



Morgan, C. A. M., Heidarivincheh, F., Craddock, I., Mcconville, R., Perello Nieto, M., Tonkin, E. L., Masullo, A., Vafeas, A. T., Kim, M., Mcnaney, R., Tourte, G. J. L., & Whone, A. L. (2021). Data labelling in the wild: annotating free-living activities and Parkinson's disease symptoms. In *2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021: 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 471-474). Article 9431017 (2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2021). Institute of Electrical and Electronics Engineers (IEEE).  
<https://doi.org/10.1109/PerComWorkshops51409.2021.9431017>  
Peer reviewed version

Link to published version (if available):  
[10.1109/PerComWorkshops51409.2021.9431017](https://doi.org/10.1109/PerComWorkshops51409.2021.9431017)

[Link to publication record on the Bristol Research Portal](#)  
PDF-document

This is the accepted author manuscript (AAM). The final published version (version of record) is available online via IEEE at [10.1109/PerComWorkshops51409.2021.9431017](https://doi.org/10.1109/PerComWorkshops51409.2021.9431017). Please refer to any applicable terms of use of the publisher.

## University of Bristol – Bristol Research Portal

### General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:  
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/>

# Data labelling in the wild: annotating free-living activities and Parkinson’s disease symptoms.

Catherine Morgan<sup>\*¶</sup>, Farnoosh Heidarvinchek<sup>†</sup>, Ian Craddock<sup>†</sup>, Ryan McConville<sup>†</sup>, Miquel Perello Nieto<sup>†</sup>,  
Emma L. Tonkin<sup>†</sup>, Alessandro Masullo<sup>†</sup>, Antonis Vafeas<sup>†</sup>, Mickey Kim<sup>†</sup>, Roisin McNaney<sup>‡</sup>,  
Gregory J. L. Tourte<sup>§</sup>, Alan Whone<sup>\*</sup>

<sup>\*</sup> Translational Health Sciences, University of Bristol, Bristol, UK.

<sup>†</sup> School of Computer Science, Electrical and Electronic Engineering, University of Bristol, Bristol, UK

<sup>‡</sup> Department of Human Centred Computing, Monash Data Futures Institute, Australia

<sup>§</sup> School of Geographical Sciences, University of Bristol, Bristol, UK

<sup>¶</sup> Contact Email: catherine.morgan@bristol.ac.uk

**Abstract**—This paper looks to explore the challenges faced when producing a set of annotations from videos produced by a pilot study evaluating 24 participants (12 with Parkinson’s disease, each accompanied by a healthy volunteer control participant) who are free-living in a house embedded with a platform of sensors. We discuss the outcome measures chosen to annotate from the videos and the controlled vocabularies formulated for this task, the tools and processes, how we intend to achieve standardisation and normalisation of the annotations, and how to improve quality and re-usability of the annotation dataset.

**Index Terms**—Parkinson’s disease, outcome measures, annotations

## I. INTRODUCTION

Parkinson’s disease (PD) is the second most common neurodegenerative disease in the UK and worldwide [1]. It is a progressive, disabling disease for which currently there are no disease-modifying treatments, despite multiple putative neuroprotective agents having been tested in double-blind randomised controlled trials [2]. The gold-standard way to measure disease progression in the clinic and in clinical trials is by using a clinical rating scale called the Movement Disorders Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale, or MDS-UPDRS. However, PD fluctuates day-to-day and hour-by-hour, so quantifying it in the ‘wild’, at home, is felt to be important by many patients and clinicians.

A home-based sensor system which continuously monitors symptoms of PD has the potential to personalise medical treatments, reduce clinic visits in both clinical care and trials, and better appreciate PD symptom fluctuations which occur away from clinic.

Of particular interest in clinical trials outcomes are those symptoms which are present pre-diagnosis/in early disease and those which gradually worsen over the early years from diagnosis. Such symptoms include bradykinesia (slowness of movement), rigidity (stiffness in the muscles), sleep disturbances, hypophonia (low/soft voice), constipation.

To improve the accuracy of outcome measures that aim to quantify progression of these symptoms in a free-living environment, excellent training data would involve multiple expert clinicians’ annotation of a participant’s symptoms [3].

Evaluation of ‘free-living’ technology-assisted outcomes in PD, in other words when a person is in the home or home-like environment and carrying out unstructured/unscripted activities, has been done by a number of groups with varying degrees of validation [4] which include video stream analysis with annotations, direct clinician observation and participant diaries. Video use to produce annotations has advantages over diaries which can have poor temporal resolution and low accuracy [5] and direct clinician observation that can impair the participant’s free-living behaviour [6]. Video use in validation attempts of free-living data was mentioned in 19 papers according to a recent systematic review [4]. We would like to discuss the challenges of producing annotations from videos taken in the wild.

A project from the University of Bristol has led to the development of a platform of sensors that are designed to allow continuous, relatively inexpensive, unobtrusive sensing of naturalistic living symptoms and activities [7]. The ‘SPHERE house’ (SPHERE stands for Sensor Platform for HEalthcare in a Residential Environment) is a 2-bedroomed terraced property with a kitchen, 2 reception rooms, bathroom and a small garden. This setting is embedded with multiple multi-modal sensors [8]. The sensors include environmental/ambient sensors which measure temperature, humidity, air pressure, light levels, presence (passive infra-red), wearable devices (worn on both wrists) with accelerometers, camera sensors which generate silhouettes of participants and bounding boxes—that when linked with the SPHERE wearable, let the system know who the participant is, appliance use sensors, water flow sensors and several receivers distributed across the house also obtain the signal strength from the wearable, which is used to estimate the location of the participant in the house. The data from all these sensors, including the wearables, synchronise into the same database sharing the same timestamps. The sensors are easily and inexpensively deployable to people’s own homes.

Common (scripted) activities of daily living (ADLs), including reaching up to cupboards and pouring a kettle and other actions such as sit-to-stand, have been annotated offline from videos taken in a home-like environment by groups within the SPHERE Project [9].

Additionally, rather than using cameras for post-hoc data

annotation, work has explored approaches to live annotations centred on a cooking activity in the participants' home [10].

Another approach to online annotation was explored by Tonkin et al. [11], whereby participants annotated their own data, during their own free-living, using a self-annotation app: users could scan a tag triggering the app to display a confirmation message. Work by McConville et al. showed participants following a scripted procedure to automatically annotate their own data [12] for localisation and activity recognition.

Further work comparing the performance of long-term activity recognition using accelerometers concluded that it is more difficult to estimate the performance of these methods in the wild than in a laboratory environment [13]. This highlights the need for ground truth to accompany free-living data sets to improve prediction accuracy in activity recognition.

This paper looks to explore the challenges faced when producing a set of annotations from videos produced by a pilot study evaluating 24 participants (12 with PD each accompanied by a healthy volunteer control participant) who are free-living in the SPHERE house whilst being sensed by the platform of sensors described above. The house also has camera sensors which can capture RGB (red-green-blue) video with which to produce a source of ground truth (clinician annotations post-hoc) for the behaviour and symptoms of the participants. Longer term, the study is looking to produce technology-assisted outcome measures which could then potentially be used to measure disease progression in clinical trials of disease-modifying therapies in PD.

We discuss the outcome measures chosen to annotate from the videos and the controlled vocabularies formulated for this task, the tools and processes we intend to use, how we intend to achieve standardisation and normalisation of the annotations, and how to improve quality and re-usability of the annotation dataset.

## II. METHODOLOGY

We have started data collection for a feasibility and acceptability study during which 12 pairs of participants are recruited: each pair comprises one person with mild to moderate PD and one healthy volunteer control participant who is likely to be a spouse, family member or friend of the person with PD. Each pair will stay and live freely for 5 days in the SPHERE house.

During the 5-day stay, researchers visit twice to conduct clinical assessments and rating scales, in addition to some pre-defined scripted activities which focus on kitchen tasks. Otherwise, the participants are encouraged to live as they would at home. They stay overnight for a total of 4 nights.

Technology-assisted outcome measures of interest for which annotations will be generated are shown in table I, along with their controlled vocabularies.

RGB video data is captured from communal rooms downstairs. This data is recorded during the time the researcher is undertaking clinical assessments of the participants, and while they are doing scripted activities in the kitchen. Furthermore, while the participants are free-living alone in the house, the cameras capture 4 hours of RGB video across the 5 days. The participants are aware of the times when the RGB will be

TABLE I  
CONTROLLED VOCABULARIES FOR ANNOTATIONS

<b>Activity level</b>
Lying down, Sitting, Sit-to-stand, Stand-to-sit, Standing still, Standing with activity (e.g., washing dishes), Light walking (<4 km/h), Moderate physical activity
<b>Activities of daily living</b>
Watching TV, Reading, Food preparation, Cleaning, Talking on phone, Eating, Playing a game, Talking to someone, Using computer/tablet/other device
<b>Global spontaneity of movement</b>
Normal: no problems. Slight: slight global slowness and poverty of spontaneous movements. Mild: mild global slowness etc. Moderate: moderate global slowness etc. Severe: severe global slowness etc.
<b>Location</b>
Kitchen, Living room, Hall, Dining room, Stairs, Garden, Outside front door.
<b>Gait</b>
Normal: no problems. Slight: independent walking with minor gait impairment. Mild: independent walking but with substantial gait impairment. Moderate: requires an assistance device for safe walking but not a person. Severe: cannot walk at all or only with another person's assistance.
<b>Sit-to-stand (impairment)</b>
Normal: no problems. Able to arise quickly without hesitation. Slight: arising is slower than normal; may need more than one attempt, may need to move forward in chair to arise. No use of chair arms. Mild: pushes self up from arms of chair without difficulty. Moderate: needs to push off; may need to try more than once to arise. Can get up without help. Severe: unable to arise without help.

captured, partly to ensure they avoid leaving the house for exercise, for example.

The spectrum of outcomes to be labelled from the RGB videos are designed to investigate a variety of patient-centred, clinical trial-relevant and potentially measurable metrics that would be possible and interesting to measure for a longer period in peoples' own homes in a future study.

A widely-available software called ELAN [14] will be used to watch up to 4 simultaneously-captured videos running concurrently which will show both study participants' movements and activities around the kitchen, dining room, living room and hall of the SPHERE house. ELAN creates annotations on multiple layers called 'tiers' which can be inter-linked, share controlled vocabularies, share timestamps, or share other categories.

The primary annotator will be a medically trained neurologist who will view the videos offline to annotate each tier. At their discretion and according to their experience and confidence with ELAN, multiple tiers could be annotated in the same video viewing session (e.g., location and activity). A second rater, a Movement Disorders specialist neurology consultant, who will annotate 10% of the randomly selected videos, blinded to the ratings produced by the first rater. The first rater will re-rate a selection of videos at an interval of >1 week from their original annotations.

This study has been approved by a Research Ethics Committee.

### III. DISCUSSION

Inherent in the use of researchers to produce labels from watching videos is the challenge posed by human error and human nature. For instance, as a single person evaluates a large quantity of video data. For example, there may be intra-rater variability induced by fatigue after several hours/days of annotating. To improve best practice in annotations, the use of a second rater, blinded to the annotations of the first rater, should be helpful to reduce intra-rater variability. However, then inter-rater variability will be introduced. Of note, it is appreciated that, even in controlled laboratory conditions with trained raters, the Parkinson's disease UPDRS motor score (a sub-score of the rating scale that has a total possible score of 56) has an inter-rater variability of up to 16 points [15].

While most machine learning methods assume that there exists a single true label for each data point, typically coming from a single source, it is possible for variability in annotations and ratings to be modelled directly by the machine learning methods. There exists a plethora of work in this area, including within the medical domain [16], where it is not uncommon for multiple experts to have disagreeing subjective opinions on a diagnosis, as well as the smart home domain [17]. The most relevant methods for this task may be those that attempt to estimate the true label from annotations by multiple experts [18], [19] and those that attempt to model directly the accuracy of annotators, for example by using Expectation Maximization [20]. Methods which incorporate multiple annotations have been shown to improve classification accuracy of supervised machine learning models [21].

The human participant(s) are likely to pose challenges for the annotators of this dataset. A participant may move out of view of the camera sensor (e.g., move to a bedroom without a camera, go behind a door or have their back to the camera), so the continuity of movement and activities would be disrupted. Knowledge of when the cameras are recording could induce bias in the participants' activities.

The hardware itself can provide challenges: hardware can be placed in locations at risk of blunt trauma, for example near the floor where they could be knocked accidentally by participants passing them. Also, the video resolution may not be enough for capturing finer-grained PD symptoms (e.g., bradykinesia of the upper limb). To mitigate this risk, we conducted several pilot exercises capturing videos of different frame rates and consulted 3 clinicians with movement disorders expertise about the minimum frame rate necessary to see the movements being evaluated.

The controlled vocabulary chosen for the annotation metrics presents its own set of unique challenges. Developing an annotation system is often an iterative process. Whilst it is helpful to ensure that the phrases and words within a controlled vocabulary are standardised to the extent that they are understood easily by other research groups and can be re-used if desired, they are also necessarily tailor-made for the use case in question including the participants, setting and activities under investigation.

For the PD symptom metrics, there is no widely-accepted controlled vocabulary for creating annotations from this kind of video data. The aim of this study is to produce technology-assisted outcome measures to complement, not to replace, the MDS-UPDRS. However, given that, by its very nature, the clinical assessment of items such as a patient's gait is subjective and non-linear with risk of observer bias, this naturally must extend to a clinician's assessment of PD symptoms on a video. When considering how best to create labels for the PD symptoms of global spontaneity of movement, gait and sit-to-stand ability, there was no natural rating scale/label set which was felt more helpful or appropriate than the 5-point MDS-UPDRS severity scale for these outcomes. This is a validated scale [22] for use in people with all severities of PD with clear examples of all sub-ratings of each symptom within training videos on the Movement Disorders Society website (for members). Therefore, despite its limitations, we are planning to annotate the videos for these outcomes using the severity scale within the MDS-UPDRS (see table I).

For the non-PD symptoms, outcomes such as room location are straightforward to create a controlled vocabulary for (e.g., kitchen, living room, hall). Less easy to quickly define were the activity recognition outcome measures. Interestingly, some groups have found their accelerometry data analytics have difficulty with distinguishing between standing still and sitting [23], so to increase the information we provide within the training dataset we have made a distinction between standing still and standing with light activity (e.g., washing dishes). We have followed an iterative path so far, preparing a controlled vocabulary which is in-line with that previously prepared by our colleagues [24] but adapting it once the annotations started being collected, according to its usefulness to the computer science team and its ease of use to the annotators.

For continuous behaviour or symptom severity (e.g., global spontaneity of movement), the sampling frequency of labels has been a subject of discussion amongst our group. Acknowledging frequent fluctuations of PD symptoms, the aim is to capture them using the technology-assisted outcome measures. However, this is balanced against both the burden of annotations needed from many dozens of hours of video recordings and the pragmatic approach that most people with PD will not fluctuate in their motor symptoms minute-to-minute. Currently, we are planning to annotate on a minute-by-minute basis where PD symptom severity is evaluated and labelled for the preceding 60 seconds of video footage.

### IV. CONCLUSION

This study is undertaking ambitious and wide-ranging annotation of video data capturing free-living symptoms and activities from people with Parkinson's disease and healthy volunteer control participants. The annotations will be undertaken post-hoc by multiple clinician raters and the challenges posed by this task will be explored and reviewed iteratively. It is hoped that this will inform future research looking to obtain ground truth from datasets in the wild in Parkinson's research, to better understand this complex and fluctuating disease.

Through participating in a workshop, we would hope to stimulate conversation around the following questions:

- What are others' experiences with working with clinician raters?
- What levels of inter-rater reliability are appropriate for a gold standard set of this kind?
- Is this the right approach to be taking with free-living annotation?
- What are some of the approaches people have taken to reduce rater fatigue?

#### ACKNOWLEDGEMENTS

We gratefully acknowledge the current and future study participants for their time and efforts in participating in this research.

This work was performed under the SPHERE Next Steps Project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/R005273/1.

This work is supported by the Elizabeth Blackwell Institute for Health Research, University of Bristol, and the Wellcome Trust Institutional Strategic Support Fund, grant code 204813/Z/16/Z, by The Cure Parkinson's Trust, grant code AW021 and by IXICO plc, grant code R101507-101.

Dr Jonathan de Pass and Mrs Georgina de Pass made a charitable donation to the University of Bristol, through the Development and Alumni Relations Office, which pays the salary of CM.

#### REFERENCES

- [1] Theo GBD 2016 Disease and Injury Incidence and Prevalence Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016." *Lancet (London, England)*, vol. 390, pp. 1211-1259, Sep. 2017.
- [2] A. E. Lang, E. Melamed, W. Poewe, and O. Rascol, "Trial designs used to study neuroprotective therapy in parkinson's disease," *Movement Disorders*, vol. 28, no. 1, pp. 86-95, 2013.
- [3] L. Lonini, A. Dai, N. Shawen, T. Simuni, C. Poon, L. Shimanovich, M. Daeschler, R. Ghaffari, J. Rogers, and A. Jayaraman, "Wearable sensors for parkinson's disease: which data are worth collecting for training symptom detection models," *npj Digital Medicine*, vol. 1, 2018.
- [4] C. Morgan, M. Rolinski, R. McNaney, B. Jones, L. Rochester, W. Maetzel, I. Craddock, and A. Whone, "Systematic review looking at the use of technology to measure free-living symptom and activity outcomes in parkinson's disease in the home or a home-like environment," *Journal of Parkinson's Disease*, vol. 10, no. 2, pp. 429-454, Apr. 2020.
- [5] J. Reimer, M. Grabowski, O. Lindvall, and P. Hagell, "Use and interpretation of on/off diaries in parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 3, pp. 396-400, 2004.
- [6] V. Robles-García, Y. Corral-Bergantiños, N. Espinosa, M. A. Jácome, C. García-Sancho, J. Cudeiro, and P. Arias, "Spatiotemporal gait patterns during overt and covert evaluation in patients with parkinson's disease and healthy subjects: Is there a hawthorne effect?" *Journal of Applied Biomechanics*, vol. 31, no. 3, pp. 189-194, Jun. 2015.
- [7] P. Woznowski, A. Burrows, T. Diethé, X. Fafoutis, J. Hall, S. Hannuna, M. Camplani, N. Twomey, M. Kozlowski, B. Tan, N. Zhu, A. Elsts, A. Vafeas, A. Paiement, L. Tao, M. Mirmehdi, T. Burdhardt, D. Damen, P. Flach, R. Piechocki, I. Craddock, and G. Oikonomou, *SPHERE: A sensor platform for healthcare in a residential environment*. United Kingdom: Springer, Dec. 2016, pp. 315-333.
- [8] N. Zhu, T. Diethé, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock, "Bridging ehealth and the internet of things: The sphere project," *IEEE Intelligent Systems*, vol. 30, no. 4, pp. 39-46, Aug. 2015.
- [9] E. Tonkin, M. Perello Nieto, H. Bi, and A. Vafeas, "Towards a methodology for acceptance testing and validation of monitoring bodyworn devices," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops 2020)*. United States: IEEE Computer Society, 2020.
- [10] E. L. Tonkin, O. Bykowska, H. Berg, and I. Craddock, "Towards estimation of cooking complexity: Free-text annotations in the kitchen environment," in *Proceedings of the 6th International Workshop on Sensor-Based Activity Recognition and Interaction*, ser. iWOAR '19. New York, NY, USA: Association for Computing Machinery, 2019.
- [11] E. Tonkin, A. Burrows, P. Woznowski, P. Laskowski, K. Yordanova, N. Twomey, and I. Craddock, "Talk, text, tag? understanding self-annotation of smart home data from a user's perspective," *Sensors*, vol. 18, no. 7, Jul. 2018.
- [12] R. McConville, D. Byrne, I. Craddock, R. Piechocki, J. Pope, and R. Santos-Rodriguez, "A dataset for room level indoor localization using a smart home in a box," *Data in Brief*, vol. 22, pp. 1044-1051, 2019.
- [13] N. Twomey, T. Diethé, X. Fafoutis, A. Elsts, R. McConville, P. Flach, and I. Craddock, "A comprehensive study of activity recognition using accelerometers," *Informatics*, vol. 5, no. 2, p. 27, May 2018.
- [14] H. Brugman and A. Russel, "Annotating multi-media/multi-modal resources with ELAN," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004.
- [15] J. L. Palmer, M. A. Coats, C. M. Roe, S. M. Hanco, C. Xiong, and J. C. Morris, "Unified parkinson's disease rating scale-motor exam: inter-rater reliability of advanced practice nurse and neurologist assessments," *Journal of advanced nursing*, vol. 66, no. 6, pp. 1382-1387, 2010.
- [16] A. W. S. Rutjes, J. B. Reitsma, A. Coomarasamy, K. S. Khan, P. M. M. Bossuyt *et al.*, "Evaluation of diagnostic tests when there is no gold standard. a review of methods," *Health Technology Assessment*, vol. 11, no. 50, 2007.
- [17] R. McConville, D. Byrne, I. Craddock, R. Piechocki, J. Pope, and R. Santos-Rodriguez, "Understanding the quality of calibrations for indoor localisation," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. IEEE, 2018, pp. 676-681.
- [18] H. Valizadegan, Q. Nguyen, and M. Hauskrecht, "Learning classification models from multiple experts," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 1125-1135, Dec. 2013.
- [19] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*, 2009, pp. 889-896.
- [20] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20-28, 1979.
- [21] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," *CoRR*, vol. abs/1703.08774, 2017.
- [22] C. Goetz, B. Tilley, S. Shaftman, G. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. Lang, A. Lees, S. Leurgans, P. Lewitt, D. Nyenhuis, and N. Lapelle, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Movement disorders : official journal of the Movement Disorder Society*, vol. 23, pp. 2129-2170, 2008.
- [23] K. Bakrania, T. Yates, A. V. Rowlands, D. W. Esliger, S. Bunnewell, J. Sanders, M. Davies, K. Khunti, and C. L. Edwardson, "Intensity thresholds on raw acceleration data: Euclidean norm minus one (enmo) and mean amplitude deviation (mad) approaches," *PloS one*, vol. 11, no. 10, p. e0164045, 2016.
- [24] P. Woznowski, E. Tonkin, and P. Flach, "Activities of daily living ontology for ubiquitous systems: Development and evaluation," *Sensors*, vol. 18, no. 7, Jul. 2018.