AN AGING FOCUSED UNOBTRUSIVE AND PRIVACY-PRESERVING DIGITAL BEHAVIOROME

Narayan Schütz¹, Samuel E.J. Knobel¹, Angela Botros¹, Michael Single¹, Bruno Pais², Valérie Santschi², Daniel Gatica-Perez^{3,4}, Philipp Bulushek⁵, Prabitha Urwyler¹, Stephan M. Gerber¹, Réne M. Müri^{1,8}, Urs Mosimann¹, Hugo Saner^{1,6,7*}, Tobias Nef^{1,8*}

¹ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

²LaSource School of Nursing Sciences, HES-SO University of Applied Sciences and Arts Western Switzerland, Lausanne, Switzerland

³Idiap Research Institute, Martigny, Switzerland

⁴School of Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

⁵DomoSafety SA, Lausanne, Switzerland

⁶Institute of Social and Preventive Medicine, University of Bern, Switzerland

⁷Sechenov First Medical State University, Moscow, Russian Federation

⁸Department of Neurology, Inselspital, Bern, Switzerland

** Shared Last Authorship*

December 23, 2021

ABSTRACT

Digital measures are increasingly used as objective health measures in remote-monitoring settings. In addition to their use in purely clinical research, such as in clinical trials, one promising application area for sensor-derived digital measures is in technology-assisted ageing and ageing-related research. In this context, digital measures may be used to measure the risk of certain adverse events such as falls, and also to provide novel research insights into ageing and ageing-related conditions, like cognitive impairment. While major emphasis has been placed on deriving one or more digital measures from wearable devices, a more holistic approach inspired by systems biology that leverages large, non-exhaustive sets of digital measures may prove highly beneficial. Such an approach would be useful if combined with modern big data approaches like machine learning. As such, extensive sets of digital measures, which may be referred to as digital behavioromes, could help characterise new phenotypes in deep phenotyping efforts. These measures could also assist in the discovery of novel digital biomarkers or in the creation of digital clinical outcome assessments. While clinical research into digital measures focuses primarily on measures derived from wearable devices, proven technology used for long-term remote monitoring of older adults is generally contactless, unobtrusive, and privacy-preserving. In this context, we introduce and describe a digital behaviorome: a large, non-exhaustive set of digital measures based entirely on contactless, unobtrusive, and privacy-preserving sensor technologies. We also demonstrate how such a behaviorome can be used to build digital clinical outcome assessments that are relevant to ageing and derived from machine learning. These outcomes included fall risk, frailty, mild cognitive impairment, and late-life depression. With the exception of late-life depression, all digital outcome assessments demonstrated a promising ability ($\text{ROC AUC} \geq 0.7$) to discriminate between positive and negative health outcomes, often in the range of comparable work with wearable devices. Finally, we highlight the possibility of using these digital behaviorome-based outcome assessments to discover novel potential digital biomarkers for each outcome. Here, we found reasonable contributors but also some potentially interesting new candidates regarding fall risk and mild cognitive impairment.

Keywords Digital Measures · Ageing · Digital Biomarkers · Remote-Monitoring · Digital Behaviorome · Older Adults

1 Introduction

Modern sensor technologies are increasingly used to obtain health information in remote-monitoring settings [1]. While no commonly accepted nomenclature exists when referring to sensor-derived health information, we try to adhere to recommendations set out by Goldsack et al. [2, 3]. As such, we use the term ‘digital measure’ to broadly describe any sensor-derived measurement of health. Applications of digital measures in general can be diverse, ranging from providing novel endpoints in clinical trials [4] over healthcare monitoring [5] to wellness monitoring with fitness trackers. Following Goldsack et al., meanwhile, we use the term ‘digital biomarkers’ to refer to individual digital measures that apply an easy-to-grasp concept and are strongly correlated to a known measurement of disease [2, 3]. An example of this is the number of steps per day, as measured by a digital device, which has been shown to be strongly associated with cardiovascular disease and mortality [6]. We use the term ‘digital clinical outcome assessment’ (COA) for digital measures that describe how a patient feels, functions, or survives [2, 3]. This latter type of digital measure may utilise different sensor modalities and other digital measures without necessitating the measurement of an underlying physical concept. For instance, digital measures of gait and finger-tapping might be used to build a machine learning-derived digital Parkinson’s disease severity measure [7].

While much recent interest in this area has centred on the use of digital measures in clinical trials, another traditional field of application has been in technology-assisted ageing and ageing-related research. With the rapid ageing of many countries’ populations, age-related social and economic challenges are starting to gain prominence [8, 9]. In this area, modern information technology may help solve, or at least alleviate, some of the associated negative effects of societal ageing, such as steadily increasing healthcare costs [10, 11]. Moreover, digital measures could provide information about an individual’s functional status and health changes, thereby facilitating earlier and more personalised interventions that may eventually allow older adults to remain independent for longer and increase their quality of life [5]. Although limited to the efforts of just a few research groups, research in this direction has been ongoing for more than a decade, amounting to numerous successful studies around the world [12, 5, 13, 14, 15, 16, 17, 18, 19, 20]. For instance, Rantz et al. show how sensor technologies linked to early alert systems led to better health outcomes among older adults [5]. Furthermore, Hayes et al. demonstrate that variations in sensor-derived gait speed and physical activity differ significantly between older adults with mild cognitive impairment (MCI) and healthy older adults [18]. Meanwhile, in our own research, we recently found that nightly movements in bed may be a promising digital measure; with further validation, these movements could potentially constitute a digital biomarker for monitoring signs of early health deteriorations in older adults [16]. In another line of research, Piau et al. have demonstrated how in-home walking speed may be a potential digital biomarker for fall risk among older adults [21]. Furthermore, Austin et al. report that digital measures were used successfully to assess loneliness in a study with 16 older adults [22]. In addition to directly improving health outcomes by means of early detection of health issues, digital measures in ageing research may also enable novel insights into ageing and ageing-related conditions [23].

Sensor technology used in populations of older adults is commonly unobtrusive or even contactless, with minimal to no participant interaction and the goal to minimise privacy intrusion. These sensor technologies contrast with the predominant use of wearable devices employed to derive digital measures in clinical research. There are multiple reasons for this: (1) older adults tend to be more wary of novel technologies [24]; (2) since monitoring durations are potentially very long or even unlimited, it is ideal if there is no interaction with the system, as there is evidence of wear-time-dependent compliance issues [25]; (3) there is a certain stigma attached to the use of wearable devices, whereby many older adults tend to fear being seen as frail if they wear a device – even if it is just an alarm clock [24]; (4) for users with potential memory issues, wearing and maintaining devices may not be feasible. Due to these factors, most long-term research using sensor technology with older adults has focused on unobtrusive and cost-effective contactless devices [18, 13, 5, 19, 23, 26, 14]. Sensors successfully used in ageing-related, real-world, long-term monitoring projects include passive infrared (PIR) motion sensors that capture an individual’s activity in a given room [5, 18, 13, 19, 23, 27], contact door sensors that can signal when a person leaves or enters the home [15, 26, 23], pressure sensors on or under a mattress that capture sleep measures [16, 23, 5], and electronic pillboxes to track medication adherence [28], along with potentially more obtrusive depth-sensing cameras that track silhouettes to detect falls and gait parameters [5]. A primary downside of using this kind of contactless, unobtrusive technologies, as opposed to wearables, is the additional environmental influences such sensor placement, apartment layouts or material properties. This may lead to more noise measurements as well as potentially worse inter-subject reliability.

Currently, digital measures for ageing and other areas are commonly evaluated individually or by combining several specific measures based on concepts of interest. While this approach is entirely reasonable, it may limit the potential of digital measures. We therefore hypothesise that a more holistic approach — inspired by systems biology and applying

digital measures — may be highly promising. This involves using larger sets of digital measures, potentially in the hundreds to thousands, which may be particularly helpful in exploratory research and may enable the creation of strong digital COAs by leveraging large-scale machine learning approaches. This approach is analogous to more classical biological settings, where measurements can assess individual blood tests or genes (for instance, by means of single-nucleotide polymorphisms) as well as whole sets, such as metabolomes or genomes, to identify new phenotypes of health and disease. Since science is primarily informed by what it can measure [29], the introduction of a new layer of information such as this could potentially lead to significant new research insights [30]. In the context of unobtrusive and contactless technologies, this could also help to counteract some of the downsides of those technologies by basing algorithms on a larger set of measures, that may bring about some redundancies.

In the context of digital measures, an extensive set may be called a digital behaviorome, in which the basic measurement unit is a digital measure – as opposed to, for instance, a gene or metabolite in a genome or metabolome, respectively. In this train of thought, Rashidisabet et al. define a digital behaviorome as ‘a subset of a person’s behavior measurable by connected devices via a systems biology approach’ [31]. However, it remains to be seen whether this definition is overly broad, as it would encompass measures for parameters such as vital signs, which are only indirectly related to behaviour. Furthermore, it should be taken into account that, unlike with biological ‘omes’, a complete set is likely never attainable, as there may be an infinite number of possible digital measures to be extracted, such that a digital behaviorome is a relatively non-exhaustive set of digital measures. Two notable real-world examples of using extensive sets of digital measures are studies by Cook et al. [27] and Chen et al. [32], which demonstrated the feasibility of using a behaviorome based on wearable and contactless sensors to predict multiple clinical scores (in the former) and mild cognitive impairment (in the latter).

Building on their work, we introduce a detailed, unobtrusive digital behaviorome, constituting a non-exhaustive set of 1268 existing and novel digital measures. We emphasise the importance of basing these measures on sensor technology that is evidently accepted by older adults, in addition to being affordable and feasible in real-world deployments. Thus, all of these measures are based on a small set of unobtrusive, contactless, and cost-effective sensors that, as shown by Baettie et al., scale well to large ageing-related remote-monitoring projects [23]; thus, these sensors should be compatible with most long-term monitoring projects in community-dwelling older adults.

In this study, we make three major contributions: (1) We introduce and describe the digital measures that comprise the proposed unobtrusive behaviorome (including detailed descriptions as part of accessible supplementary online material); (2) We showcase the potential for using this behaviorome to create ageing-relevant COAs for fall risk, frailty, late-life depression, and MCI; (3) We highlight the possibility of discovering novel digital biomarkers on the basis of the introduced COAs.

2 Results

2.1 An Unobtrusive Digital Behaviorome

We introduce a set of 94 hypothesis-driven base measures, from which we further derive a total of 1268 digital measures using aggregation and frequency analysis. All measures are obtainable through unobtrusive and privacy-preserving contactless sensors that do not require any participant interaction. Of these 1268 digital measures, 223 were extracted by means of PIR sensors in essential rooms (the entrance, bathroom, living room, bedroom, and kitchen) and magnetic door sensors on the refrigerator and entrance door. An additional 1046 measures were extracted on the basis of sleep data from a quasi-piezoelectric bed sensor placed under the mattress. Detailed descriptions and derivations, as well as associated hypotheses, are provided in the supplementary material, together with a high-level summary table¹ of all presented digital measures. Furthermore, an extensive online version with interactive visualisations, along with additional data including measure distributions and correlations with various ageing-relevant health indicators and outcomes, is available online at GitHub² and serves as an online supplementary to this article. An example of averaged behavioromes is shown in Figure 1.

2.2 Machine Learning based Digital Clinical Outcome Assessments

Here, we demonstrate how the introduced digital behaviorome could be useful for for ageing and ageing related research. To this end, we created machine learning-derived digital clinical outcome assessments, aimed at automatically classifying ageing-relevant health outcomes based on the behaviorome. We created five datasets, one based on each clinical

¹<https://narayanschuetz.github.io/digital-behaviorome/SummaryTable.xlsx>

²<https://narayanschuetz.github.io/digital-behaviorome>

Figure 1: Depicts an example of z-normalised, averaged behavioromes of participants with mild cognitive impairment (MCI) (based on a Montreal Cognitive Assessment screening < 23 points). Digital measures > 0 (in blue) indicate above-average values for that group, while < 0 (red) indicates below-average values. Many digital measures visually differ in both examples. It should be noted that this is a down-scaled visualisation, as not all measures would fit in the figure. For the complete and interactive version, see the *supplementary online version*.

Table 1: Results for the creation of machine learning-derived digital COAs based on clinical assessments for fall risk (TUG & POMA), frailty (EFS), late-life depression (GDS), and mild cognitive impairment (MoCA).

TUG: Timed Up and Go Test

POMA: Performance-Oriented Mobility Assessment

EFS: Edmonton Frail Scale

MoCA: Montreal Cognitive Assessment

Dataset	ROC AUC ($\pm$ 95 CI)			PrAUC ($\pm$ 95 CI)		
	Demographics	Behaviorome	Behaviorome + Demographics	Demographics	Behaviorome	Behaviorome + Demographics
TUG	0.662 $\pm$ 0.041	0.786 $\pm$ 0.037	0.783 $\pm$ 0.036	0.681 $\pm$ 0.035	0.816 $\pm$ 0.033	0.805 $\pm$ 0.034
POMA	0.777 $\pm$ 0.028	0.782 $\pm$ 0.035	0.805 $\pm$ 0.027	0.615 $\pm$ 0.046	0.650 $\pm$ 0.047	0.678 $\pm$ 0.044
EFS	0.564 $\pm$ 0.034	0.704 $\pm$ 0.039	0.698 $\pm$ 0.038	0.523 $\pm$ 0.041	0.625 $\pm$ 0.045	0.603 $\pm$ 0.044
GDS	0.602 $\pm$ 0.046	0.620 $\pm$ 0.048	0.616 $\pm$ 0.048	0.397 $\pm$ 0.051	0.477 $\pm$ 0.054	0.458 $\pm$ 0.053
MoCA	0.430 $\pm$ 0.038	0.780 $\pm$ 0.039	0.757 $\pm$ 0.038	0.653 $\pm$ 0.027	0.863 $\pm$ 0.028	0.839 $\pm$ 0.028

assessment, including fall risk, frailty, late-life depression, and MCI. This analysis is based on remote-monitoring data from two observational longitudinal pilot studies in Switzerland, where independently living, community-dwelling older adults were equipped with pervasive computing systems and monitored over the course of a year, while simultaneously being subject to regular visits and clinical assessments. The results on predicting ageing-relevant positive and negative health outcomes are summarised in Table 1. The differences between using the behaviorome alone versus using the behaviorome in addition to demographics were minimal and, judging by overlapping 95% CIs, non-significant. The highest discriminative power, in terms of ROC AUC (area under the receiver-operating characteristic curve), was achieved with the Tinetti Performance-Oriented Mobility Assessment (POMA)-based fall-risk dataset when including both demographic and behaviorome information (ROC AUC = 0.805). Notably, however, demographic information alone was sufficient to produce good performance (ROC AUC = 0.777) in this particular case. Performances on the fall-risk related Timed Up and Go Test (TUG), the MCI-related Montreal Cognitive Assessment (MoCA), and the frailty-related Edmonton Frail Scale (EFS) datasets were also relatively high, with ROC AUC values of 0.786, 0.780, and 0.704, respectively, when using only the behaviorome. The worst performance was achieved with the dataset based on the Geriatric Depression Scale (GDS) (ROC AUC = 0.620), when using only the behaviorome. Here, the difference between using only demographics versus using the behaviorome was also minimal, with a slight but non-significant advantage in favour of the behaviorome-only scenario. Overall, though, the addition of the behaviorome resulted, in all cases, in higher ROC AUC and PrAUC (Area Under the Precision-Recall Curve) values than those obtained when using only demographic information. These differences were statistically significant in all but the POMA and GDS datasets (which were based on rather conservative, non-overlapping CI intervals). The largest differences were found with the MoCA and TUG datasets, for which the objectives were to identify participants with an indication of mild cognitive impairment or increased fall risk, respectively.

2.3 Evaluation of Individual Digital Measure Importance

The importance of individual digital measures with respect to the COAs were evaluated by means of SHapley Additive exPlanations (SHAP) values [33, 34]. In Figure 2, we present the most important digital measure, based on global SHAP values, across all 100 simulations of each single dataset – corresponding to the different COAs. A more detailed table, highlighting the top 10 highest-ranked measures based on global SHAP values, is available in the Appendix C.III. Furthermore, in Figure 3, we display *beeswarm* plots of SHAP values for the individual COAs. These show how the nine most important digital measures – as well as the sum of all other measures combined – influence the log odds ratio of having a negative health outcome on the various datasets. Values shown in Figure 3 mostly align with the global SHAP importance rankings, although there are minor differences. Since the global SHAP rankings are based on 100 iterations (and thus 100 models), and since the importances shown in Figure 3 are based on a single model, we generally place greater emphasis on the global SHAP values. Nonetheless, the *beeswarm* plots are still useful, as they provide insights into the direction of effects.

3 Discussion

In this work, we introduce a digital behaviorome based on unobtrusive and privacy-preserving contactless sensor technology that is uniquely suitable for use in the population of community-dwelling older adults. We then demonstrate

Figure 2: Displays descriptions and density plots of the most important digital measure for each outcome. Across all density plots, blue indicates a positive/neutral outcome, while orange indicates a negative outcome.

Figure 3: Shows *beeswarm* plots of the 9 most important digital measures based on SHAP values on all outcome datasets: TUG (Timed Up and Go) & POMA (Performance Oriented Mobility Assessment = fall risk, GDS (Geriatric Depression Scale) = late-life depression, EFS (Edmonton Frail Scale) = frailty, MoCA (Montreal Cognitive Assessment) = mild cognitive impairment. Finally, the contributions of the sum of the remaining measures is displayed. Digital measures are ordered according to their importance, from top to bottom. The x-axis represents log odds, where values above zero indicate relevance for a negative outcome. Colouring further shows the direction of this association, where blue indicates lower values of a given measure and red indicates higher respective measure values. Detailed explanations of the individual measure names are given in the supplementary material or on the supplementary website. Note that these plots are based on models trained on the whole respective dataset and are therefore slightly different from the global importances shown in Table C.III, which are based on 100 simulation iterations.

the feasibility of utilising the introduced behaviorome to create machine learning-based digital COAs for common ageing-relevant outcomes, including fall risk, frailty, late-life depression, and mild cognitive impairment. Finally, we determine which individual digital measures were most relevant for a given COA.

While the idea of a sensor-derived behaviorome as a non-exhaustive set of digital measures characterising a person's behaviour is not entirely new, existing work remains scarce and is not primarily focused on technologies that can be reliably used in the population of community-dwelling older adults. The existing research often concerns the use of wearable devices over brief pre-defined time intervals, which may not be optimal for long-term monitoring (ranging from several years to potentially decades) with older adults; this technology contrasts somewhat with the type of sensor systems that have been successfully used for this purpose. Thus, our introduced digital behaviorome is not only the most detailed (to our knowledge) to date but also derivable by a set of simple, privacy-preserving, and low-cost contactless sensors. Consequently, it is compatible with technologies used in a majority of long-term remote-monitoring projects involving older adults, including the recently established Collaborative Aging Research Using Technology (CART) initiative, which seeks to make ageing-related digital health approaches more accessible to the broader research community [23]. Overall, then, we believe that the presented digital behaviorome has the potential to serve as a baseline set of measures that may be calculated over long time frames (ranging from years to potentially decades), but which could also be supplemented (potentially over shorter time periods) with digital measures based on more specific sensors, such as pillbox door sensors, wearables, or smartphones, depending on the specific needs, circumstances, and conditions.

Our results indicate that a digital behaviorome derived from simple and unobtrusive sensors could be suitable for creating useful ageing-relevant digital COAs. This is reflected by the good discriminative performances across all COAs except for late-life depression, resulting in ROC AUC values of ≥ 0.7 . Notably, these results are based on a very limited sample size, which makes it highly probable that this is a conservatively low estimate of what could be possible. In all cases, results based on the behaviorome led to higher ROC AUC and PrAUC values than those obtained from only using demographics information such as (see Table 1). These differences were significant in three (TUG, EFS, and MoCA) out of the five outcome datasets, indicating that the behaviorome captures information beyond just simple demographics. It is also notable that adding demographic information to the behaviorome did not result in significantly better performance across outcomes, which may indicate that this type of information is already implicitly present in the behaviorome. While we used COAs here as an example for feasibility purposes, COAs could be very useful for continuously assessing an older adult's health status with respect to specific outcomes and may allow for the implementation of early preventive interventions. For instance, if an older adult exhibits increased fall risk, it may be reasonable for them to see a fall prevention specialist. Furthermore, COAs such as those shown could aid research into complex ageing-relevant conditions such as cognitive impairment by providing objective information on how such impairment impacts a person's daily life.

Putting these results into perspective, passive sensor-based fall-risk assessments were shown to yield AUC values in the range of 0.65–0.89 [35]. It should be noted that these values were obtained using wearables by means of accelerometry and predominantly with very few digital measures of gait (and sometimes with accelerometer signal characteristics). Furthermore, none of the studies mentioned in this paper used long-term data, and some were performed under laboratory conditions far removed from real life. Our behaviorome-only results, with ROC AUC values of 0.786 (TUG dataset) and 0.805 (POMA dataset), are thus in line and satisfactory by comparison. Meanwhile, in terms of important digital measures, it is notable that not only physical activity and broadly gait-related measures (such as the number of room transitions) but also sleep and rhythmicity measures, such as activity in bed, bed-exit count, or activity fragmentation, were of major importance in discriminating between participants with high and low fall risk (see Figure 3). Although prospective data would be necessary to form clear conclusions, this may suggest that behavioural data beyond just gait and physical activity may be relevant for fall-risk assessments in older adults. Moreover, results from Piau et al. show that PIR array-based gait speed may help identify future fallers [21]. In this regard, the inclusion of gait-speed information that is more accurate yet still unobtrusive into a behaviorome would likely further increase performance on fall-risk assessments.

In terms of frailty, comparable studies report ROC AUC values between 0.72 and 0.86, based on wearable sensors [36, 37, 38]. These results were primarily obtained on the basis of gait and physical activity measures. Additionally, frailty definitions, study type, and participant characteristics differ quite significantly, so at best this gives a broad idea of what is possible. With a behaviorome-only ROC AUC value of 0.704, our results are on the lower end of this spectrum. However, given the types of sensors on which our behaviorome is based, this seems realistic. Indeed, some of the most important measures related to frailty were related to fridge usage, physical activity (room-transition counts),

and sleep duration (see Figure 3), all of which seem plausible as indicators.

For late-life depression, comparable studies are lacking. Several studies have demonstrated the utility of using wearable-based digital measures in assessing general depression [39, 40, 41]. Furthermore, one instance reported on the assessment of late-life depression using PIR-derived information on activities of daily living (ADL) [42]. However, it is unclear whether their methodology prevented data leakage, judging by the unusually high ROC AUC values ≥ 0.95 . Our own results, by contrast, show modest performance in assessing late-life depression, with a behaviorome-only ROC AUC value of 0.620. While this may be due to the low number of participants with a GDS score above 5 in our cohorts, it could also indicate some inherent difficulties in measuring this outcome. Many of the most important individual measures for late-life depression assessments are related to sleep duration (see Figure 3), which is known to be associated with depression. More interestingly, variations in fridge usage and behaviour complexity were relevant; however, due to their relatively low discriminative power, further interpretation may not be meaningful.

Regarding the distinction between healthy older adults and those with MCI, recent work has shown ROC AUC values of 0.62–0.80, based on comparable time intervals [43, 32]. These values were achieved with wearable devices but also when using additional modalities such as smartphone and computerised assessments. Our behaviorome-only result, with a ROC AUC value of 0.780, is thus aligned with similar research and shows that good discriminative performance may potentially be achieved through an entirely passive and unobtrusive set of sensors. Further supporting the plausibility of our results, a respectable body of literature shows how individual digital measures based on PIR and door sensors – such as variability in PIR array-based gait speed [44, 45, 20], ADL regularity [13], regularity in physical activity [18], sleep disturbances [46], and outing duration [47] – differed between older adults with MCI and healthy controls. Finally, with regards to MCI, highly important measures include those related to physical activity, such as the number of room-transitions or the total amount of PIR-based activity. Moreover, sleep-related measures such as sleep duration, activity in bed, and variation in in-bed activity were found to be important (see Figure 3). Regarding MCI, however, the most noteworthy finding is the inclusion of various sleep-related heart-rate measures – most importantly, the variation in nightly heart-rate dipping behaviour, where unusually high variation seems indicative of MCI (see Figure 2). This is especially interesting, as it has not been previously reported in connection with digital measures. However, it is known that heart-rate dipping is associated with cardiovascular disease [48] and that cardiovascular risk factors may be involved in cognitive decline [49]. As such, it could be beneficial to further investigate the relationship between nightly heart rate and mild cognitive impairment.

The presented behaviorome is a novel, thoroughly documented, and extensive set of digital measures obtainable with unobtrusive contactless sensor technologies. Sensor technologies appear to be the most successful in real-world use in ageing-related long-term remote-monitoring projects [14]. With growing interest in digital health solutions and remote monitoring, a digital behaviorome may be critical to the future of technology-assisted ageing and may provide value in both healthcare and research settings. As one potential use case, we demonstrate the feasibility of using the introduced behaviorome to create ageing-relevant COAs with real-world data from naturalistic samples of older adults. Our results often align with comparable results obtained from more obtrusive (but potentially more accurate) wearable sensors. Finally, we have found multiple individually interesting digital measures that may merit further investigation. These measures show how a behaviorome may also be of potential use in discovering novel ageing-relevant digital biomarkers. Another notable observation is that, across all outcomes, the sum of the remaining SHAP values – that is, all digital measures except for the 9 most important ones combined – was highly important in explaining model outcomes (see Figure 3). This supports the intuition that an extensive set of digital measures, such as the introduced behaviorome, is likely beneficial in terms of COA performance, as opposed to individual or several digital measures. It should also be mentioned that, by relying entirely on privacy-preserving, unobtrusive, and contactless technologies, large sets of digital measures may be derived without ethical concerns related to burdening subjects with unnecessary sensing modalities. The best practices laid out by Goldsack et al., for instance, discourage efforts towards sensor-symptoms mapping, which is to some degree what a digital behaviorome-based approach is doing [3]. However, since in this case sensor technologies respect privacy and do not add any additional burden, there are scarcely any downsides, as would occur with adding additional wearables or tasks requiring interaction.

Future research should emphasise further analytical and clinical validation of the included digital measures. Moreover, analysing long-term temporal dynamics would be essential, as it would enable the identification of trajectories of certain digital measures or even whole groups of measures. Evaluating trajectories, in turn, could be extremely valuable, as was shown by Akl et al., for instance, who produced impressive results regarding MCI classification based on long-term trajectories of several individual digital measures [20]. In addition, future research could combine a digital behaviorome such as the one we utilised with traditional multi-omics data in a deep phenotyping effort. This may enable a wide

range of new research insights into ageing and ageing-related conditions, as it adds a new layer of objective information for characterising phenotypes of health and disease. As a final caveat, a digital behaviorome such as this should never be assumed to be complete or fixed. Future research will add new digital measures while old measures may be merged if they exhibit closely correlated behaviour. As such, in the immediate future, it may be of major value to add more accurate gait-related digital measures to the introduced behaviorome, as these have consistently been shown to be very important across many ageing-relevant health outcomes. Consequently, they are likely to add significant value.

There are four major limitations to this work. First, some of the introduced measures have not been validated beyond the scope of this research (the use or validation of a digital measure in other studies is indicated in the supplementary material). This implies that some digital measures may not quantify what we hypothesise, which could lead to inaccurate interpretations and conclusions. Second, our feasibility demonstration involved a relatively small number of participants. Therefore, these results should be interpreted with caution, although general tendencies are likely to be valid. Third, in the feasibility demonstration, we use a cross-sectional approach that fails to leverage the temporal trajectories of sequentially collected behaviorome slices. Indeed, we strongly believe that data collected over multiple years would be necessary to fully explore the utility of behaviorome-based approaches. Finally, one drawback of simple PIR sensors is that not all digital measures based on this technology can be calculated when more than one person is living in an apartment. This is not a problem for the sleep sensor, and certain PIR sensor-derived measures may be valid nonetheless, but it does warrant consideration.

4 Conclusion

In this work, we introduce an unobtrusive digital behaviorome for use in the older adult demographic. Overall, the behaviorome consists of 1268 digital measures derived from 94 hypothesis-driven base measures. All included digital measures are derived from a small set of privacy-preserving contactless sensors that have been successfully used in multiple ageing-related, long-term, remote-monitoring projects around the world. For each measure, we provide a detailed description, background information, and additional real-world data as supplementary online material. While use cases for the introduced behaviorome are diverse, we demonstrate a potential use case by creating multiple machine learning-derived digital COAs and evaluating their discriminative performance. To this end, we show how ageing-relevant outcomes such as fall risk, frailty, and mild cognitive impairment could be assessed through an extensive set of unobtrusively obtained digital measures. Our results with these measures not only show that information from the behaviorome significantly outperformed basic demographic information, but also that behaviorome-based COAs could match the performance reported in comparable studies employing more obtrusive wearable sensors, which may not be optimal for long-term monitoring among older adults. Finally, we highlight the possibility of using a digital behaviorome to potentially discover novel digital biomarkers, using a model explainability approach on the basis of models used to create the aforementioned COAs. The respective results show that the most important digital measures are reasonable, while also revealing two potentially relevant insights. First, while fall risk may be primarily associated with gait and physical activity, it also potentially exhibits strong associations with sleep-related measures. Second, unusually high variation in nocturnal heart-rate dipping was uniquely related to mild cognitive impairment.

5 Methods

5.1 Minimally-Obtrusive Digital Behaviorome

5.1.1 Sensors

The introduced digital measures are based on three sensor types that have been commonly used in remote-monitoring projects with older adults: PIR sensors, contact door sensors, and a sleep sensor. The PIR motion sensors were placed in the essential rooms of older adults' apartments. Essential rooms included the living room, bedroom, entrance, bathroom/toilet, and kitchen. The employed PIR sensors sampled with 0.5 Hz, and thus reported activity on or off states every 2 seconds. The reed switch-based door sensors, meanwhile, were placed at the entrance and the refrigerator door. Both PIR and door sensors were part of the DomoCare® (DomoSafety SA, Lausanne, Switzerland) home-monitoring system. Finally, for the sleep sensor, we used an EMFIT QS ferroelectret sensor (Emfit Ltd, Vaajakoski, Finland), which was fixed beneath the mattress at approximately chest height. A summary of these three devices, as well as their respective source data streams, is given in Table 2.

Table 2: A summary of the contactless pervasive computing devices used, including the respective data streams they provide.

Sensor	Device	Location	Device Type	Data Stream	Sampling Rate	Channels
PIR Motion	DomoCare	Wall/ Ceiling	Contactless	Activity	0.5 Hz	Entrance Toilet Bedroom Livingroom Kitchen Others
Magnetic Door Reed Switches	DomoCare	Door	Contactless	Open/ Close State	Events	Entrance Fridge
Bed Sensor	EMFIT QS	Beneath Mattress	Contactless	Heart Rate	0.25 Hz	N/A
				Respiration Rate	0.25 Hz	N/A
				Activity	0.25 Hz	N/A
				Toss & Turns	Events	N/A
				Sleep Phases	0.25 Hz	Awake REM Deep Light
				Raw Low Frequency Band	50 Hz	N/A
				Raw High Frequency Band	100 Hz	N/A
				Bed Exits	Events	N/A
				Summary Statistics	Daily	N/A

5.1.2 Creation of Digital Measures

The creation of the digital measures was mostly hypothesis-driven or based on measures from the existing literature. The majority of measures were calculated on a daily or nightly basis (for instance, daily total activity, daily outing duration, or average heart rate during a night). Subsequently, we calculated derivatives of those measures by means of descriptive statistics over non-overlapping bi-weekly segments. While bi-weekly segments may be somewhat arbitrary, two weeks is a sufficient period to capture variation in behaviour that often follows daily or weekly cycles while still being short enough to capture temporally limited behaviours. Additionally, it serves to increase the number of data points (data augmentation) and facilitates the process of working with sensor recordings of various lengths or with data gaps. For cases of certain behavioural or rhythmicity measures, such as Cosinor regression-based measures, raw data from the whole bi-weekly segment was used directly. To avoid the inclusion of measures with insufficient data, we set a minimum number of 10 days for which raw source data was available throughout a given bi-weekly segment; otherwise, the measure was set as missing. The criteria for including a day's worth of data for each sensor type are explained in detail in the supplementary material 1.

For all daily or sub-daily base measures, derivatives based on summary statistics were calculated over the bi-weekly segments. Summary statistics include various quantiles, denoted as qn (e.g. $q10$), the interquartile range (iqr), the mean, median ($=q50$), coefficient of variation ($coefvar$), and robust measures of kurtosis and skewness ($kr3$ and $sk3$, respectively), following the naming convention proposed by Kim and White [50]. Figure 4 summarises this workflow visually. Eventually, this left 1268 dimensional vectors (one per bi-weekly segment). Of those, 223 dimensions are related to PIR and door sensors, while the remaining 1045 are based on sleep-sensor data. Detailed information regarding the exact calculation of each measure, as well as individual distributions across our cohort, can be found at *online*.

5.1.3 Example Visualisation

We provide an example of averaged behavioromes with respect to MCI. These visualisations were created by first averaging the behavioromes on a per-participant basis, followed by z-normalisation. After that, the behavioromes were split into the positive and negative outcome groups and averaged once more. Finally, heatmaps for both conditions were created. As a result, the values of individual digital measures > 0 indicate above-average values, while those < 0 indicate below-average values.

Figure 4: Shows a broad summary of how digital measures were calculated, starting with raw sensor data from PIR sensors, door sensors, and bed sensors. Raw data streams were first segmented into non-overlapping bi-weekly segments. Then, for each bi-weekly segment, digital measures were calculated. If, for a given measure, less than 10 days of data were present, the measure was encoded as missing, which eventually left 1268 dimensional vectors – one per bi-weekly segment.

Table 3: Participant Characteristics

Characteristic	Cohort 1 (n=24)	Cohort 2 (n=21)	Pooled (n=45)	Differences (p-value)
Age (years)	88±7	86±7	87±7	0.50
Sex female (%)	79	52	67	0.06
TUG	14.0±10.0	15.3±4.8	14.5±8.2	0.58
POMA	21.4±7.4	24.1±3.3	22.5±6.1	0.13
EFS	4.6±3.6	5.0±1.7	4.7±2.9	0.65
GDS	2.3±2.3	4.2±3.2	3.1±2.9	0.04
MoCA	20.5±5.1	20.6±3.7	20.5±4.7	0.95

5.2 Machine Learning based Digital Outcome Assessments

5.2.1 Participants

To test the feasibility of using the previously described behaviorome to create ageing-relevant COAs, we used real-world remote-monitoring data from two cohorts of older Swiss adults (pooled age years=87±7; sex 67% [30/45] female). The original studies were both pilots designed to assess novel computing technologies for ageing-in-place scenarios in the German- and French-speaking cantons of Switzerland [51, 52]. They were conducted between 2017 and 2018 and monitored participants over one year with a set of pervasive computing devices and clinical assessments [51, 52]. The inclusion criteria between cohorts were similar in the sense that both aimed to recruit a natural sample of community-dwelling older adults (aged ≥ 70 years) who lived alone and without pets. On the other hand, the exclusion criteria between cohorts differed. For cohort 1, the only exclusion criterion was an unwillingness to comply with the study protocol. But, for cohort 2, the exclusion criteria were as follows: (1) severe cognitive impairment rendering the individual unable to follow study protocol (clock-drawing score ≥ 4); (2) skin problems such as irritations, itching, or serious redness; (3) undergoing dialysis; (4) unwillingness to comply with the study protocol; (5) an inability to understand the study aim; or (6) hospitalisation planned within a short period of time [51]. Both studies were conducted based on principles declared in the Declaration of Helsinki and approved by the Ethics Committees of the cantons of Bern and Vaud (KEK-ID: 2016-00406 and CER-VD ID: 2016-00762, respectively). All subjects signed and returned informed consent forms before participating in the study. Detailed participant characteristics and cohort differences are shown in the Appendix 3. The differences between cohorts were statistically examined on the basis of unpaired, two-sided, two-sample t-tests ($\alpha = 0.05$). In every analysis involving participant data, all participants with any available data (depending on sensor data and the availability of clinical assessments) were included; this also applies to participants that dropped out of the studies.

Table 4: Overview of clinical assessments used to evaluate the health status of study participants, including the measurement interval and cut-off points used to divide participants into groups with positive/neutral or negative health outcomes.

Assessment	Assessment Interval		Cut-Off Points	Outcome
	Cohort 1	Cohort 2		
TUG	6 weeks	6 weeks	≥ 12 seconds [58]	Fall-Risk
POMA	half year	half year	< 19 points [59]	Fall-Risk
EFS	begin/end	begin/end	> 5 points [60]	Frailty
GDS	begin/end	begin/end	≥ 5 points [61]	Depression
MoCA	begin/end	end	< 23 points [62]	MCI

Table 5: Dataset Characteristics

Dataset	Participants Total	Participants with Negative Outcome	Number Bi-Weekly Segments
TUG	28	14	277
POMA	28	10	277
EFS	28	10	277
GDS	28	7	277
MoCA	25	16	260

5.3 Clinical Assessments

Participants in both cohorts were subject to an overlapping set of standardised clinical assessments. These include the following six assessments: (1) the Timed Up and Go Test (TUG), which is often used in geriatrics to assess fall risk [53]; (2) the Tinetti Performance-Oriented Mobility Assessment (POMA), which, as with the TUG, also measures balance and gait characteristics that are often indicators for elevated fall risk among older adults [54]; (3) the Edmonton Frail Scale (EFS), a frequently used measure of frailty among older adults [55]; (4) the short version (15-item) of the Geriatric Depression Scale (GDS), a commonly used late-life depression screening tool [56]; (5) the Montreal Cognitive Assessment (MoCA), which measures cognitive function and is often used as a brief screening tool for the detection of MCI in older adults [57]. In each cohort, these assessments were planned to be conducted at least once during the one-year study duration. Detailed assessment intervals are summarised in Table 4.

5.4 Dataset Creation

To evaluate the potential for creating COAs that may help differentiate between positive and negative ageing-relevant health outcomes based on the proposed behaviorome, we categorise participants into one of the two categories for each clinical assessment. This was done on the basis of validated cut-offs for each assessment in 5.3. The respective cut-off values for the negative groups are shown in Table 4. Next, we calculated the digital measures as presented in 5.1.2 for all participants. For each assessment, we then combined the positive/negative labels with the bi-weekly segments of a given participant. If multiple records of the same clinical assessments were obtained throughout the study, we assigned the target label corresponding to the assessment closest in time. After this procedure, we obtained one dataset per assessment.

Note that measures derived from PIR and door sensors stem from one sensor system (meaning that technical failure usually affect both sensor types, except for instances where an individual sensor unit failed, which happened rarely), while sleep stems from another sensor; thus, for a bi-weekly segment to be valid, at least 30% of measures from both sensor systems must be valid. This led to a significant reduction in the number of bi-weekly segments, as a large number of sleep sensor data were missing due to technical issues, as has been discussed in prior work [16]. These two issues — lacking sensor data from both PIR/door and bed sensors and the unavailability of respective assessments — are responsible for the generally lower numbers of participants who could be included in this analysis (the exact numbers with regards to each assessment are given in Table 5). In Figure 5, we present the high-level flowchart of dataset creation.

5.5 Machine Learning Analysis

A small forenote aimed at a more technically oriented audience: what we call digital measures throughout this work can be seen as synonymous to the more abstract and general term "features". To evaluate COAs based on the behaviorome,

Figure 5: Highlights the workflow of creating datasets, subsequently used for the creation of digital clinical outcome assessments. First, digital measures were separately calculated for the PIR + door and bed sensors and segmented into non-overlapping, bi-weekly segments. After that, the measures from bi-weekly segments, where the percentage of missing digital measures from either sensor system was $< 30\%$, were combined. Next, the clinical assessments from each participant were matched with the respective bi-weekly digital measure vectors to combine 5 datasets — one for each assessment.

we largely followed the approach set out by Chen et al., albeit with some minor changes [32]. As such, we use the gradient boosting-based XGBoost algorithm [63] as a classifier, since it generally performs impressively on tabular data, tends to deal reasonably well with high-dimensional feature spaces (even in $p \gg n$ -type scenarios, as here), and can inherently deal with missing values, all of which means it is close to being the gold standard for this kind of application [32, 64, 39]. Furthermore, gradient boosting-based tree approaches tend to be more easily explainable than modern neural network approaches such as convolutional neural networks, while also retaining high accuracy, especially on tabular data structures [34]. To better account for stochasticity in participant selection, we further adapted the simulation strategy of Chen et al. [32], in which 70% of participants were repeatedly drawn from the entire participant pool to form a training set, while the remaining 30% were used as a test set (the splits are stratified for the respective clinical assessment labels). This procedure was repeated for 100 iterations. Note, that this way each new draw represents a shuffling of the dataset without introducing data leakage between training and test splits. Throughout each iteration, hyperparameters were first optimised within the training split by means of stratified 3-fold cross-validation coupled with random search (consisting of 50 search trials). For more detailed explanation of this strategy, we refer to the original article by Chen et al., where it is demonstrated in detail [32]. Eventually, for each iteration, we calculated the Area Under the Receiver-Operating Curve (ROC AUC) and the Area Under the Precision Recall Curve (PrAUC) on the test set, where multiple bi-weekly segments from a single participant were combined into one score by averaging their predictions (soft voting), as was done in [32]. Likewise, if multiple assessment results were available, they were first averaged, which should have also reduced some of the inherent noise; these results were then dichotomised on the basis of the previously introduced cut-off points (see Table 4) to yield a single label per participant. We removed three digital measures (Measure IDs: *iqr_entrance_door_tod_first*, *q50_entrance_door_tod_first*, and *q50_fridge_door_tod_middle*) from the full set for this portion of the analysis, as they were biased towards identifying one of the two cohorts (to account for further less-obvious biases in this regard, we included cohort information in the demographics). Note that the PrAUC is sensitive to label distribution, which means it only lends itself to comparisons within the same dataset. For each assessment, we ran three different scenarios, one with only demographic information (age, sex, and cohort membership) as baseline, one with only the behaviorome, and one with both the behaviorome and demographic information combined. Differences between these scenarios were deemed statistically significant if the 95% CIs of two conditions do not overlap. Model hyperparameter ranges were given in the supplementary material. We used the original Python (version 3.6) implementation of XGBoost (version 1.3.3). Model training was performed on UBELIX (<http://www.id.unibe.ch/hpc>), the HPC cluster at the University of Bern.

5.6 Identifying Digital Measures of Interest

To better understand the role of individual digital measures in machine learning-based COAs, we used SHapley Additive exPlanations (SHAP), a game-theoretic approach for explaining complex machine learning models. With this approach, exact solutions can be found in the case of tree-based models [33, 34]. SHAP values have been used fairly extensively in recent biomedical applications [65, 32, 66, 67]. For each of the assessments, we provide overall global SHAP values across all 100 simulations. That is, we give the mean absolute value of the SHAP values for a given digital measure m in a single simulation, summed up over all simulations

$$SHAP_m^{global} = \sum_{i=1}^{100} mean(|SHAP_m^i|).$$

While global SHAP values reveal the overall importance of a given digital measure, they do not say anything about the direction in which the digital measure influences a model. Therefore, we additionally calculate beeswarm plots of the SHAP values. These are based on a model trained over the entire respective dataset, with manually set hyperparameters (reported in the supplementary material). SHAP values were calculated using Python (version 3.6) with the *shap* package (version 0.39.0).

References

- [1] Caroline Marra, Jacqueline L Chen, Andrea Coravos, and Ariel D Stern. Quantifying the use of connected digital products in clinical research. *NPJ digital medicine*, 3(1):1–5, 2020.
- [2] Jennifer C Goldsack, Ariel V Dowling, David Samuelson, Bray Patrick-Lake, and Ieuan Clay. Evaluation, acceptance, and qualification of digital measures: from proof of concept to endpoint. *Digital Biomarkers*, 5(1):53–64, 2021.
- [3] Jennifer C Goldsack and Ieuan Clay. It takes a village: development of digital measures for computer scientists. In *Proceedings of the 2021 Workshop on Future of Digital Biomarkers*, pages 38–44, 2021.
- [4] Philip Boehme, Arne Hansen, Ronenn Roubenoff, Joseph Scheeren, Maximilian Herrmann, Thomas Mondritzki, Jan Ehlers, and Hubert Truebel. How soon will digital endpoints become a cornerstone for future drug development? *Drug discovery today*, 24(1):16–19, 2019.
- [5] Marilyn J Rantz, Marjorie Skubic, Mihail Popescu, Colleen Galambos, Richelle J Koopman, Gregory L Alexander, Lorraine J Phillips, Katy Musterman, Jessica Back, and Steven J Miller. A new paradigm of technology-enabled ‘vital signs’ for early detection of health change for older adults. *Gerontology*, 61(3):281–290, 2015.
- [6] Katherine S Hall, Eric T Hyde, David R Bassett, Susan A Carlson, Mercedes R Carnethon, Ulf Ekelund, Kelly R Evenson, Deborah A Galuska, William E Kraus, I-Min Lee, et al. Systematic review of the prospective association of daily step counts with risk of mortality, cardiovascular disease, and dysglycemia. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1):1–14, 2020.
- [7] Andong Zhan, Srihari Mohan, Christopher Tarolli, Ruth B Schneider, Jamie L Adams, Saloni Sharma, Molly J Elson, Kelsey L Spear, Alistair M Glidden, Max A Little, et al. Using smartphones and machine learning to quantify parkinson disease severity: the mobile parkinson disease score. *JAMA neurology*, 75(7):876–880, 2018.
- [8] James R Knickman and Emily K Snell. The 2030 problem: caring for aging baby boomers. *Health services research*, 37(4):849–884, 2002.
- [9] David E Bloom, David Canning, and Alyssa Lubet. Global population aging: Facts, challenges, solutions & perspectives. *Daedalus*, 144(2):80–92, 2015.
- [10] Sabine Koch. Healthy ageing supported by technology—a cross-disciplinary research challenge. *Informatics for Health and Social Care*, 35(3-4):81–91, 2010.
- [11] Hon Prof John R Beard and David E Bloom. Towards a comprehensive public health response to population ageing. *Lancet (London, England)*, 385(9968):658, 2015.
- [12] Bayard E Lyons, Daniel Austin, Adriana Seelye, Johanna Petersen, Jonathan Yeagers, Thomas Riley, Nicole Sharma, Nora Mattek, Katherine Wild, Hiroko Dodge, et al. Pervasive computing technologies to continuously assess alzheimer’s disease progression and intervention efficacy. *Frontiers in aging neuroscience*, 7:102, 2015.
- [13] Prabitha Urwyler, Reto Stucki, Luca Rampa, René Müri, Urs P Mosimann, and Tobias Nef. Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living. *Scientific reports*, 7(1):1–9, 2017.

- [14] Antoine Piau, Katherine Wild, Nora Mattek, and Jeffrey Kaye. Current state of digital biomarker technologies for real-life, home-based monitoring of cognitive function for mild cognitive impairment to mild alzheimer disease and implications for clinical care: systematic review. *Journal of medical Internet research*, 21(8):e12785, 2019.
- [15] Narayan Schütz, Hugo Saner, Beatrice Rudin, Angela Botros, Bruno Pais, Valérie Santschi, Philipp Buluschk, Daniel Gatica-Perez, Prabitha Urwyler, Laura Marchal-Crespo, et al. Validity of pervasive computing based continuous physical activity assessment in community-dwelling old and oldest-old. *Scientific reports*, 9(1):1–9, 2019.
- [16] Narayan Schütz, Hugo Saner, Angela Botros, Bruno Pais, Valérie Santschi, Philipp Buluschk, Daniel Gatica-Perez, Prabitha Urwyler, René M Müri, Tobias Nef, et al. Contactless sleep monitoring for early detection of health deteriorations in community-dwelling older adults: Exploratory study. *JMIR mHealth and uHealth*, 9(6):e24666, 2021.
- [17] Narayan Schutz, Angela Amira Botros, Sami Ben Hassen, Hugo Saner, Philipp Buluschk, Prabitha Urwyler, Bruno Pais, Valerie Santschi, Daniel Gatica-Perez, Rene M Muri, et al. A sensor-driven visit detection system in older adults homes: Towards digital late-life depression marker extraction. *IEEE Journal of Biomedical and Health Informatics*, 2021.
- [18] Tamara L Hayes, Francena Abendroth, Andre Adami, Misha Pavel, Tracy A Zitzelberger, and Jeffrey A Kaye. Unobtrusive assessment of activity patterns associated with mild cognitive impairment. *Alzheimer's & Dementia*, 4(6):395–405, 2008.
- [19] Jeffrey Kaye, Nora Mattek, Hiroko H Dodge, Ian Campbell, Tamara Hayes, Daniel Austin, William Hatt, Katherine Wild, Holly Jimison, and Michael Pavel. Unobtrusive measurement of daily computer use to detect mild cognitive impairment. *Alzheimer's & dementia*, 10(1):10–17, 2014.
- [20] Ahmad Akl, Babak Taati, and Alex Mihailidis. Autonomous unobtrusive detection of mild cognitive impairment in older adults. *IEEE transactions on biomedical engineering*, 62(5):1383–1394, 2015.
- [21] Antoine Piau, Nora Mattek, Rachel Crissey, Zachary Beattie, Hiroko Dodge, and Jeffrey Kaye. When will my patient fall? sensor-based in-home walking speed identifies future falls in older adults. *The Journals of Gerontology: Series A*, 75(5):968–973, 2020.
- [22] Johanna Austin, Hiroko H Dodge, Thomas Riley, Peter G Jacobs, Stephen Thielke, and Jeffrey Kaye. A smart-home system to unobtrusively and continuously assess loneliness in older adults. *IEEE journal of translational engineering in health and medicine*, 4:1–11, 2016.
- [23] Zachary Beattie, Lyndsey M Miller, Carlos Almirola, Wan-Tai M Au-Yeung, Hannah Bernard, Kevin E Cosgrove, Hiroko H Dodge, Charlene J Gamboa, Ona Golonka, Sarah Gothard, et al. The collaborative aging research using technology initiative: An open, sharable, technology-agnostic platform for the research community. *Digital Biomarkers*, 4(1):100–118, 2020.
- [24] Sebastiaan TM Peek, Eveline JM Wouters, Joost Van Hoof, Katrien G Luijkx, Hennie R Boeije, and Hubertus JM Vrijhoef. Factors influencing acceptance of technology for aging in place: a systematic review. *International journal of medical informatics*, 83(4):235–248, 2014.
- [25] Susan L Murphy. Review of physical activity measurement using accelerometers in older adults: considerations for research design and conduct. *Preventive medicine*, 48(2):108–114, 2009.
- [26] Jeffrey A Kaye, Shoshana A Maxwell, Nora Mattek, Tamara L Hayes, Hiroko Dodge, Misha Pavel, Holly B Jimison, Katherine Wild, Linda Boise, and Tracy A Zitzelberger. Intelligent systems for assessing aging changes: home-based, unobtrusive, and continuous assessment of aging. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66(suppl_1):i180–i190, 2011.
- [27] Diane J Cook and Maureen Schmitter-Edgecombe. Fusing ambient and mobile sensor features into a behaviorome for predicting clinical health scores. *IEEE Access*, 9:65033–65043, 2021.
- [28] Tamara L Hayes, John M Hunt, Andre Adami, and Jeffrey A Kaye. An electronic pillbox for continuous monitoring of medication adherence. In *2006 international conference of the IEEE engineering in medicine and biology society*, pages 6400–6403. IEEE, 2006.
- [29] Cathryn M Delude. Deep phenotyping: the details of disease. *Nature*, 527(7576):S14–S15, 2015.
- [30] E Dorsey, Larsson Omberg, Emma Waddell, Jamie L Adams, Roy Adams, Mohammad Rafayet Ali, Katherine Amodeo, Abigail Arky, Erika F Augustine, Karthik Dinesh, et al. Deep phenotyping of parkinson's disease. *Journal of Parkinson's Disease*, 10(3):855–873, 2020.
- [31] Homa Rashidisabet, Paul J Thomas, Olusola Ajilore, John Zulueta, Raeanne C Moore, and Alex Leow. A systems biology approach to the digital behaviorome. *Current Opinion in Systems Biology*, 2020.

- [32] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2145–2155, 2019.
- [33] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [34] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.
- [35] Ruopeng Sun and Jacob J Sosnoff. Novel sensing technology in fall risk assessment in older adults: a systematic review. *BMC geriatrics*, 18(1):1–10, 2018.
- [36] Danya Pradeep Kumar, Nima Toosizadeh, Jane Mohler, Hossein Ehsani, Cassidy Mannier, and Kaveh Laksari. Sensor-based characterization of daily walking: a new paradigm in pre-frailty/frailty assessment. *BMC geriatrics*, 20(1):1–11, 2020.
- [37] Michael Schwenk, Jane Mohler, Christopher Wendel, Mindy Fain, Ruth Taylor-Piliae, Bijan Najafi, et al. Wearable sensor-based in-home assessment of gait, balance, and physical activity for discrimination of frailty status: baseline results of the arizona frailty cohort study. *Gerontology*, 61(3):258–267, 2015.
- [38] Catherine Park, Ramkinker Mishra, Jonathan Golledge, and Bijan Najafi. Digital biomarkers of physical frailty and frailty phenotypes using sensor-based physical activity and machine learning. *Sensors*, 21(16):5289, 2021.
- [39] Mariko Makhmutova, Raghu Kainkaryam, Marta Ferreira, Jae Min, Martin Jaggi, and Ieuan Clay. Prediction of self-reported depression scores using person-generated health data from a virtual 1-year mental health observational study. In *Proceedings of the 2021 Workshop on Future of Digital Biomarkers*, pages 4–11, 2021.
- [40] Jonah Meyerhoff, Tony Liu, Konrad P Kording, Lyle H Ungar, Susan M Kaiser, Chris J Karr, David C Mohr, et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: Longitudinal cohort study. *Journal of Medical Internet Research*, 23(9):e22844, 2021.
- [41] Ipsit V Vahia and Daniel D Sewell. Late-life depression: a role for accelerometer technology in diagnosis and management. *American Journal of Psychiatry*, 173(8):763–768, 2016.
- [42] Jung-Yoon Kim, Na Liu, Hwee-Xian Tan, and Chao-Hsien Chu. Unobtrusive monitoring to detect depression for elderly with chronic illnesses. *IEEE Sensors Journal*, 17(17):5694–5704, 2017.
- [43] Jia Li, Yu Rong, Helen Meng, Zhihui Lu, Timothy Kwok, and Hong Cheng. Tatc: predicting alzheimer’s disease with actigraphy data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 509–518, 2018.
- [44] Tamara L Hayes, Stuart Hagler, Daniel Austin, Jeffrey Kaye, and Misha Pavel. Unobtrusive assessment of walking speed in the home using inexpensive pir sensors. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7248–7251. IEEE, 2009.
- [45] HH Dodge, NC Mattek, Daniel Austin, TL Hayes, and JA Kaye. In-home walking speeds and variability trajectories associated with mild cognitive impairment. *Neurology*, 78(24):1946–1952, 2012.
- [46] Tamara L Hayes, Thomas Riley, Nora Mattek, Misha Pavel, and Jeffrey A Kaye. Sleep habits in mild cognitive impairment. *Alzheimer disease and associated disorders*, 28(2):145, 2014.
- [47] Johanna Petersen, Daniel Austin, Nora Mattek, and Jeffrey Kaye. Time out-of-home and cognitive, physical, and emotional wellbeing of older adults: a longitudinal mixed effects model. *PloS one*, 10(10):e0139643, 2015.
- [48] Kazuo Eguchi, Satoshi Hoshide, Joji Ishikawa, Thomas G Pickering, Joseph E Schwartz, Kazuyuki Shimada, and Kazuomi Kario. Nocturnal non-dipping of heart rate predicts cardiovascular events in hypertensive patients. *Journal of hypertension*, 27(11):2265, 2009.
- [49] Alin Ciobica, Manuela Padurariu, Walther Bild, and Cristinel Stefanescu. Cardiovascular risk factors as potential markers for mild cognitive impairment and alzheimer’s disease. *Psychiatria Danubina*, 23(4.):340–346, 2011.
- [50] Tae-Hwan Kim and Halbert White. On more robust estimation of skewness and kurtosis. *Finance Research Letters*, 1(1):56–73, 2004.
- [51] Bruno Pais, Philipp Bulushek, Guillaume DuPasquier, Tobias Nef, Narayan Schütz, Hugo Saner, Daniel Gatica-Perez, and Valérie Santschi. Evaluation of 1-year in-home monitoring technology by home-dwelling older adults, family caregivers, and nurses. *Frontiers in public health*, 8:538, 2020.

- [52] Hugo Saner, Narayan Schütz, Angela Botros, Prabitha Urwyler, Philipp Bulushek, Guillaume Du Pasquier, and Tobias Nef. Potential of ambient sensor systems for early detection of health problems in older adults. *Frontiers in cardiovascular medicine*, 7:110, 2020.
- [53] Diane Podsiadlo and Sandra Richardson. The timed “up & go”: a test of basic functional mobility for frail elderly persons. *Journal of the American geriatrics Society*, 39(2):142–148, 1991.
- [54] Mary E Tinetti, T Franklin Williams, and Raymond Mayewski. Fall risk index for elderly patients based on number of chronic disabilities. *The American journal of medicine*, 80(3):429–434, 1986.
- [55] Darryl B Rolfson, Sumit R Majumdar, Ross T Tsuyuki, Adeel Tahir, and Kenneth Rockwood. Validity and reliability of the edmonton frail scale. *Age and ageing*, 35(5):526–529, 2006.
- [56] Emerson L Leshner and Jeffrey S Berryhill. Validation of the geriatric depression scale-short form among inpatients. *Journal of clinical psychology*, 50(2):256–260, 1994.
- [57] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.
- [58] Heike A Bischoff, Hannes B Stähelin, Andreas U Monsch, Maura D Iversen, Antje Weyh, Margot Von Dechend, Regula Akos, Martin Conzelmann, Walter Dick, and Robert Theiler. Identifying a cut-off point for normal mobility: a comparison of the timed ‘up and go’ test in community-dwelling and institutionalised elderly women. *Age and ageing*, 32(3):315–320, 2003.
- [59] Marjan J Faber, Ruud J Bosscher, and Piet CW van Wieringen. Clinimetric properties of the performance-oriented mobility assessment. *Physical therapy*, 86(7):944–954, 2006.
- [60] Simone Perna, Matthew D’Arcy Francis, Chiara Bologna, Francesca Moncaglieri, Antonella Riva, Paolo Morazzone, Pietro Allegrini, Antonio Isu, Beatrice Vigo, Fabio Guerriero, et al. Performance of edmonton frail scale on frailty assessment: its association with multi-dimensional geriatric conditions assessed with specific screening tools. *BMC geriatrics*, 17(1):1–8, 2017.
- [61] Filipi Leles da Costa Dias, Antônio Lúcio Teixeira, Henrique Cerqueira Guimarães, Maira Tonidandel Barbosa, Elisa de Paula França Resende, Rogério Gomes Beato, Karoline Carvalho Carmona, and Paulo Caramelli. Accuracy of the 15-item geriatric depression scale (gds-15) in a community-dwelling oldest-old sample: the pietà study. *Trends in psychiatry and psychotherapy*, 39:276–279, 2017.
- [62] Nicole Carson, Larry Leach, and Kelly J Murphy. A re-examination of montreal cognitive assessment (moca) cutoff scores. *International journal of geriatric psychiatry*, 33(2):379–388, 2018.
- [63] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [64] Diane Cook. Digitally mapping the human behaviorome. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 1–1. IEEE, 2020.
- [65] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.
- [66] Pål V Johnsen, Signe Riemer-Sørensen, Andrew Thomas DeWan, Megan E Cahill, and Mette Langaas. A new method for exploring gene–gene and gene–environment interactions in gwas with tree ensemble methods and shap values. *BMC bioinformatics*, 22(1):1–29, 2021.
- [67] Richard Li, Ashwin Shinde, An Liu, Scott Glaser, Yung Lyou, Bertram Yuh, Jeffrey Wong, and Arya Amini. Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. *JCO Clinical Cancer Informatics*, 4:637–646, 2020.

Appendix

1 Individual Day and Night Data Inclusion Criteria

1.1 PIR and Door Sensors

Sensors associated with the DomoCare system (PIR and door) were marked invalid based on whole days. A day is deemed invalid, if less than 300s of raw PIR activity was present for the whole apartment. We found that in general,

days with less activity were corresponding to technical issues or extremely long outings (such as related to holidays or hospitalizations). In Figure A.1a, this cut-off value is indicated. It is visible, how there is a second modality in the distribution appearing close to zero, distinct from the main modality, referring to the normal range of daily activity sums.

1.2 Bed Sensor

For night inclusion, we adopt the same procedure as used in previous work [16], displayed in A.1b. Nights without any data were marked as invalid beforehand.

Figure A.1: a) Shows raw PIR sensor system activity distribution for individual days, across all participants of both cohorts. The red vertical line denotes the cut-off, corresponding to the minimal activity in seconds for a day to be valid. Days below this value likely correspond to non-normal days, such as days with technical issues or days, where participants were away from home for almost the whole day. b) Depicts the inclusion criteria for nights to be valid based on sleep sensor data. If no data is present at all, the whole night is marked invalid. (the depicted flowchart is based on our original article in <https://mhealth.jmir.org/2021/6/e24666>).

2 Hyperparameter Search Space

In Table B.I we list the hyperparameter search space that was optimized with 3-fold cross-validation within each training split. Remaining hyperparameters were left at default values.

Table B.I: Hyperparameters of models used to produce beeswarm SHAP plots.

Hyperparameter	Value
subsample	[0.3, 0.5, 0.7]
max_depth	[1, 2, 3]
n_estimators	[100, 300, 600, 900]
gamma	[1, 2, 3]
eta	[0.01, 0.1, 0.5]

3 Beeswarm SHAP Values Plot Hyperparameters

In table C.II we list the XGBoost hyperparameters, used to train the models the SHAP *beeswarm* plots are based upon. Note, these hyperparameters are based on manual fine tuning and resulted in good performance across all outcome datasets. Hyperparameters not mentioned were left at default values.

Table C.II: Hyperparameters of models used to produce beeswarm SHAP plots.

Hyperparameter	Value
subsample	0.5
max_depth	2
n_estimators	900
eta	0.1

Table C.III: Displays digital measures sorted on the basis of highest global SHAP importances.

Digital Measure ID	Global SHAP Value
<i>TUG Dataset</i>	
q50_bed_act_rel_bandpower0.000000-0.002501_Hz	55.755
q50_fridge_door_tod_first	23.808
q50_bed_act_rel_bandpower0.025014-0.027515_Hz	21.877
q75_bed_duration	15.889
q25_bed_nocturia_count	10.104
q90_bed_bedexit_count	9.163
q50_bed_act_rel_bandpower0.030017-0.032518_Hz	8.994
q75_outing_duration	7.800
q90_ActivityFragmentationPercentage	7.800
q90_bed_duration_in_sleep	7.435
<i>POMA Dataset</i>	
q75_transition_count	37.061
q50_bed_act_rel_bandpower0.025014-0.027515_Hz	23.863
q75_fridge_usage_5min_epochs	21.461
q90_ActivityFragmentationPercentage	17.651
q90_transition_count	13.879
q90_fridge_usage	13.645
iqr_transition_count	12.639
lfhf	11.947
q75_fridge_usage	11.107
q50_bed_deep_sleep_hr	10.599
<i>EFS Dataset</i>	
q50_fridge_door_tod_first	55.915
entropy_fridge	49.930
q50_transition_count	19.039
q75_transition_count	18.332
q90_bed_duration	15.546
q10_transition_count	13.577
q75_bed_duration	13.066
q90_fridge_usage_5min_epochs	11.982
iqr_outing_duration	11.275
q50_fridge_door_tod_last	10.826
<i>GDS Dataset</i>	
q50_bed_duration_in_bed	44.054
q75_bed_duration_in_bed	25.409
q50_bed_duration_in_light	23.934
iqr_fridge_door_tod_last	22.334
iqr_fridge_door_tod_first	21.360
reconstruction_error	18.220
q75_bed_awake_sleep_hr	14.743
coefvar_fridge_usage_5min_epochs	14.063
q50_bed_duration	13.913
q75_bed_duration_in_light	12.611
<i>MoCA Dataset</i>	
iqr_bed_hr_q75_q25	31.798
q50_bed_act_rel_bandpower0.000000-0.002501_Hz	18.528
q50_transition_count	16.585
q90_bed_duration	14.927
q90_bed_activity_1min	14.562
q50_bed_hr_rel_bandpower0.007504-0.010006_Hz	11.331
q75_raw_activity_sum_seconds	9.295
q90_bed_bedexit_count	8.658
q50_bed_hr_rel_bandpower0.005003-0.007504_Hz	8.461
coefvar_bed_activity_1min	8.261