will be presented to users, and further filtered before to HCW, and thus be effective in reducing HCW workload, when deployed at mass scale.

#6 “Wider system” has some elements relevant here.

#7 Embedding over time was found too difficult to apply here.

It is not clear how the acronym NASSS is derived from these 7 domains (which do not form an acronym.)

In short, applying all 22 issues to these four new data sources was not appropriate for this small-scale project, and it was decided to develop a more heuristic Technical Framework, that would expand #2 Technology with the significant factors addressed in this work.

6.2.3 Targeted List-of-Criteria

This uses only mHealth device features as target criteria in a structured list to be used with each device; it does not attempt to compare multiple devices within the diagram.

The 5-blob ordinal ranking was replaced by 3-item Red / Amber / Green indicators for display as RAG , representing discrete items rather than a 5-point scale applied to one item i.e. negative features show as red to distinguish from a lowered score of positive features.

This should be used in contexts identified with the six clinical scenarios for managing COVID-19 per Section 2.4.

The specific criteria were organised by Research Question into 3 sets.

6.2.3.1 RQ1 What is Computable Data? see Xmind RQ1

1. Structure of data? If numeric — RAG
    - i. is it continuous? — G
    - ii. Is it ordinal or cardinal? — A
    - a. If text, is it
        - i. coded? — G
        - ii. semi-structured e.g. scales — A
        - iii. narrative text — R
2. Can it be machine-processed? — G
3. IG: Is data reuse
    - i. controlled by user? — G
    - ii. FOC — G
    - iii. chargeable? — A
4. Device features
    - a. Can it be deployed one-per-user / be not shared? — G
    - b. Does it require — HCW support OR user alone?
        - i. Finance — A — A
        - ii. Install, config — R — G
        - iii. Maintain incl. recharge — R — G
5. Metrology
    - a. Device Certification else — G
    - b. Can the Data Quality be quantified? — A
        - a. Precision — G
            - i. Reliable — A
            - ii. Repeatable — A
            - iii. Reproducible — A
        - b. Accuracy calibration / normalisation / pairing to individual — G
        - c. Valid other DQ features — A
6. Usability (real-world metrology when in consumer use) — G

RAG legend

R	red	problematic, stop
A	amber	uncertain
G	green	good to go

6.2.3.2 RQ2 Extending data see Xmind RQ2

1. Device features RAG
    a. How is output reusable
        i. manual R
        ii. automatic A
        iii. PAN > smartphone A
        iv. mHealth portal > Internet G

2. Associated systems: PAN, Internet
    a. Do these need supply and deploy? R

3. Can data be co-processed to form derivatives like Risk Predictors? A

6.2.3.3 RQ3 Everyday Data see Xmind RQ3 --Data/device features

1. Wearable retention for RAG
        i. 1 day A
        ii. <5d G
        iii. >= 5d G

2. Operation of data acquisition
    a. Is data output
        i. ad-hoc A
        ii. scheduled A
        iii. continuous? G
    b. does it require user operation? A
        i. does it require HCW assistance? R
        ii. can it be scheduled for 5d? G
        iii. fully automatic? G
    c. is wearability stable enough? G

Can data be co-processed to form derivatives like Risk Predictors? A

The process of constructing these criteria helped to prioritise them.

However, the level of detail knowledge required to answer them is beyond most potential users. User evaluations may be much influenced by promotional material from suppliers. This will highlight benefits but not defects, such as the issues of fixation for all wrist-worn devices, or power management.

This was evaluated here as a Framework initially called “PADIS”:

Precision, Accuracy, Deployment, Interoperability, Sharing

6.3 Acceptability of mHealth data into routine practice

This opportunity to evaluate PADIS was a live conference presentation, which required further condensation of issues for rapid use, and was an opportunity to interact with a group of front-line HCW[50] [51] despite COVID-19 disruption, and for their evaluation. An attendance of around 100 had been expected as in earlier years, but only 18 contributed. The presentation can be accessed near the end of this (long) page showing all conference presentations, as #15 titled “New data for new diseases”

Figure 15: New Data for New Diseases: a Framework

SNUG Members Day 2021 | Scottish National Users Group (snughealth.org.uk)

Four online polls were conducted, and showed:

[50] from which role the author had retired 6y earlier

[51] A new skill was learned: to use online Microsoft Forms® during a live Powerpoint® presentation while hosted on the ENU account (and so a guest account for the NHS-hosted conference organisers)

1. Have you ever at any time worn a Fitness Tracker or Smartwatch?

Yes / No / Maybe 13 / 4 / 1

This showed much more familiarity in Scottish Primary Care attendees for this topic, than in a US general population.

2 . How often do you recheck with your own device a reading from a patient's device? Consider in the last year

Never / seldom / sometimes / often / almost always 2 /1 /3 /2 /1

For confidence in data from a patient's device. this shows a full range, from zero to full.

3 Which patient devices do you *always* trust as a valid part of your assessment? Select any

1. thermometer – digital 6
2. blood pressure - digital wrist or arm 4
3. glucose – continuous glucose monitor system 4
4. weight – spring scale 3
5. weight – digital 5
6. fitness tracker / smartwatch 2
7. finger pulse oximeter 5
8. other (unspec) 1

This shows lower trust in fitness tracker/smartwatch than any other device, and that digital thermometer had the highest trust.

This is surprising, as these devices have analogue sensor and processing up to the digital driver for the 7-segment display, and are usually unable to interoperate via the internet; this applies also to all but the latest "digital" BP, weight, and CGM devices. This may be due to the long-term promotional misuse of the term "digital" per Section 2.7.1: all devices here are essentially analogue with digital displays, but only the spring scale is fully so.

The one response "other" did not specify another device, so may represent No Trust in any of these.

4. For any mHealth device to be trusted by a clinician, please rate these features:

	Mean score (range is 0 - 4)
1. precision	3.4
2. cited in published work	3.5
3. brand name	2.0
4. approval by Medicines and Healthcare products Regulatory Agency MHRA / FDA etc	3.6
5. user's competence using device	3.1
6. validity for healthcare	3.4
7. high cost to patient	2.0
8. other (text)	

This shows that of these trust criteria

- the most valued is official certification
- the least valued are brand name and high cost to patient (closely related)
- citation in published work is well valued

  (there was no option for “citation” by recommendation in social media)
- precision was rated equal to accuracy (described as “validity for healthcare”)

These are very small numbers, and although the presentation had discussed the meaning of these terms, in a live session more discussion would be expected to clarify. It also suggests that

- unbranded inexpensive devices, such as “digital” thermometers, can have high trust
- fitness tracker/smartwatch devices do not have high trust, despite familiarity to these participants

This may relate to the duration of wide public use: “digital” thermometers have been in use for 30y, whereas these fitness tracker/smartwatch devices are still only 10y old, with the first mobile app in 2014, and mass consumer use only in in recent years.

6.4 User requirements

As the global industry uses consumer-driven business models, the criteria it uses to develop new devices will differ from the needs of healthcare. However during research the many consumer-oriented publications did generate criteria also of high healthcare value, specifically usability and power management.

Cost is ever-present but to be evaluated differently since bulk purchase by national health systems can bear little relation to prices paid by individual consumers, and other finance models may apply, such as leasing. An example of consumer criteria is

6.4.1 Consumer research online survey for Suunto®

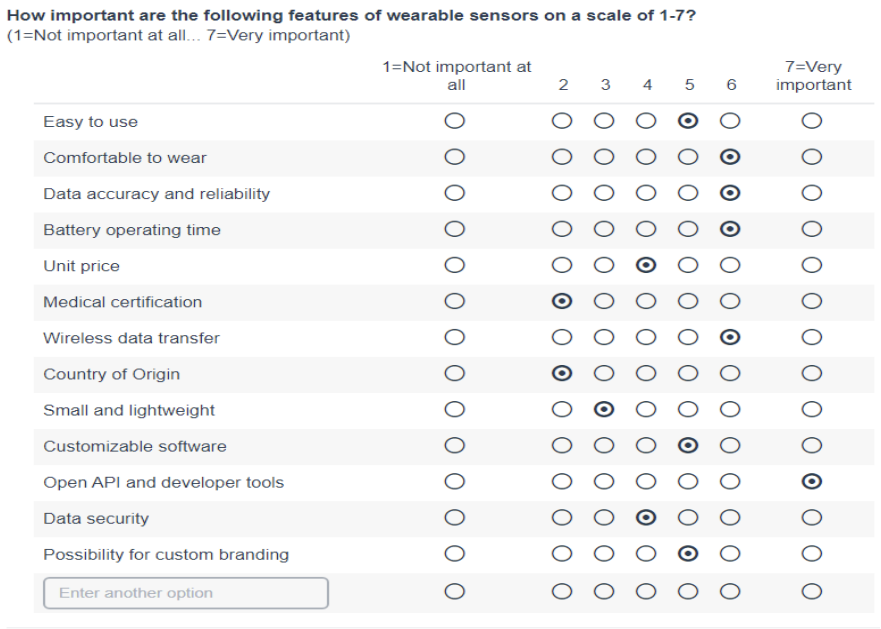


Figure 16: consumer survey for Suunto®
Start the Survey by ToinenPHD for www.suunto.com 24.3.2021

Table 7: Suunto® map to PAIDUR

Suunto criterion	Precise	Accurate	Interoperate	Deploy	Use	Reuse
Easy -to-use					1	
Comfortable					1	
Accuracy* + reliability	1	1				
Battery					1	
Med. certification		1				
Wireless data xfer				1	1	1
Country of origin				?		
Small + lightweight					1	
Customise s/w						1
Open API + tools			1			
Data security						1
Custom branding						?

* "Accuracy" was used here by Suunto in the common sense of Precision + Accuracy.

The earlier PADIS Framework was first applied to this sample list of criteria.
This and earlier evaluations then suggested separating Usability from Deploy as a major criterion, giving the current version "PAIDUR":

6.5 openEHR peer discussion on Processing of ordinal "numbers"

Please see https://discourse.openehr.org/t/revisiting-symptom-sign/1867/38?u=colinbro and following.

This discussion is inconclusive at time of writing: there is recognition that ordinal "numbers" should not be calculated, but also that so many RP tools do this that it's not the remit of HealthIT developers to question it.

6.6 PAIDUR Derivation

from Research Questions is shown here.

Table 8: PAIDUR derivation and NASSS

	RQ1 Data types	RQ2 Ext-end	RQ3 Automate	RQ4 Interop	NASSS = (Non)adoption, Abandonment, Spread, Scale-up, and Sustainability factors
Precision					
design, QA	1	1			2C data type, knowledge generated
resolution		1	1		
Accuracy					
design, QA	1	1			2E IP owner, 6B regulation
personal device	1	1	1	1	-
Interoperation					
values	1			1	-
data models				1	-
access		1			-
uncertainty		1		1	-
Deployment					2D supply models
trust	1			1	-
cost		1			3B cost
initialise			1		2B knowledge to use it
connectivity			1	1	-
Use					4 Adopters
automation			1		2C, usable 4A, 4B , 4C or solo
UI, UX			1	1	2B knowledge to use it
reliable			1		-
power			1		-
wearable, fixable		1	1		2A material factors
Reuse					6 Wider system
user			1	1	3B patient value
NHS			1	1	-
business	1	1		1	3A supplier value

6.7 PAIDUR comparison to NASSS

Table 8 above also shows PAIDUR compared to the NASSS framework per Section 3.5

PAIDUR is unique in considering these domains

- Precision incl. resolution
- Accuracy incl. user-device pairing,
- Interoperability of data in 4 subdomains

And features in more detail

- Deployment incl. trust, connectivity, HCW support (dynamic factors of embedding and adaptation)
- Usability incl. reliability, power, wearability and fixability (adopters)
- Reuse incl. user-friendly privacy control and value to NHS (wider system)

Compared to NASSS, PAIDUR does not specify co-morbidities, socio-cultural factors, or organisational capacity factors.

6.8 PAIDUR use for comparison of devices: L8 and Mi6 examples

Precision and Accuracy: Design

L8 uses Nordic® nnRF52932, with published SDK (*NRF5 SDK - Nordicsemi.Com*, n.d.)

Accelerometer sensors are said to be by Bosch®

Mi6 uses Dialog® DA1469X, with published SDK, used by (Dialog SDK10 Awesome OS, n.d.) and (Dialog SDK10 GitHub, n.d.)

Both use proprietary operating systems, but relationship to widely-used OS such as WearOS is unknown.

Both offer $Sp0_2$ data, user-initiated while sensor is firmly fixed on the wrist. Despite this, in this user's experience both are inaccurate even at normal values.

L8 offers BP data, again requiring user initiation and co-operation to stabilise the sensors.

They are already worth the attention of the global "white hat" community, but further hardware evaluation of these devices is out-of-scope here.

Precision and Accuracy: Quality Assurance of manufacture

The classic U-curve of reliability (Shewhart, 2012) should apply, and as the devices are not serviceable, for early failures replacement under guarantee is the only remedy, and for end-of-life failure, upgrade to latest version at cost to consumer. These are factored into the business model of these value devices.

Batch management probably applies, as to other digital technology: the grade of CPU batches used in the unbranded L8 or the value-branded Mi6 may be reflected in their price. The risks of faulty devices are now managed by the supply chain, instead of the manufacturer's quality assurance.

Resolution is unknown, as is error in the numeric data.

Neither device offers a means of accessing native data directly – though SDK may enable its exposure via API.

Neither device has FDA or other 3rd party assurance; Mi6 reports CE compliance only.

Accuracy as Healthcare Veracity

Many studies check the accuracy of specific mHealth devices, such as (Wallen et al., 2016) showing error of <9% compared to ECG. But there are so many devices that an explosion of studies is required, and is unlikely to address the value-branded devices.

Accuracy is a key criterion for consumer reviews, which usually compare the device data to that from reputable devices, such as sleep data from Polar® devices, that Mi6 is said to match (TechRadar, n.d.) . However, this is at the anecdotal level of the single reviewer's unblinded serial observations, so of low accuracy.

However, user-device pairing also applies, so the confidence expressed by each reviewer may accumulate for that device, as for any user, such that serial data can show trends with healthcare veracity.

L8's BP function supports user calibration with a known reference device, so that preset values are more accurate.

MiFit software is more sophisticated, including a Physical Activity Index of 35y provenance (Kieffer et al., n.d.) with graphical displays, and several levels of educational text in the app. This standardised measure was validated on user self-report, with a "secret" algorithm for PAI itself; its derivation on-device, from HR and activity data, was also proprietary, initially to Mio® devices, and may differ again for this device.

Interoperation

Both devices output native data as standard numeric values for activity, HR, RHR, BP and SpO_2. Both devices also output activity and sleep data in graphical formats and classes that are proprietary, and some numeric data is derived by proprietary algorithms, such as 1hr inactivity alerts.

Only Mi6 outputs numeric PAI data.

Data models for text are not used here, as symptom scales or PROMs are not implemented, though easy to do for an always-on device with a UI.

All this non-standard data relies on the "human interoperability bridge" to interoperate.

Deployment

Trust in each device is an integrating derivative of the above factors.

If there is no trust in the device, then all this evaluation is void.

Consumer price in Jun 21was £25 for L8 in Feb 20, and £40 for Mi6: both inexpensive enough to be supplied by NHS without charge to patients, and be considered disposable.

Initialisation is easy for a motivated consumer, but less so to a patient who would require HCW assistance.

Connectivity is poor for L8, with frequent Bluetooth pairing faults, but good for Mi6.

Use

Automation is similar: they need user action for many functions, and HCW support.

UI/UX is superior for the Mi6 + MiFit, but subject to the above limitation.

Reliability cannot be assessed on a single sample of each.

Power management is better for Mi6, with duration of a week including sleep monitoring, where L8 was only 2d. Neither offer wireless recharge.

Wearable/fixable rating was similar for both.

Reuse

L8 shares only with Weibei service, available only in China.

MiFit shares with GoogleFit and Apple Health.

Neither address any differential offer of data to other services, such as native HR to NHS.

6.9 PAIDUR Strengths and Weaknesses

Strengths

- applies metrological engineering principles to mHealth data quality
- identifies user-device pairing as a procedure essential to accuracy
- analyses real-world healthcare usage for further data quality procedures
    - repetition, triangulation, serialisation
- applies healthcare scenarios appropriate to accounts of mHealth technologies
- identifies asynchronous on-device processing as
    - adding meaning
    - liberating HCW time
- considers Propagation of Errors as a manageable engineering problem
- distinguishes Confidence and Trust
- introduces “Veracity” as a specific term for qualitative accuracy and certainty
- aligns “clinical judgment” with co-processing
- proposes how new data may contribute to RP tools
    - such as RHR for COVID-19
- identifies risks of category error in group data
    - suggests management as inaccuracy risk for populations only
- framework development embeds its application to COVID-19
- value proposition considered from user, business/industry and NHS viewpoints
- promotes user control of their own data
- acknowledges that technology development is usually industry-led and
    - shows users and designers what is now possible if healthcare-led

Weaknesses

- propagation of error metadata relies on APIs or other means
- error metadata may be considered commercially sensitive by manufacturer
- provenance of data is not normally made available
- industry drivers to design for lifestyle purposes are of greater scale and more profitable, so re-purposing these to healthcare use is a big task.

7 DISCUSSION OF OUT-OF-SCOPE OR FUTURE WORK

7.1 PAIDUR application to future devices

The three potential new devices at Section 3.4.8 may be evaluated by PAIDUR.

One traditional device, a simple digital thermometer, is also included for comparison

These are sample values to outline the concept, that normally require text to evaluate.

Table 9: PAIDUR and future devices

	Digital thermometer	CO_2 sensor	Cough phonics	Smell
Precision				
design and QA i)	••	•••••	••••	•••
resolution	••••	••••	••	••
Accuracy				
design and QA ii)	•••••	•••	••••	•••••
personal device pairing	•••••	•••••	•••••	•••••
Interoperation	-	•••••	••••	•••••
numeric values	•••••	•••••	?	ordinal
data models	•••••	•••••	?	?
access	•••••	•••••	?	?
uncertainty	•••••	•••	?	?
Deployment				
trust	•••••	?	?	?
cost	•••••	?	?	•••••
initialise / HCW	•••••	?	?	••••• / 0
connectivity	?	•••••	•••••	•••••
Use				
automation	••	•••••	••	••
UI, UX	•••••	•••••	•••••	•••••
reliable	•••••	?	?	?
power	•••••	•••••	•••••	•••••
wearable, fixable	•••	•••••	•••••	•••••
Reuse	-			
User	•••••	•••••	•••••	•••••
NHS	•••••	?	?	?
Business	•••••	?	?	?

7.2 Concepts introduced that need further research

7.2.1 Engineering

Network infrastructure capacity for central processing of mHealth data is feasible, as shown by the three platforms studied.

Quality Assurance of device is the remit of the manufacturer until specified by the purchaser e.g. for NHS use, immediate replacement of faulty devices is needed.

Power management for healthcare needs more endurance and ease of recharge.

7.2.2 Informatics

Healthcare-standard “NHS” data should be defined, such as RHR, step count, HROS, inactivity and sleep parameters.

National action, such as by standards (DIN, IEC, ISO) or healthcare organisations (MHRA, FDA, EMA) is required to standardise healthcare-standard data.

This “Standard NHS data” should be fully supported by openEHR archetypes for generic reuse.
OpenEHR templates and guidance should also be reviewed for misuse of calculations on ordinal data, as below.

Quantified error metadata should be made available via API.

Re-use across platforms needs national-level management of these businesses, promoted by pilots/demos of live interoperability between major platforms.

Narrative data of high semantic quality needs PROM presented as ePRO apps, which are now feasible for these everyday mHealth devices with a UI.

7.2.3 Information Governance

Major category errors of misidentified, missing or ecological data are here identified and need systematic IG guidance.

The “Not for Healthcare” disclaimer is widespread in mHealth: its legal status needs clarification as a defence for supplier’s liability, rather than a limitation on use.

Platforms, business models and privacy all need more work, such as whether people at high risk of COVID-19 might choose GPS tracking for social mobility despite privacy risks.

7.2.4 Healthcare management

RP tools are under continuous live development for all healthcare including COVID-19. e.g. Is rise in glucose a valid early sign of COVID-19 for current users of CGM systems?

Use of native data instead of derived classes should be promoted as good practice.

Healthcare veracity is lacking to date for

- Inactivity e.g. for periods other than 1 hour
- Resting Heart Rate e.g. utility for recovery monitor / diagnosis of Long Covid

The “Operational Paradigm” is a novel term, but its essence - to create Actions – is common to business process modelling. Task Planning, such as to identify the specific selection and timing of structured direct care actions, is a major work area in openEHR (Tom Beale, personal comms).

Validation by PAIDUR of new data in healthcare needs new Policies and Procedures.

7.2.5 Mathematical/Philosophical

The use of classes where native continuous data was machine-processable is not explained in those research papers seen, several of which are detailed here.

Alternative serial but non-numeric symbols for ordinals should be promoted, such as via peer review, though this is more difficult to apply to historic research.

Propagation of Uncertainty is shown here as a major cause of imprecision in RP tools.

“Veracity” for trueness or truth of qualitative data needs more semantic work.

8 SUMMARY and CONCLUSION

This work is a multi-disciplinary investigation of how new data from mHealth devices may be combined with current healthcare data in Risk Predictor tools for a range of conditions, and may apply to the early diagnosis of COVID-19.
The initial evidence for the earliest signs of viral infections was from FitBit® devices, and this work shows how data from multiple devices on two mHealth Personal Health Record platforms, Google Fit® and Apple Health®, can all contribute to three research platforms (DETECT at Scripps, Stanford and Robert Koch Institute) to detect COVID-19 early.

MHealth devices are commonly selected by consumers for their everyday lifestyle, and each device is paired to only one user. But when everyone needs healthcare, the criteria to select devices change. Further the workload of healthcare staff requires that only the minimum quantity of maximum quality mHealth data should be presented to them. On-device processing can output high-value derivatives of native data, such as Inactivity-for-1hr, 24-hr Resting Heart Rate, or Heart-rate-over-steps, that maximise the meaning and minimise the scale of data; but for reuse native numeric data retains the key metrological qualities of precision and accuracy, with uncertainty that should be quantified as metadata.

MHealth data are set in the context of six healthcare scenarios of the course of COVID-19 in the community, which vary the healthcare meaning of data. For this “Veracity” is introduced as a qualitative component of Data Quality. Veracity of qualitative data is an issue familiar to clinicians and managed pragmatically, principally by serialisation of data in time to standardise for bias errors, and by pairing each user with one device. All data can be time-serialised to form trends that can also contribute to algorithmic Risk Predictor tools.

Healthcare workers currently must mentally process multiple data to predict risk for conditions, such as was prototyped for frailty using algorithms which were originally simple enough for mental arithmetic. Increasingly, however, Risk Predictor tools could use machine processed data, which can be from local devices, or from remote devices if made interoperable using these PAIDUR principles.

The metrological precision and accuracy of mHealth devices far exceeds that of older technologies, yet data from those had long been included in early Risk Predictor tools by clinicians with expertise in interpreting their validity. This expertise can also apply to new data from new devices. On-device processing creates new output types and classifications, but these are not standardised, so less valid, and so less trusted.
The trust of a group of clinicians in new and older devices has been evaluated.

Deployment of devices on a mass scale is here already due to the global consumer uptake of mHealth devices for lifestyle, which has driven prices down to $25 or less for unbranded products. These use the same digital technologies, and often specific components, as branded and healthcare-certified products, and implications of these business factors for precision and accuracy are discussed. Data from multiple devices, processed on-device to simple parameters of high certainty, can be collated not only by each manufacturer's platforms but also by other online platforms specialising in acquiring data from multiple mHealth devices for personal health records.

However, the costs of these platforms are not only financial, but to privacy, which is built-in to the business models. Valid consent requires individuals to understand both benefits and risks, and they can actively manage privacy so healthcare benefit can be free-at-point-of-care, as shown by Robert Koch Institute.
Interoperability at technical level has been demonstrated by the three research platforms examined, but semantic interoperability is harder. For measurement data, reusing numeric data in standard units of measure is simpler than the non-standard and text-based class outputs that consumers prefer.

The PAIDUR framework:

Precision / Accuracy / Interoperation / Deployment / Use / Reuse

is presented to structure the functional and quality features required by new mHealth data sources to address social-technical barriers and to interoperate new data with healthcare systems.

8.1 What does “good” look like?

An mHealth device with a good PAIDUR profile would thus feature

- **P**recision of design and manufacture
- **A**ccuracy of data, with veracity for healthcare purposes
- **I**nteroperation of data in other healthcare settings that is semantically valid
- **D**eployment is cost-effective, and feasible for the unwell user
- **U**se is automated enough for varied user expertise
- **R**euse of their own data is simple for users to understand, control and trust.

9 REFERENCES

Addison, P. S., Watson, J. N., Mestek, M. L., Ochs, J. P., Uribe, A. A., & Bergese, S. D. (2014). Pulse oximetry-derived respiratory rate in general care floor patients. *Journal of Clinical Monitoring and Computing*, *29*(1), 113–120. https://doi.org/10.1007/s10877-014-9575-5

Alexa can help make strides in telecare | HeraldScotland. (n.d.). Retrieved December 3, 2020, from https://www.heraldscotland.com/opinion/17808476.alexa-can-help-make-strides-telecare/

Anon. (2021). *Metrology of Body Comp scales Consumer view*. Amazon Website. https://www.amazon.co.uk/gp/customer-reviews/R18K59J1L4EU15/ref=cm_cr_dp_d_rvw_ttl

Apperta Foundation. (n.d.). Retrieved March 2, 2020, from https://apperta.org/coPHR/

ARMED. (n.d.). *WILL ARTIFICIAL INTELLIGENCE SET THE STANDARD FOR FALLS DETECTION ?*

Asher, R. (1959). TALKING SENSE. *The Lancet*, *274*(7100), 417–419. https://doi.org/https://doi.org/10.1016/S0140-6736(59)90413-1

Atypical Covid-19 presentations in older people – the need for continued vigilance | British Geriatrics Society. (n.d.). Retrieved May 14, 2020, from https://www.bgs.org.uk/blog/atypical-covid-19-presentations-in-older-people-—-the-need-for-continued-vigilance

Basu, B., Riya, P. A., Issac, J., Parvathy, S., Nair, B. S., Tokdar, P., Kumar, D. S., Kulkarni, P. R., Hanumanram, G., Jagadeesan, M., Suthakaran, P. K., Nair, L. D. V., Vennila, R., Kannan, R., Kaimal, B., Anoop, G. K., Joseph, I., Nair, R., George, S., … Pillai, M. R. (2020). COVID Anosmia Checker. *MedRxiv*, *12*, 2020.10.28.20221200. https://doi.org/10.1101/2020.10.28.20221200

Beale, T. (2018). *Why the platform will replace today's interoperability standards in healthcare | Woland's cat*. https://wolandscat.net/2018/07/21/why-the-platform-will-replace-todays-interoperability-standards-in-healthcare/

Bhalla, N., Jolly, P., Formisano, N., & Estrela, P. (2016). Introduction to biosensors. *Essays in Biochemistry*, *60*(1). https://doi.org/10.1042/EBC20150001

BIPM Joint Committee for Guides in Metrology. (2008). JCGM 200 : 2008 International vocabulary of metrology — Basic and general concepts and associated terms (VIM). In *International Organization for Standardization Geneva ISBN* (Vol. 3, Issue Vim). http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches. *Communications of the Association for Information Systems*, *34*, 257–286. https://doi.org/10.17705/1CAIS.03412

Bramley, R. (n.d.). *Digitising the National Early Warning Score NEWS2 – Reflections of a CCIO*. Retrieved March 4, 2021, from https://clinicalwebportal.home.blog/2019/08/02/digitising-the-national-early-warning-score-news2/

CarePredict, I. (n.d.). *The Tempo Medical Alert From CarePredict*. Retrieved October 2, 2021, from https://www.caring.com/best-medical-alert-systems/tempo-medical-alert/#features-of-the-tempo-series-3-from-carepredict

Catalonia openEHR. (n.d.). Retrieved July 31, 2021, from https://openehr.org/news_events/openehr_news/341

Chao, Y.-S. S., Wu, H.-C. C., Wu, C.-J. J., & Chen, W.-C. C. (2018). Index or illusion: The case of frailty indices in the Health and Retirement Study. *PLoS ONE*, *13*(7), 1–19. https://doi.org/10.1371/journal.pone.0197859

Chute, C., & French, T. (n.d.). *Care 4.0*.

Clegg, A., Bates, C., Young, J., Ryan, R., Nichols, L., Ann Teale, E., Mohammed, M. A., Parry, J., & Marshall, T. (2016). Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age and Ageing*, *45*(3), 353–360. https://doi.org/10.1093/ageing/afw039

Clegg, A., Young, J., Iliffe, S., Rikkert, M. O., & Rockwood, K. (2013). Frailty in elderly people. *The Lancet*, *381*(9868), 752–762. https://doi.org/10.1016/S0140-6736(12)62167-9

Clinical Knowledge Manager. (n.d.). Retrieved January 13, 2020, from https://www.openehr.org/ckm/archetypes/1013.1.3334/mindmap

CO_2 Sensors in Smart Phones: a ULISSES Webinar. (n.d.). Retrieved July 31, 2021, from https://www.ulisses-project.eu/news/?post_id=15&title=co2-sensors-in-smart-phones:-a-ulisses-webinar

Cohrs, R. J., Martin, T., Ghahramani, P., Bidaut, L., Higgins, P. J., & Shahzad, A. (2015). Translational medicine definition by the European society for translational medicine. In *New Horizons in Translational Medicine* (Vol. 2, Issue 3, pp. 86–88). Elsevier Ltd. https://doi.org/10.1016/j.nhtm.2014.12.002

Coiera, E. (2015). Guide to Health Informatics. In *Guide to Health Informatics, Third Edition* (3rd ed.). CRC Press. https://doi.org/https://doi.org/10.1201/b13617

Complex adaptive systems Evidence scan. (2010).Retrieved 30.4.2021 from https://www.health.org.uk/publications/complex-adaptive-systems

Corona Apps: This is how Countries are Dealing with Privacy (in Times of Covid-19) | VPNoverview.com. (n.d.). Retrieved August 30, 2021, from https://vpnoverview.com/news/corona-apps-worldwide-implementation-and-privacy/

Create a Question and Answer Bot with Amazon Lex and Amazon Alexa | AWS Machine Learning Blog. (n.d.). Retrieved February 10, 2020, from https://aws.amazon.com/blogs/machine-learning/creating-a-question-and-answer-bot-with-amazon-lex-and-amazon-alexa/

Dambha-Miller, H., Everitt, H., & Little, P. (2020). Clinical scores in primary care. *British Journal of General Practice*, *70*(693), 163–163. https://doi.org/10.3399/BJGP20X708941

Data Mining - Classification & Prediction - Tutorialspoint. (n.d.). Retrieved March 26, 2021, from https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

De Dombal, F. T., Leaper, D. J., Staniland, J. R., Mccann, A. P., & Horrocks, J. C. (1972). Computer-aided Diagnosis of Acute Abdominal Pain. *British Medical Journal*, *2*(5804), 9–13. https://doi.org/10.1136/bmj.2.5804.9

DHC Wales. (n.d.). *Wales contract for clinical data repository*. Retrieved July 9, 2021, from https://htn.co.uk/2021/07/07/digital-health-and-care-wales-awards-contract-for-clinical-data-repository/

Dialog SDK10 Awesome OS. (n.d.). *Ble_sdk10_examples*. Retrieved October 28, 2021, from https://awesomeopensource.com/project/dialog-semiconductor/BLE_SDK10_examples

Dialog SDK10 GitHub. (n.d.). *GitHub - dialog-semiconductor/BLE_SDK10*. Retrieved October 28, 2021, from https://github.com/dialog-semiconductor/BLE_SDK10_examples#readme

Exposure Notifications Frequently Asked Questions Preliminary-Subject to Modification and Extension. (2020).

Fargas, B. C., & Petersen, M. N. (2017). GPS-free geolocation using LoRa in low-power

WANs. *GIoTS 2017 - Global Internet of Things Summit, Proceedings*. https://doi.org/10.1109/GIOTS.2017.8016251

Fiorentino, F., Prociuk, D., Belen, A., Gonzalez, E., Neves, A. L., Husain, L., Ramtale, S., Mi, E. E., Mi, E. E., Macartney, J., Anand, S., Sherlock, J., Saravanakumar, K., Mayer, E., de Lusignan, S., Greenhalgh, T., Delaney, B., Espinosa Gonzalez, A. B., Neves, A. L., … Delaney, B. (2021). An early warning risk prediction tool for patients diagnosed with COVID-19: the statistical analysis plan for RECAP V1. *JMIR Research Protocols*. https://doi.org/10.2196/preprints.30083

Flaatten, H., & Clegg, A. (2018). Frailty: we need valid and reliable tools in critical care. *Intensive Care Medicine*, *44*(11), 1973–1975. https://doi.org/10.1007/s00134-018-5404-5

FluPhone project: Computer Laboratory. (n.d.). Retrieved August 11, 2021, from https://www.cl.cam.ac.uk/research/srg/netos/projects/archive/fluphone2/

Fowler's Dictionary of Modern English Usage. (2015). In *Fowler's Dictionary of Modern English Usage*. Oxford University Press. https://doi.org/10.1093/ACREF/9780199661350.001.0001

Fried, L. P., Tangen, C. M., Walston, J., Newman, A. B., Hirsch, C., Gottdiener, J., Seeman, T., Tracy, R., Kop, W. J., Burke, G., & McBurnie, M. A. (2001). Frailty in Older Adults: Evidence for a Phenotype. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *56*(3), M146–M157. https://doi.org/10.1093/gerona/56.3.M146

Google. (2021). *Soli: Home*. https://atap.google.com/soli/

Greenhalgh, T., A'Court, C., & Shaw, S. (2017). Understanding heart failure; explaining telehealth - A hermeneutic systematic review. *BMC Cardiovascular Disorders*, *17*(1), 156. https://doi.org/10.1186/s12872-017-0594-2

Greenhalgh, T., Procter, R., Wherton, J., Sugarhood, P., Hinder, S., & Rouncefield, M. (2015). *What is quality in assisted living technology? The ARCHIE framework for effective telehealth and telecare services*. https://doi.org/10.1186/s12916-015-0305-8

Greenhalgh, T., Wherton, J., Papoutsi, C., Lynch, J., Hughes, G., A'Court, C., Hinder, S., Fahy, N., Procter, R., & Shaw, S. (2017). Beyond adoption: A new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of Medical Internet Research*, *19*(11), e367. https://doi.org/10.2196/jmir.8775

Guidelines, American, T. N., & Guidelines. (1996). Guidelines Heart rate variability. *European Heart Journal*, *17*, 354–381. https://doi.org/10.1161/01.CIR.93.5.1043

Hanlon, P., Nicholl, B. I., Jani, B. D., Lee, D., McQueenie, R., & Mair, F. S. (2018). Frailty and pre-frailty in middle-aged and older adults and its association with multimorbidity and mortality: a prospective analysis of 493 737 UK Biobank participants. *The Lancet Public Health*, *3*(7), e323–e332. https://doi.org/10.1016/S2468-2667(18)30091-4

Hayano, J., Yoshida, Y., & Yuda, E. (2017). Characteristics of basal heart rate during daily life: Relationships with age, sex, and mean heart rate. *ACM International Conference Proceeding Series*, 492–495. https://doi.org/10.1145/3154862.3154900

Health and Social Community Care | SPARRA | SPARRA Online | Health Topics | ISD Scotland. (n.d.). Retrieved January 13, 2020, from https://www.isdscotland.org/Health-Topics/Health-and-Social-Community-Care/SPARRA/SPARRA-Model/index.asp?Co=Y

Healthcare Improvement Scotland. (2020). *Hospital at Home*. 2020205-hospital-at-

home-guiding-principles.pdf (ihub.scot)
Heard, S. (2010). The use of standard content specifications in a national health interoperability framework. *Electronic Journal of Health Informatics*.
Heneghan, C., Venkatraman, S., Russell, A., Heneghan, C., Venkatraman, S., & Russell, A. (2019). Investigation of an estimate of daily resting heart rate using a consumer wearable device. *MedRxiv*, 19008771. https://doi.org/10.1101/19008771
HL7_FHIR. (n.d.-a). *Condition definition*. Retrieved December 10, 2020, from https://www.hl7.org/fhir/condition.html
HL7_FHIR. (n.d.-b). *Observation*. Retrieved December 10, 2020, from https://www.hl7.org/fhir/condition.html
Hodkinson, A., Kontopantelis, E., Adeniji, C., van Marwijk, H., McMillan, B., Bower, P., & Panagioti, M. (2019). Accelerometer- and Pedometer-Based Physical Activity Interventions Among Adults With Cardiometabolic Conditions. *JAMA Network Open*, *2*(10), e1912895. https://doi.org/10.1001/jamanetworkopen.2019.12895
Hubbard, R. E., Meier, A., Hilmer, S., Naganathan, V., Etherton-Beer, C., & Rockwood, K. (2020). *Frailty in the Face of COVID-19 | Age and Ageing | Oxford Academic*. Age and Ageing. https://academic.oup.com/ageing/advance-article/doi/10.1093/ageing/afaa095/5831147
IEEE. (n.d.). *IEEE Standard Computer Dictionary*. Retrieved October 19, 2021, from https://ieeexplore.ieee.org/document/182763
InvenSense. (n.d.). *Gyroscopic motion sensors*. Retrieved May 8, 2021, from https://web.archive.org/web/20141022104500/http://www.invensense.com/mems/principles.html
Is the Brain Analog or Digital? | NeuroLogica Blog. (n.d.). Retrieved March 5, 2021, from https://theness.com/neurologicablog/index.php/is-the-brain-analog-or-digital/
Issues of validity in qualitative research. - PsycNET. (n.d.). Retrieved July 26, 2021, from https://psycnet.apa.org/record/1989-98606-000
Jacob, S. N., & Nieder, A. (2008). The ABC of cardinal and ordinal number representations. In *Trends in Cognitive Sciences* (Vol. 12, Issue 2, pp. 41–43). Elsevier Current Trends. https://doi.org/10.1016/j.tics.2007.11.006
Johansson, M. A., Quandelacy, T. M., Kada, S., Prasad, P. V., Steele, M., Brooks, J. T., Slayton, R. B., Biggerstaff, M., & Butler, J. C. (2021). SARS-CoV-2 Transmission From People Without COVID-19 Symptoms. *JAMA Network Open*, *4*(1), e2035057. https://doi.org/10.1001/jamanetworkopen.2020.35057
Johnston, L., Hockley, J., Henderson, D. A. G., & Shenkin, S. D. (2020). The Development of a Care Home Data Platform in Scotland: Insights from the Care Home Innovation Partnership, Lothian. *MedRxiv*, 2020.08.17.20176503. https://doi.org/10.1101/2020.08.17.20176503
Jouffroy, R., Jost, D., Prunet, B., R, J., D, J., B, P., Jouffroy, R., Jost, D., & Prunet, B. (2020). Prehospital pulse oximetry: a red flag for early detection of silent hypoxemia in COVID-19 patients Romain. *Critical Care (London, England)*, *24*(1), 313. https://pubmed.ncbi.nlm.nih.gov/32513249/
Kahneman, D. (2015). Thinking, Fast and Slow. *College Music Symposium*, *55*. https://doi.org/10.18177/sym.2015.55.ca.10990
Kang, M., Xin, H., Yuan, J., Ali, S. T., Liang, Z., Zhang, J., Hu, T., Lau, E. H. Y., Zhang, Y., Zhang, M., Cowling, B. J., Li, Y., & Wu, P. (2021). Transmission dynamics and epidemiological characteristics of Delta variant infections in China. *MedRxiv*, 2021.08.12.21261991. https://doi.org/10.1101/2021.08.12.21261991
Kieffer, S., Zisko, N., Coombes, J., Nauman, J., & Wisloff, U. (n.d.). *Personal Activity*

Intelligence and Mortality: The HUNT study. Retrieved October 28, 2021, from https://reader.elsevier.com/reader/sd/pii/S0025619618303215?token=D0C9AAED6CDC420F71E7B5ABCF5E99597A137A848BC98A4B2A84CFFD28D961CEAA73813835670C7E487C622DA479ADA1&originRegion=eu-west-1&originCreation=20211028130948

Koikkalainen, H. (2019). *Modelling Anticipatory Care Plans using openEHR,*. Edinburgh Napier Uni.

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 275–281. https://doi.org/10.1109/ojemb.2020.3026928

Larrabee, R. D., & Postek, M. T. (1993a). Precision, accuracy, uncertainty and traceability. *Solid State Electronics*, *36*(5), 673–684. https://doi.org/10.1016/0038-1101(93)90234-H

Larrabee, R. D., & Postek, M. T. (1993b). Precision, accuracy, uncertainty and traceability and their application to submicrometer dimensional metrology. In *Solid State Electronics* (Vol. 36, Issue 5). Pergamon. https://doi.org/10.1016/0038-1101(93)90234-H

Lau, F., Maida, V., Downing, M., Lesperance, M., Karlson, N., & Kuziemsky, C. (2009). Use of the Palliative Performance Scale (PPS) for End-of-Life Prognostication in a Palliative Medicine Consultation Service. *Journal of Pain and Symptom Management*, *37*(6), 965–972. https://doi.org/10.1016/j.jpainsymman.2008.08.003

Living with COVID-19. (2020). https://doi.org/10.3310/THEMEDREVIEW_41169

Logan, N., Reilly, J., Grant, S., & Paton, J. (2000). Resting heart rate definition and its effect on apparent levels of physical activity in young children. *Medicine and Science in Sports and Exercise*, *32*(1), 162–166. https://doi.org/10.1097/00005768-200001000-00024

Maley, C. J. (2011). Analog and digital, continuous and discrete. *Philosophical Studies*, *155*(1), 117–131. https://doi.org/10.1007/s11098-010-9562-8

McCarthy, J. (2019). *One in Five U.S. Adults Use Health Apps, Wearable Trackers*. Www.Gallup.Com. https://news.gallup.com/poll/269096/one-five-adults-health-apps-wearable-trackers.aspx

McKinstry, B., Alexander, H., Maxwell, G., Blaikie, L., Patel, S., Guthrie, B., & Group, T. (2021). The Use of Telemonitoring in Managing the COVID-19 Pandemic: Pilot Implementation Study. *JMIR Form Res 2021;5(9):E20131 Https://Formative.Jmir.Org/2021/9/E20131*, *5*(9), e20131. https://doi.org/10.2196/20131

McMahon, DJ. (2013). *There-is-no-such-thing-as-a-SpO_2-simulator* (Vol. 510). Accessed 21.5.2021 https://www.flukebiomedical.com/resource/theres-no-such-thing-spo_2-simulator

Microwear L8 Smartwatch – Specs Review - SmartWatch Specifications. (n.d.). Retrieved November 14, 2020, from https://www.smartwatchspecifications.com/Products/microwear-l8-smartwatch-specs-review/

Miles, A., Webb, T. E., Mcloughlin, B., Mannan, I., Rather, A., Knopp, P., & Davis, D. (2020). Outcomes from COVID-19 across the range of frailty: excess mortality in fitter older people. *MedRxiv*, 2020.05.22.20110486. https://doi.org/10.1101/2020.05.22.20110486

Mishra, T., Wang, M., Metwally, A. A., Bogu, G. K., Brooks, A. W., Bahmani, A., Alavi, A., Celli, A., Higgs, E., Dagan-Rosenfeld, O., Fay, B., Kirkpatrick, S., Kellogg, R., Gibson,

M., Wang, T., Hunting, E. M., Mamic, P., Ganz, A. B., Rolnik, B., … Snyder, M. P. (2020). Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering*, *4*(12), 1208–1220. https://doi.org/10.1038/s41551-020-00640-6

Morley, J E, Malmstrom, T. K., & Miller, D. K. (2012). A simple frailty questionnaire (FRAIL) predicts outcomes in middle aged African Americans. *The Journal of Nutrition, Health & Aging*, *16*(7), 601–608. https://doi.org/10.1007/s12603-012-0084-2

Morley John. (n.d.). *Academic Phrasebank*. Retrieved July 27, 2021, from https://www.phrasebank.manchester.ac.uk/

Morley, John E., Vellas, B., Abellan van Kan, G., Anker, S. D., Bauer, J. M., Bernabei, R., Cesari, M., Chumlea, W. C. C., Doehner, W., Evans, J., Fried, L. P., Guralnik, J. M., Katz, P. R., Malmstrom, T. K., McCarter, R. J., Gutierrez Robledo, L. M., Rockwood, K., von Haehling, S., Vandewoude, M. F., & Walston, J. (2013). Frailty Consensus: A Call to Action. *Journal of the American Medical Directors Association*, *14*(6), 392–397. https://doi.org/10.1016/j.jamda.2013.03.022

Morris, D. E., Pepe, M. S., & Barlow, W. E. (2010). Contrasting two frameworks for ROC analysis of ordinal ratings. *Medical Decision Making*, *30*(4), 484–498. https://doi.org/10.1177/0272989X09357477

Mung (computer term) - Wikipedia. (n.d.). Retrieved July 30, 2021, from https://en.wikipedia.org/wiki/Mung_(computer_term)

Nandy, S., Parsons, S., Cryer, C., Underwood, M., Rashbrook, E., Carter, Y., Eldridge, S., Close, J., Skelton, D., Taylor, S., & Feder, G. (2004). Development and preliminary examination of the predictive validity of the Falls Risk Assessment Tool (FRAT) for use in primary care. *Journal of Public Health*, *26*(2), 138–143. https://doi.org/10.1093/pubmed/fdh132

Naska, A., Oikonomou, E., Trichopoulou, A., Psaltopoulou, T., & Trichopoulos, D. (2007). Siesta in Healthy Adults and Coronary Mortality in the General Population. *Archives of Internal Medicine*, *167*(3), 296–301. https://doi.org/10.1001/ARCHINTE.167.3.296

Natarajan, A., Su, H.-W., & Heneghan, C. (2020). Assessment of physiological signs associated with COVID-19 measured using wearable devices. *Npj Digital Medicine*, *3*(1), 156. https://doi.org/10.1038/s41746-020-00363-7

National Early Warning Score (NEWS) 2 | RCP London. (n.d.). Retrieved March 2, 2021, from https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

NIST. (n.d.). *Propagation of error*. Retrieved June 15, 2021, from https://www.itl.nist.gov/div898/handbook/mpc/section5/mpc55.htm

NIST TN 1297: Appendix A. Law of Propagation of Uncertainty | NIST. (n.d.). Retrieved June 23, 2021, from https://www.nist.gov/pml/nist-technical-note-1297/nist-tn-1297-appendix-law-propagation-uncertainty

Nokia, W. (n.d.). *Under-mattress sleep tracker - Sleep Analyzer | Withings*. Retrieved October 24, 2021, from https://www.withings.com/uk/en/sleep-analyzer

nRF5 SDK - nordicsemi.com. (n.d.). Retrieved August 21, 2021, from https://www.nordicsemi.com/Products/Development-software/nrf5-sdk

Oura. (n.d.). *Charging Your Ring & Best Practices – Oura Help*. Retrieved September 23, 2021, from https://support.ouraring.com/hc/en-us/articles/360046218953

Palliative Performance Scale (PPSv2) version 2. (2001). https://victoriahospice.org/reprint-and-use-information/

Patient Safety improvement.nhs.uk/resources/patient-safety-alerts. (n.d.). Retrieved

August 10, 2021, from http://rc.rcjournal.com/content/52/6/727
Pepe, M. S. (n.d.). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series :28. https://doi.org/10.1198/jasa.2005.s19
Polar Inc. (2021). *Polar Activity Tracking: Background, Feedback and Validity*. www.polar.com
Practice Team Information Data Quality Model. (2012). *March*, 1–13. Practice Team Information Data Quality Model (isdscotland.org)
Quer, G., Radin, J. M., Gadaleta, M., Baca-Motes, K., Ariniello, L., Ramos, E., Kheterpal, V., Topol, E. J., & Steinhubl, S. R. (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine*, *27*(1), 73–77. https://doi.org/10.1038/s41591-020-1123-x
Radin, J. M., Quer, G., Ramos, E., Baca-Motes, K., Gadaleta, M., Topol, E. J., & Steinhubl, S. R. (2021). Assessment of Prolonged Physiological and Behavioral Changes Associated with COVID-19 Infection. In *JAMA Network Open* (Vol. 4, Issue 7, p. e2115959). American Medical Association. https://doi.org/10.1001/jamanetworkopen.2021.15959
Radin, J. M., Wineinger, N. E., Topol, E. J., & Steinhubl, S. R. (2020). Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *The Lancet Digital Health*, *2*(2), e85–e93. https://doi.org/10.1016/S2589-7500(19)30222-5
Reddy, S., Rogers, W., Makinen, V.-P., Coiera, E., Brown, P., Wenzel, M., Weicken, E., Ansari, S., Mathur, P., Casey, A., & Kelly, B. (2021). Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health & Care Informatics*, *28*(1), e100444. https://doi.org/10.1136/BMJHCI-2021-100444
Rockwood, K., Andrew, M., & Mitnitski, A. (2007). A Comparison of Two Approaches to Measuring Frailty in Elderly People. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *62*(7), 738–743. https://doi.org/10.1093/gerona/62.7.738
Rockwood, Kenneth, Song, X., MacKnight, C., Bergman, H., Hogan, D. B., McDowell, I., & Mitnitski, A. (2005). A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal*, *173*(5), 489–495. https://doi.org/10.1503/cmaj.050051
Rodríguez-Mañas, L., Féart, C., Mann, G., Viña, J., Chatterji, S., Chodzko-Zajko, W., Gonzalez-Colaço Harmand, M., Bergman, H., Carcaillon, L., Nicholson, C., Scuteri, A., Sinclair, A., Pelaez, M., Van Der Cammen, T., Beland, F., Bickenbach, J., Delamarche, P., Ferrucci, L., Fried, L. P., … Vega, E. (2013). Searching for an operational definition of frailty: A delphi method based consensus statement. the frailty operative definition-consensus conference project. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, *68*(1), 62–67. https://doi.org/10.1093/gerona/gls119
Rooney, L., Rimpiläinen Sanna, Morrison, C., & Lange, S. L. (2018). *Review of Emerging Trends in Digital Health and Care: A report by the Digital Health and Care Institute*. *November*. https://doi.org/10.17868/67860
Ryle, G. (1984). *The concept of mind*. Hutchinson London 1949 pp334.
Scheffer, M. (2010). Foreseeing tipping points. *Nature*, *467*(7314), 411–412. https://doi.org/10.1038/467411a
Sedgwick, P. (2011). *The ecological fallacy*. *343*(7817). https://doi.org/10.1136/bmj.d4670
Seshadri, D. R., Davies, E. V., Harlow, E. R., Hsu, J. J., Knighton, S. C., Walker, T. A.,

Voos, J. E., & Drummond, C. K. (2020). Wearable Sensors for COVID-19: A Call to Action to Harness Our Digital Infrastructure for Remote Patient Monitoring and Virtual Assessments. *Frontiers in Digital Health*, *2*, 8. https://doi.org/10.3389/fdgth.2020.00008

Sets of Numbers. (n.d.). Retrieved February 14, 2021, from https://www.math24.net/sets-numbers/

Seyhan, A. A. (2019). Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*, *4*(1), 1–19. https://doi.org/10.1186/s41231-019-0050-7

Shannon, C. E. (n.d.). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 623–656.

Sharma, V., Ali, I., Van Der Veer, S., Martin, G., Ainsworth, J., & Augustine, T. (2021). Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. In *BMJ Health and Care Informatics* (Vol. 28, Issue 1, p. 100253). BMJ Publishing Group. https://doi.org/10.1136/bmjhci-2020-100253

Sherlaw-Johnson, C., Crellin, N., Hutchings, R., Oung, C., Rolewicz, L., Elias, L., Kumpunen, S., Castle-Clarke, S., Cummins, L., Fulop, N., Doherty, P., Harrison, A., & Scobie, S. (n.d.). *Evaluation of the Care City Wave 2 test bed Final report*.

Shewhart, W. A. (2012). *Statistical Method from the Viewpoint of Quality Control*. Dover Publications. https://books.google.co.uk/books?id=5utN7ulkCCAC

Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*, *383*(25), 2477–2478. https://doi.org/10.1056/nejmc2029240

Smales, A. (n.d.). *Technology for Preventative Healthcare*.

Stolz, E., Hoogendijk, E. O., Mayerl, H., & Freidl, W. (2020). Frailty changes predict mortality in four longitudinal studies of aging. *The Journals of Gerontology: Series A*. https://doi.org/10.1093/gerona/glaa266

Sudre, C. H., Lee, K. A., Lochlainn, M. N., Varsavsky, T., Murray, B., Graham, M. S., Menni, C., Modat, M., Bowyer, R. C. E., Nguyen, L. H., Drew, D. A., Joshi, A. D., Ma, W., Guo, C. G., Lo, C. H., Ganesh, S., Buwe, A., Pujol, J. C., du Cadet, J. L., … Ourselin, S. (2020). Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom study app. In *medRxiv* (p. 2020.06.12.20129056). medRxiv. https://doi.org/10.1101/2020.06.12.20129056

Sundvall, E., Nyström, M., Karlsson, D., Eneling, M., Chen, R., & Örman, H. (2013). *Applying representational state transfer (REST) architecture to archetype-based electronic health record systems*. https://doi.org/10.1186/1472-6947-13-57

Susanto, M., Hubbard, R. E., & Gardiner, P. A. (2018). Validity and Responsiveness of the FRAIL Scale in Middle-Aged Women. *Journal of the American Medical Directors Association*, *19*(1), 65–69. https://doi.org/10.1016/J.JAMDA.2017.08.003

TechRadar. (n.d.). *Xiaomi Mi Smart Band 6 review*. Retrieved October 28, 2021, from https://www.techradar.com/uk/reviews/xiaomi-mi-smart-band-6

The Checklist Manifesto | Atul Gawande. (n.d.). Retrieved May 5, 2021, from http://atulgawande.com/book/the-checklist-manifesto/

Theou, O., Brothers, T. D., Mitnitski, A., & Rockwood, K. (2013). Operationalization of frailty using eight commonly used scales and comparison of their ability to predict all-cause mortality. *Journal of the American Geriatrics Society*, *61*(9), 1537–1551. https://doi.org/10.1111/jgs.12420

Thyssen, J. P., & Menné, T. (2010). Metal allergys-A review on exposures, penetration,

genetics, prevalence, and clinical implications. *Chemical Research in Toxicology*, *23*(2), 309–318. https://doi.org/10.1021/tx9002726

Tremblay, M. (2012). Letter to the editor: Standardized use of the terms "sedentary" and "sedentary behaviours." *Applied Physiology, Nutrition and Metabolism*, *37*(3), 540–542. https://doi.org/10.1139/H2012-024

Turner, G., & Clegg, A. (2014). Best practice guidelines for the management of frailty: a British Geriatrics Society, Age UK and RCGP report | Age and Ageing | Oxford Academic (oup.com) *43*(6).

Using the National early warning score (NEWS2) in primary care. London: RCGP 2020. (n.d.). National Early Warning Score (NEWS) 2 | RCP London

Valencell. (n.d.). *Optical Heart Rate Monitoring Technology: What You Need to Know*. Retrieved August 20, 2021, from Optical Heart Rate Monitoring Technology: What You Need to Know (valencell.com)

Valencell | Virtual Patent Marking. (n.d.). Retrieved February 17, 2020, from https://valencell.com/patents/

Vision. (n.d.). *The Edmonton Frail Scale*. Retrieved December 13, 2019, from http://help.visionhealth.co.uk/Vision_Consultation_Manager_Help_Centre/Content/ConMgr/69849.htm

Wallen, M., Gomersall, S., Keating, S., Wisløff, U., & Coombes, J. (2016). Accuracy of Heart Rate Watches. *PloS One*, *11*(5). https://doi.org/10.1371/JOURNAL.PONE.0154420

Wang, W., Shen, M., Tao, Y., Fairley, C. K., Zhong, Q., Li, Z., Chen, H., Ong, J. J., Zhang, D., Zhang, K., Xing, N., Guo, H., Qin, E., Guan, X., Yang, F., Zhang, S., Zhang, L., & He, K. (2021). Elevated glucose level leads to rapid COVID-19 progression and high fatality. *BMC Pulmonary Medicine*, *21*(1), 64. https://doi.org/10.1186/s12890-021-01413-w

What Are the Best Marathon Times? (n.d.). Retrieved April 30, 2021, from https://www.verywellfit.com/what-is-a-good-time-for-running-a-marathon-2911426

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, 1–10. https://doi.org/10.1145/2601248.2601268

Wu, C.-T., Lidsky, P. V, Xiao, Y., IT, L., R, C., T, N., S, J., J, D., RJ, B., CA, C., RL, W., AK, S., B, Z., H, C., Y, G., A, T., JV, N., GP, N., MS, M., … PK, J. (2021). SARS-CoV-2 infects human pancreatic β cells and elicits β cell impairment. *Cell Metabolism*, *33*(8), 1565-1576.e5. https://doi.org/10.1016/J.CMET.2021.05.013

Yen, P., Kelley, M., Lopetegui, M., Osado, A., Migliore, E., Chipps, E., & Buck, J. (2016). Understanding and Visualizing Multitasking and Task Switching Activities: A Time Motion Study to Capture Nursing Workflow. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2016*, 1264–1273. https://europepmc.org/articles/PMC5333222

Yuda, E., Furukawa, Y., Yoshida, Y., & Hayano, J. (2018). Characteristics of basal heart rate during daily life: Influences of age, gender, and seasons. *EAI Endorsed Transactions on Self-Adaptive Systems*, *0*(0), 155882. https://doi.org/10.4108/eai.5-2-2018.155882

Zhu, G., Li, J., Meng, Z., Yu, Y., Li, Y., Tang, X., Dong, Y., Sun, G., Zhou, R., Wang, H., Wang, K., & Huang, W. (2020). Learning from Large-Scale Wearable Device Data for Predicting Epidemics Trend of COVID-19. *Discrete Dynamics in Nature and Society*, *2020*, 1–8. https://doi.org/10.1155/2020/6152041

10 SUPPLEMENTARY APPENDICES

10.1 Expansions of sections

10.1.1 Interoperability: how a platform can include mHealth

OpenEHR systems use standard persistent data models for data, known as archetypes in a 2-level design; the archetypes form the Clinical Knowledge level, and depend on the underlying information structure level, known as the Reference Model as it is essentially fixed. Archetypes are like healthcare knowledge, so are persistent, whereas the data they model are updated in use.

A realistic national platform architecture is for multiple datastores that are vendor-neutral and may be federated. in Scotland, the NHS is developing a National Digital Platform (NDP) for health and social care, using EHRBase, and in July 2021 NHS Wales procured an openEHR clinical data repository using the Better® platform (DHC Wales, n.d.). In July 2021 Catalonia, having standardised on Snomed CT terminology, committed to its use within an openEHR architecture (*Catalonia OpenEHR*, n.d.) These follow Open Platform design to interoperate with legacy systems, and the key feature is that the clinical data repository models the data with standard openEHR archetypes. This data could include high-value derivatives of mHealth data, such as the last 30d of inactivity, sleep or RHR data, as shown by RKI and other platforms examined here.

When data becomes semantically interoperable, we can design for its reuse by multiple devices whenever required, as the transformation between formats is now eliminated or trivial, and can be done live.

However, to retain the trust of HCW, only the minimum quantity of data of maximum veracity should be shared. The infrastructure service should be continuously functional, automatically and silently to humans, while processing the data streaming from mHealth devices to derive numeric or class data that is of sufficiently high quality and veracity to present only when requested by human HCW.

This aligns with the vision for Care 4.0 by (Chute & French, n.d.)

10.1.2 Quantities in Quantitative data

Most healthcare types of data can be rendered as Rational Numbers (*Sets of Numbers*, n.d.) All are essentially processable by machine with resolution, scale and power far beyond human ability.

A1 Rational numbers (*Rational Numbers*, n.d.) are widely used to quantify a continuous observation from an analogue device, and are usually in decimal format (number base 10.) As all biological parameters are continuously variable, their values must support being rendered as rational numbers, and so are theoretically capable of indefinite resolution. Resolution is usually limited to a number of significant digits practical for the context.

The value range is entirely determined by Units of Measure, which must be specified e.g. a temperature value of 40 is meaningless until "degrees Celsius" are added as UoM, setting the context for its human range at 25 to 45. A value may be zero, or a change may be negative.

A2 Binary numbers are rational numbers denoted in binary (0,1) the format most used by digital technology. However, when the range of selected values of a system is limited to two, then such selection is also described as binary (IEEE, n.d.). Binary data are often known as digital data, but this is one of many meanings of "digital" per Section 2.7.1.

However, rational numbers also include discrete Whole Numbers of two further types, used for counting and ordering, that must be processed in specific logical ways:

B Cardinal are used for a simple tally of discrete items, by addition; they may also be compared with that for other sets e.g. this set is twice the size of that set. The value may be zero.

C Ordinal show the order, or rank, of discrete items in a sequence. This is often used to create meaning in a set of qualitative statements rendered in text. The value is never zero.

Quantitative data may thus be of 3 types: Continuous, Binary, or Whole (Cardinal or Ordinal) – as is described for Test results by table 1.2 p3 of Pepe, (n.d.) .

- Continuous data are rendered as Rational numbers, in formats from binary to hexadecimal (used internally in processors) or decimal (standard for human use.)

- When continuous data are transformed to binary classes, with values 0,1 known commonly as true/false, or present/absent, it becomes discrete or non-continuous.
- Ordinal data are rendered as a series of discrete serial symbols. The most common serial symbol is a whole number, even though this is also used to render a rational or a cardinal number, per Section 4.1.8.

  Other serial symbol sets are available, such as Roman numerals or alphabetic
- Cardinal data are rendered as one of an infinite set of whole numbers. It is not found in test results, but is found in Risk Predictor tools, per Section 4.2.4.

10.1.3 Heuristic use of data in healthcare

10.1.3.1 Identification of citizen, user, subject or patient as an individual person

This is the essential pre-condition for any processing of Personal Data: mis-identification is a category error that may be catastrophic, and is seen here as 100% inaccuracy.

It is further briefly discussed at Information Governance, but is out-of-scope.

10.1.3.2 Conditions pertaining to that person

This is a broad generic term with open-ended definition: "A Clinical condition, problem, diagnosis, or other event, situation, issue, or clinical concept that has risen to a level of concern" (*Condition - FHIR v4.0.1*, n.d.)." The definition is intended to include all types of data used for an individual's health and care regardless of origin. Since "a level of concern" is not externally referenced for veracity, it can be considered to be so by any party. The condition may be past or current i.e. did or does now exist, but may also be future; each of these statuses carry a probability to exist.

Probability is quantifiable, and a set of probable conditions forms a Differential Diagnosis, of which very few can be true; yet these are all essential in the interim processes of direct care, until all are refuted but those which are the definitive diagnoses of the conditions[52]. Diagnostic options are further qualified by time metadata.

[52] Where probability is high for future status, they may reach the level of validity known as a Prognosis.

The current probability of any condition being true is the logical expression of the key function of diagnosis in healthcare. Further, since quantitative data always has error that can also be expressed as a probability, these probabilities could be considered together as if metadata. When that is qualitative, this co-processing of the risks of different diagnoses with the certainty of the supporting data is difficult for even the most experienced clinician, whose expertise here is known as "clinical judgment." However, where these probabilities can be quantified, machine processing may apply.

10.1.3.3 Information Models for Conditions

These are three examples at high-level using their own sub-classes of "Condition":

10.1.3.3.1 Classic: Diagnoses, Diseases, Disorders and Syndromes

These four variants of condition have these sub-types often used as content headings in paper records:

Presenting complaint / Symptoms / Signs / Results / Impressions / *Plan*

10.1.3.3.2 Problem-Oriented Medical Records

is another paradigm promoted by Weed (*Lawrence Weed - Wikipedia*, n.d.) and discussed further by (Wright et al., 2012)

The definition of "Problem" is usually left open, as it lacks a definition than can be agreed across every context: this is analogous to the above informal and heuristic description (not a definition) of "condition."

Problems are each further described by the SOAP sub-types:

Subjective / Objective / Analysis / *Plan*

10.1.3.3.3 SBAR

In nursing, social care and management, another paradigm is known as SBAR (Academy, n.d.), with sub-types:

Situation / Background / Assessment / *Recommendation*

10.1.3.3.4 The Operational Paradigm

A common feature of these information models is that their outcome is a Plan or Recommendation which promotes Action.

This suggests a more heuristic paradigm for computable Health and Care records: the essential function or operational purpose is to *Create a Plan*.

A good plan should use data from trusted devices, with high confidence in their data due to its high certainty, to minimise further processing by HCW for sense or safety. Confidence in its certainty is maximised by use of data that is computable.
Note some data can only be processed by HCW, such as narrative text, and some only by computer e.g. data of high volume or variety, mathematical transforms, per Section 2.7.

10.1.4 Data mining

Classification is the statistical process to assign each member of a group as a member to at least one of several discrete sub-groups identified by text labels, by recognising patterns of similarity in their features.

This pattern recognition appears to be intuitive for the mind, but machines must learn, by supervised learning, to follow algorithms to either classify into classes, or make numerical predictions from continuous data (*Data Mining - Classification & Prediction - Tutorialspoint*, n.d.) A rule-based classifier can be built principally by two methods:

1. Rule Extraction extracts rules such as IF-THEN from a *known* decision tree, where one rule is created for each path from the root to the leaf node, each rule antecedent forms a splitting criterion using logical AND, and the leaf node holds the prediction for the class, forming the rule consequent.
2. Rule Induction: a Sequential Covering Algorithm extracts logical rules such as IF-THEN from training data, without generating a decision tree first. Each rule for a class covers many of the ordered lists, or tuples, of that class. When we have enough domain knowledge for binary yes/no decisions to be testable, rule extraction using an algorithmic decision tree can be our first option.

In this linear algorithm, a logical cascade uses only binary choices, at a scale memorisable by domain experts and small enough to be rendered as a static diagram:

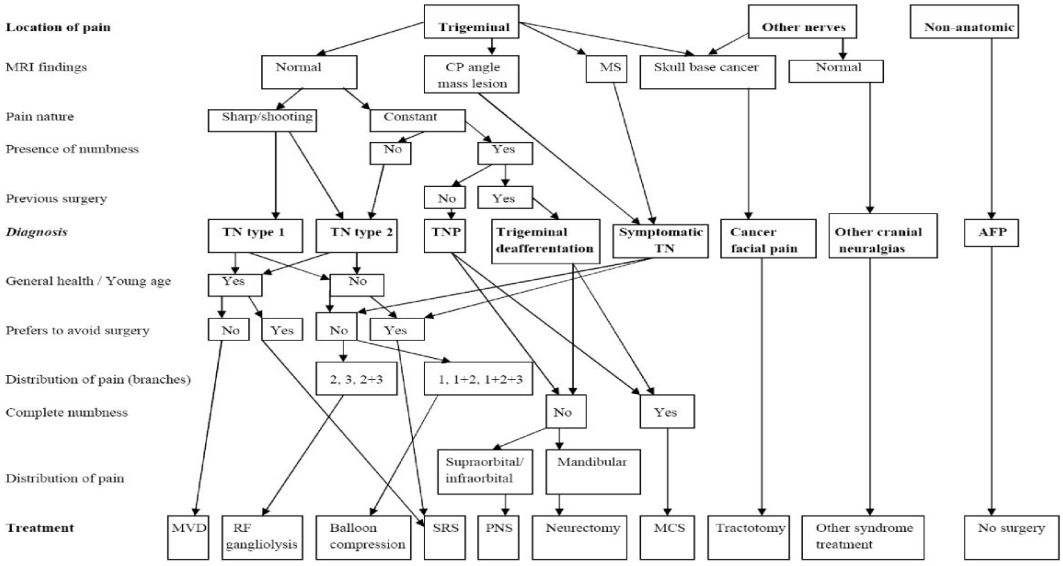


Figure 17: rule-extraction algorithm as a static diagram
Current algorithm for the surgical treatment of facial pain | Head & Face Medicine
https://doi.org/10.1186/1746-160X-3-30

However, rule induction can generate multiple classifications, by using continuous probabilities in a Bayesian network; and for machines there is no limit to the number of classes or of numeric predictions that can be so created, and simultaneously.

One of the first clinical examples of rule induction by computer algorithm was for surgeons when diagnosing appendicitis. While surgical decision-making clearly ultimately requires a binary output, to operate or not, even these early examples presented Bayesian probabilities as continuous variables and needed machine power to calculate probabilities of differential diagnoses (De Dombal et al., 1972)

The published versions do not attempt to print any (static) diagrams: only the machine can process it all, and only the expert surgeon can convert this continuous data to a final discrete binary decision.

10.1.5 Frailty and Phenotypic models

“Frailty” need not even include “elderly”: the international consensus conference (John E. Morley et al., 2013) explicitly excluded age from its definition: "a medical syndrome with multiple causes and contributors that is characterized by diminished strength, endurance, and reduced physiologic function that increases an individual's vulnerability for developing increased dependency and/or death." Age is also omitted from an earlier consensus that “frailty is a multidimensional syndrome characterized by decreased reserve and diminished resistance to stressors.” (Rodríguez-Mañas et al., 2013). As qualitative accuracy or veracity is elusive, it is unsurprising that quantifying it in Risk Predictor tools is even more so.

Thus the SPARRA Index (*Health and Social Community Care | SPARRA | SPARRA Online | Health Topics | ISD Scotland*, n.d.) models the hard outcome of risk of hospital admission, but this is only partly due to frailty, being due more directly to specific diagnoses, medications and events such as falls, which significant events are likely to be already recorded.

Phenotypic models are small datasets in which data points are specifically constructed for the purpose and must be assessed and entered by the clinician. For example, of the original 5 item FRAIL index per (Fried et al., 2001) only weight is commonly already available, but the others must be calculated and added. This adds to the HCW workload.

A new FRAIL scale, also based on 5 components: fatigue, resistance (inability to climb stairs), ambulation (inability to walk a certain distance), illnesses, and loss of weight is a simple questionnaire consisting of only 5 self-reported yes/no items. (J E Morley et al., 2012)

The Edmonton Frail Scale is an 11-item scale of PT type, which has been integrated into the VisionHealth Primary Care system (Vision, n.d.), and has also been implemented as an openEHR archetype (*Clinical Knowledge Manager*, n.d.)

A modified FRAIL index was shown to be valid in a massive study by (Hanlon et al., 2018) of nearly 0.5m subjects, despite requiring all subjects from age 37 upwards to attend over four years at special clinics to supply data, and its validity in routine Primary Care settings was confirmed by (Susanto et al., 2018)

Analysis of the comparative review of 8 such frailty indices listing all of their data as at 2013 by (Theou et al., 2013) shows that the PT type of RP is intrinsically lossy of information: for example, in the Groningen scale, a binary value for “Fitness” is derived from self-report of frequency of four modest activities, although these activities are not individually recorded, losing resolution.

For the Groningen Mental Health concept of “Emptiness,” also a binary value, the burden for HCW to accurately use the score seems to exceed that for direct care. This “lossiness” is a major source of the low-quality data from a tool, that would be known in metrology as “low resolution.”

10.1.6 Framework Development

Another evaluation framework, for implementation of AI systems, has just been published by (Reddy et al., 2021) (eleven authors.) This framework “Translational Evaluation of Healthcare AI (TEHAI)” is said to address the more translational components - functional, utility and ethical. Reporting components were identified and independently reviewed by an international panel of eight experts to form consensus. Three further translational features - safety, translational value and generalisability - are claimed to support practical application. –

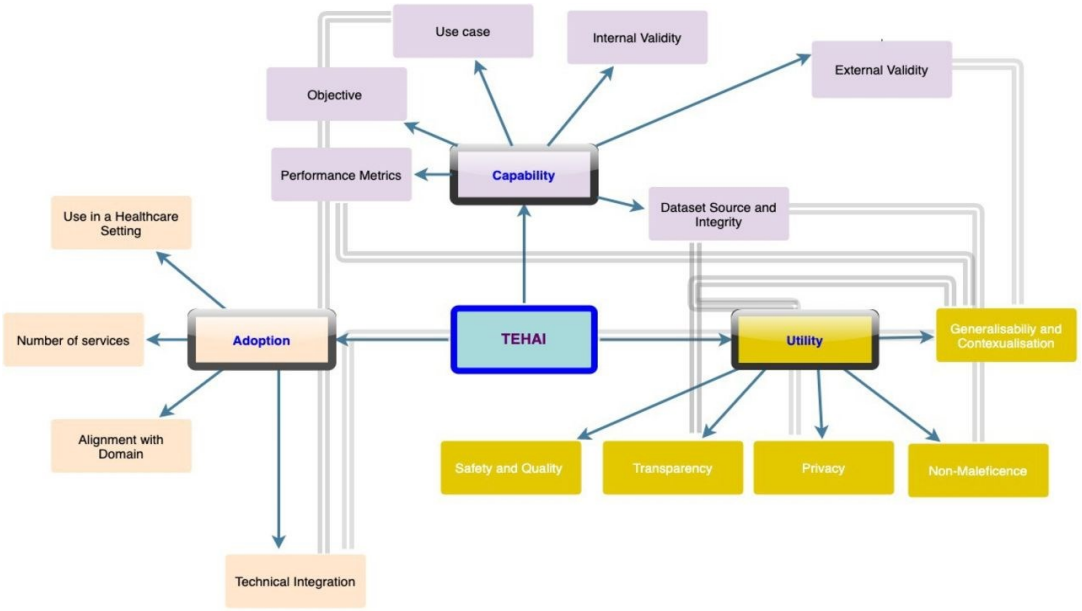


Figure 18: TEHAI framework
Reddy et al doi: 10.1136/bmjhci-2021-100444

The major domains are Adoption, Utility and Capability.

Mapping was attempted to other frameworks.

However, as the subject is AI only, it was considered unhelpful here.

10.2 Other mHealth technologies

10.2.1 Electrode contacts

on a chest belt, can collect accurate data for HR (pulse rate) and RRi (heart rate variability (HRV)) monitoring, and are now widely used by athletes to monitor training regimes. However a chest belt is impractical for everyday use.

A single contact on rear of a smartwatch is now used by some premium brands for HR, and display of a section of ECG waveform is also in development. However, continuous contact wearing eventually causes dermatitis due to nickel/chromium sensitivity in up to 17% of women and 3% of men (Thyssen & Menné, 2010).

10.2.2 GPS

tracks location, using 1 of 3 world-wide satellite systems for determining absolute location to a consumer accuracy of 3m. The timing metadata of location represents mobility. It was considered further at Section 3.4.10.1.

It is unreliable indoors due to attenuation of satellite signal.

10.2.3 Beacons

are fixed radio devices defining an area commonly around a person's home or indoors as a fixed reference location. They are used to monitor short-range mobility inside a large building where GPS is not reliable, and is more cost-effective for multiple subjects in one location e.g. a Care Home.

Power consumption is up to 10x lower than for WiFi or GPS location tracking.

Both types of data can be useful in predicting illness, including COVID-19.

10.2.4 Bio-Impedance Monitor

Dual-RF bio-impedance scales e.g. by Tanita™ measures several body composition components, of which muscle mass and hydration are reliable predictors of frailty.

10.2.5 "Digital" spring gauge

e.g. by Camry™ meter, a mechanical device measuring grip strength requires the user to operate for 3 readings/session, so not suitable for automation.

Data may be shown by "digital" display i.e. 7-segment LED, and is manually entered into an app – so this is not really a digital device.

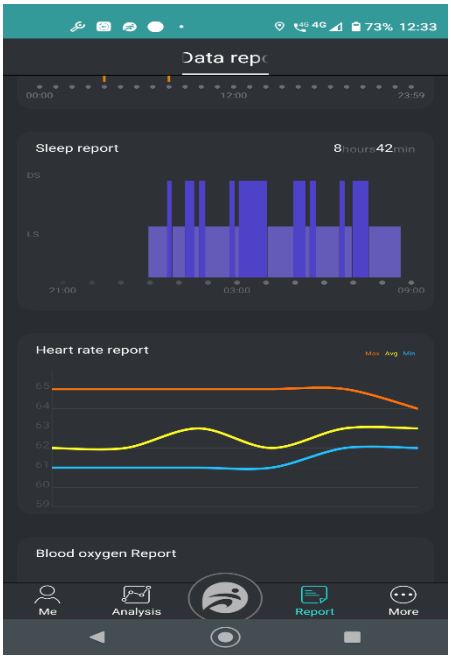
Both grip strength and bio-impedance data are reliable proxy measures for the onset of frailty (Smales, n.d.).

10.3 Device display details

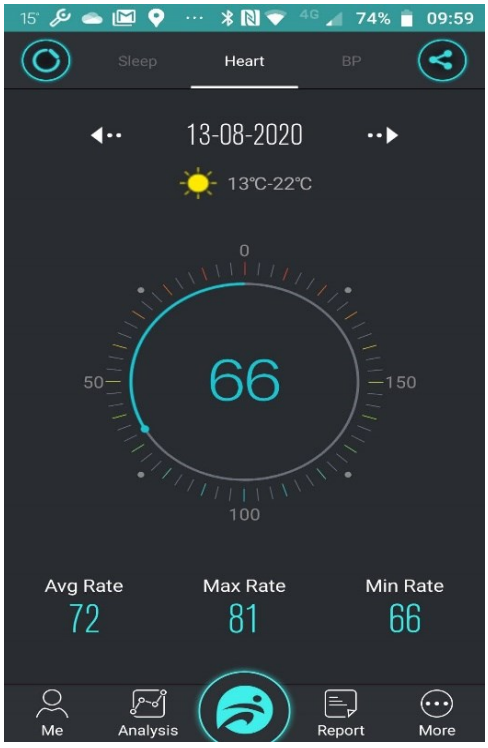
All screenshots are by author on personal Android smartphone Motorola GPro.

[L8 Multi UI Display ... Smart Watch]

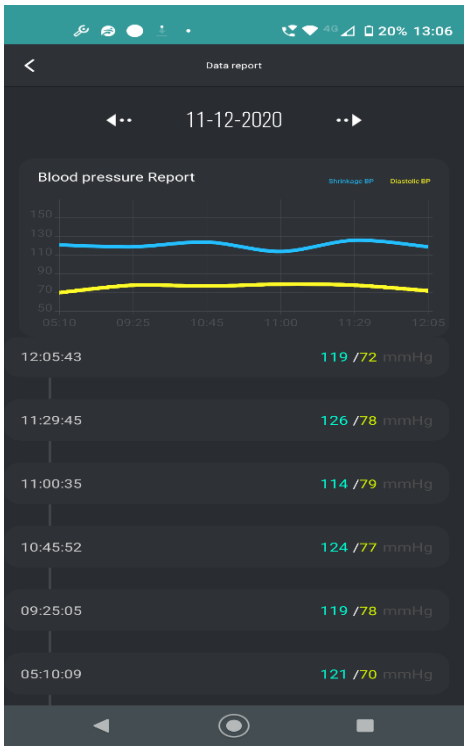
L8 smartwatch + Fundo® software



Graphic displays for Sleep and Heart Rate



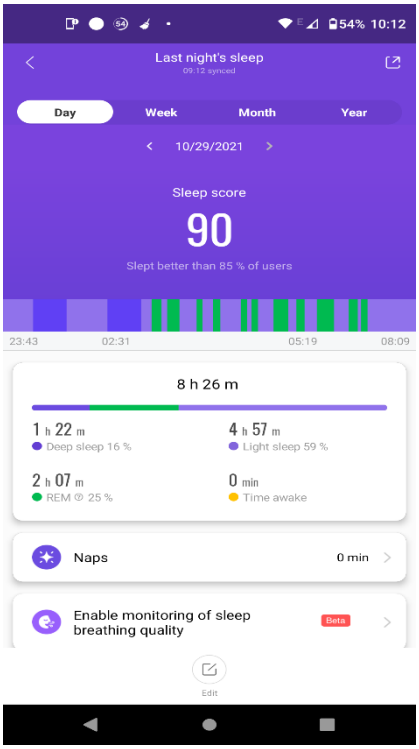
Numeric Heart rate “Min Rate” = RHR



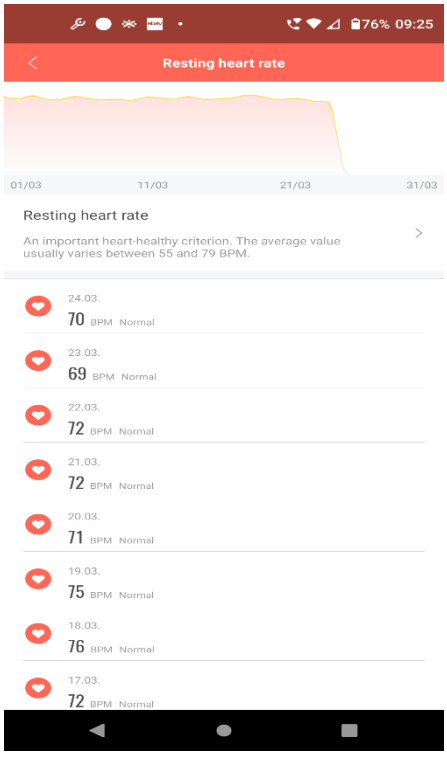
Mixed display Blood Pressure

Huami / Xiaomi Mi-Band 6 + MiFit software

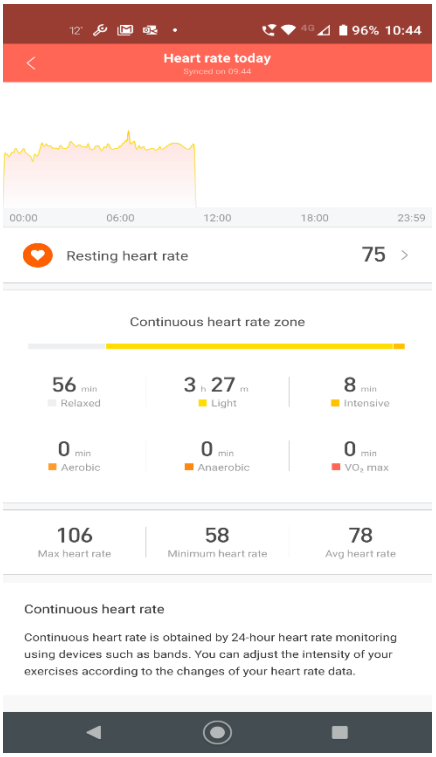
[Image per marketing on www.ebay.com]



Mixed display Sleep (high quality)



Numeric serial display RHR (stable)



Mixed display Heart Rate

10.4 Method details

10.4.1 Software used

Microsoft Office 365®

Mendeley® v1.19.8

Xmind Pro® v8u9

Better® Archetype Designer

10.4.2 Wordcount

Word® has automatic counts per section, and Tables are by individual count to exclude

	Oct 2021	Sept 2022
Intro	695	547
Background	5706	5653
Lit Reviews	15297	15275
Methods	2039	2030
Discussion	3195	3268
Evaluations or critical exam	4423	4492
Future work	283	598
Summary, conclusion	733	722
Tables x 9	-870	-870
	31501	31715

For MRes, “the normal range is 20-30,000 words” and “The word count does not include acknowledgements, the abstract, tables and diagrams, headings to tables and diagrams, the bibliography or list of references, appendices”
(per Research Development Framework)

10.4.3 Mindmaps

These are the first three Research Questions mapped in Xmind, mid-project.

Each green cloud is a connector to one of seven more mindmaps.

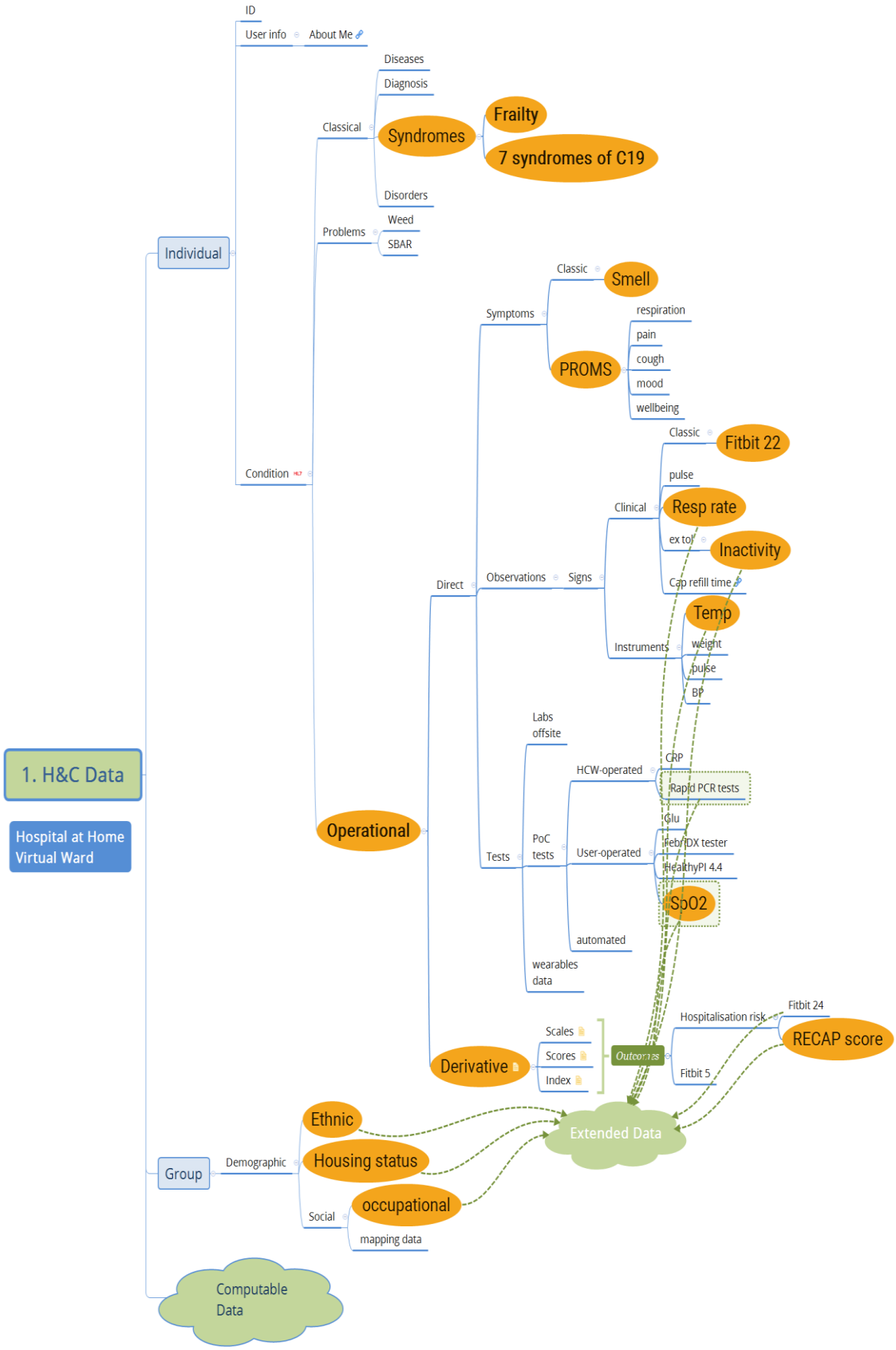


Figure 19: RQ1 data sources

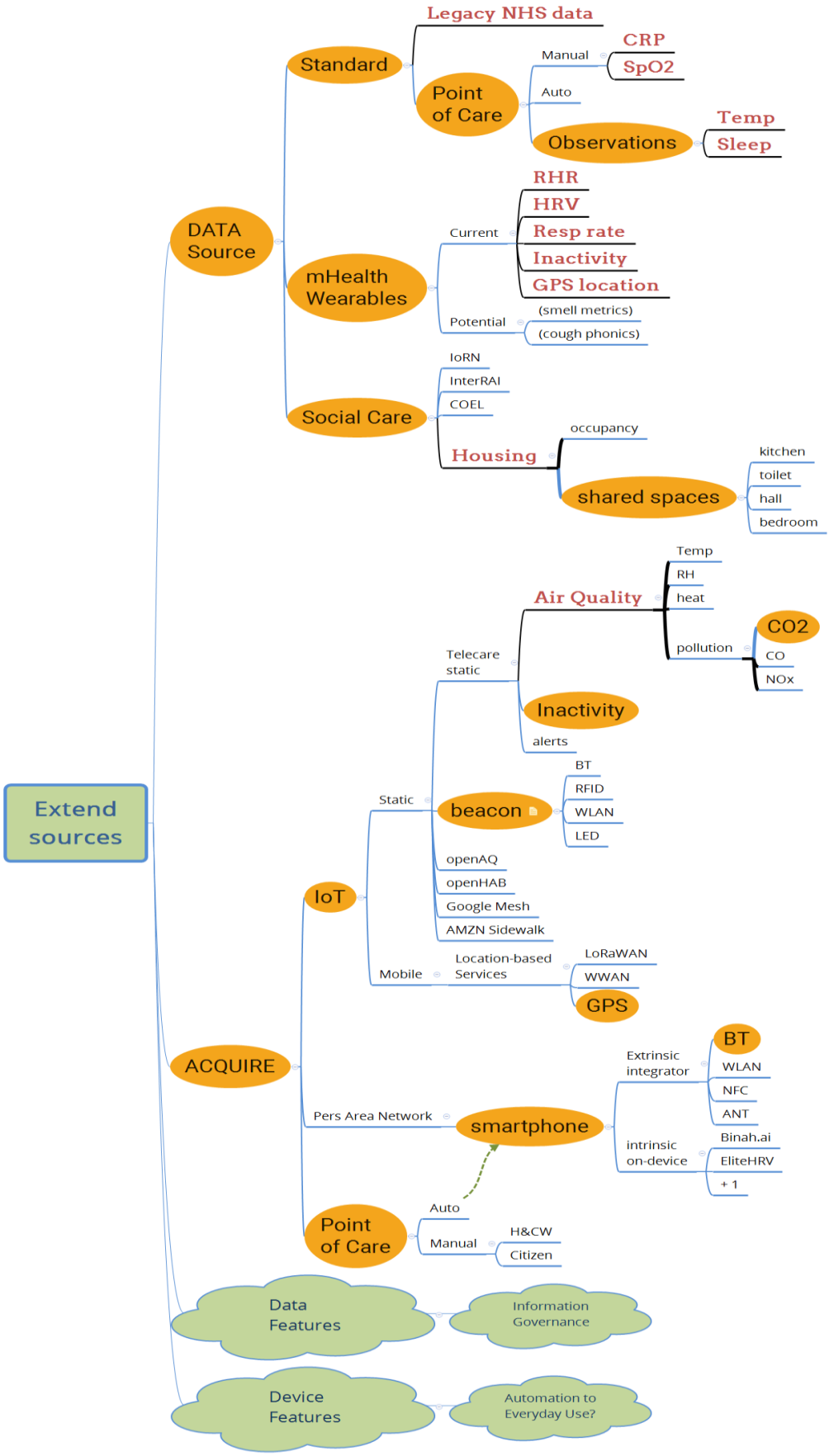


Figure 20: RQ2 extend data

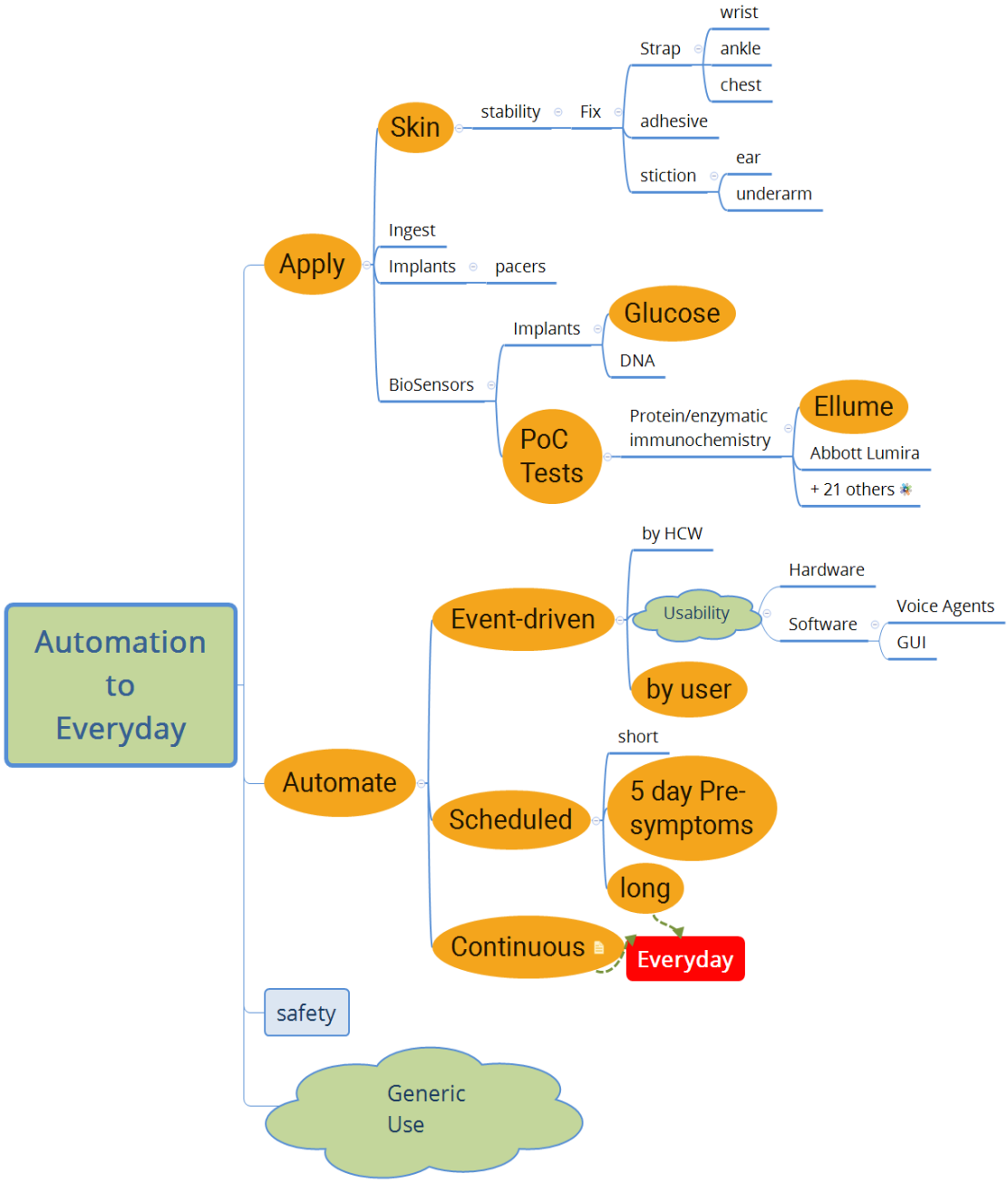


Figure 21: RQ3 automate data

10.4.4 Courses attended for education and training

May	SCIMP Open Platform Workshop	published documentation
	FHIR workshop	London
Oct	BCS Primary Healthcare Open Platform	Stratford
Nov	DH&C Conference	Glasgow
Oct	DHI Masters student induction day	DHI Glasgow
	BCS Primary Healthcare 2019 Conf	Stratford
	Mendeley ENU	Webinar
	ENU Research methodology induction day	Edinburgh
Nov	openEHR day: openEHR sites UK, Slovenia	London
Feb 2020	Armed, DH&C Conf Webinar	Edinburgh
Feb	Digital Product Forge: web-based products	Edinburgh
	All meetings from here are virtual, showing sponsor	
Mar	Armed Digital Telecare Conference	Scotgov Dig Health
Jun	DHI Masters Research day	DHI
Aug	Computable Knowledge sources	FCI
	Clinical Safety training update	NHS Digital
	CareTech 3 seminar	Scottish Care
Sept	NHS Scotland Silver Teams	Scot Dig Health Network
	NHS IT QA and Fault management	Dig Health
Sept	How can citizens better manage their data?	SDHN / Chute
	NHS Scotland Silver Teams	SDHN
	NHS IT QA and Fault management	Dig Health Networks /FCI
	Health and Social Care data exchange England by GDE sites	
Oct	Adrian Smales	mtg
	https://www.gov.scot/policies/digital/digital-identity-scotland/	
Nov	Towards a Data Strategy	SDHN
	Personal Health Records	DHI / SDHN
Dec	Silver Digital GP Review x2	SDHN
Jan 2021	COVID Oximetry @home: an overview for primary care	BMJ
	Testing for C19 in asymptomatic ppl	BMJ
Feb	Scot Joint GP IT Group	BMA
	HK Frailty archetypes etc.	mtg
Mar	openEHR Forum post #1: analogue nature of NEWS2 data	
	Covid home management update	UK Hospital-at-Home Soc
	HK PPSv2 archetype populate with AD	mtg x 4
	Health apps for long COVID self-management	FCI
	Digital Health ReWired	Dig Health Networks
	Remote Health Pathways update	ScotGov Dig Health
May	Conference Presentation	Scot Natl User GP
Jun	Webinar discussions x 2	Evergreen Health
	Webinar	FCI
Sept	Scot Joint GP IT Group	BMA
	Digital PROMS	FCI
Oct	2021 Conf Stratford	BCS Primary Healthcare
	HIGHMed openEHR symposium Berlin	openEHR Foundation

Presentations were made for:

24.6.20	DHI Research presentations day
26.5.21	SNUG Annual Conference