IPEM-Translation 1 (2022) 100004

Contents lists available at ScienceDirect

IPEM-Translation

journal homepage: www.elsevier.com/locate/ipemt

Technical Note

Characterizing sensor accuracy requirements in an artificial intelligence-enabled medical device

Kristin A. Bartlett*, Katharine E. Forth, Stefan I. Madansingh

Balance Research Institute, ZIBRIO Inc., 2450 Holcombe Blvd. Suite X, Houston, TX 77021, USA

ARTICLE INFO

Keywords:
Medical device design
Requirements engineering
Artificial intelligence
Machine learning
Sensory accuracy

ABSTRACT

Artificial intelligence and machine learning applications are increasingly prevalent in the healthcare industry. In some cases, medical devices use sensor-collected data to feed into algorithms which generate scores or risk assessments that are used to inform patient care. The process of determining sensor accuracy requirements which will ensure that the algorithm generates reliable scores is not straightforward or well-defined. In this paper, we describe a simulation-based method to characterize sensor accuracy requirements for a device that uses a machine-learning algorithm to generate a postural stability score – the ZIBRIO Stability Scale. The results of the simulation are described, as is the application to sensor selection in preparation for manufacturing of the device. Other medical device developers may be able to use this method or similar methods in their requirements engineering process.

Introduction

The healthcare industry is increasingly making use of artificial-intelligence (AI) or machine learning based methods [1–3]. Multiple trends are driving this growth, including a shortage of physicians in the midst of increasing demand for healthcare services [4], and an increased focus on patient-centered care [1]. Artificial-intelligence applications can help make healthcare more personal, predictive, preventative, and participatory [3], and can enable cost savings for healthcare systems [4]. Popular application areas include machine learning for biomarker discovery, autonomous robotic surgery, clinical outcome prediction and patient monitoring, inferring health status through wearable devices, and image-based diagnosis [5]. The most common types of data discussed in AI literature for healthcare include imaging data, genetic data, and electrophysiological data [2].

Alongside the benefits that machine learning healthcare applications promise, they also bring ethical concerns [6–8]. Some of these concerns include bias in training data sets, privacy of personal data, lack of accountability for poor outcomes [6], and amplification of existing health inequities [7]. In some cases, even “fair” or “unbiased” systems can be inherently unethical [8]. Algorithmic bias is an important concern not only in healthcare, but across all sectors that leverage data-driven algorithms [9]. Algorithmic bias is “what we experience when a machine-learning model produces a systematically wrong result” [10]. Algorithms can be morally, statistically, or socially biased [9], and there is already plenty of evidence of the occurrence of bias in AI [10]. Because of the concern of algorithmic bias, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems is developing a standard for algorithmic bias considerations [11].

Many have discussed the importance of minimizing bias in the data used to train algorithms. For example, due to inequality in access to healthcare services, training datasets may contain more data of affluent, overserved patients [12]. However, the literature on bias in healthcare AI has given less attention to another potential source of bias: measurement error in data collected from sensors. In some cases, sensors themselves may have inbuilt bias and not work as well for some populations as for others. For example, racial bias was demonstrated in pulse oximetry measurement, where readings were more accurate for white patients than for Black patients [13]. Therefore, any healthcare AI application that uses pulse oximetry data as an input should consider the impact that this measurement bias will have on the AI application’s outputs or recommendations.

Fig. 1 illustrates a flow that is relevant to many AI applications that involve sensor-collected data. The sensor, which may be embedded in an imaging device, wearable device, or other medical device, outputs a measurement, which is then fed into an AI or machine-learning-powered algorithm, which produces a result or score that is used to inform patient care. If initial sensor measurements are inaccurate, how does this impact the final output from the algorithm, and what is the resulting impact on the advice or recommendation given to a patient? There is a possibility that any error in sensor measurement, due to bias, noise, or another source, feeds inaccurate data into an algorithm and reduces the accuracy of the algorithm’s final output.

* Corresponding author.
E-mail address: kristi@zibrio.com (K.A. Bartlett).

https://doi.org/10.1016/j.ipemt.2022.100004
Received 15 February 2022; Received in revised form 3 April 2022; Accepted 4 April 2022
Available online 10 April 2022

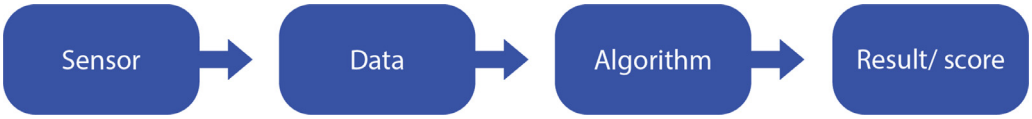

Fig. 1. Flow of data from sensor to algorithmic output.

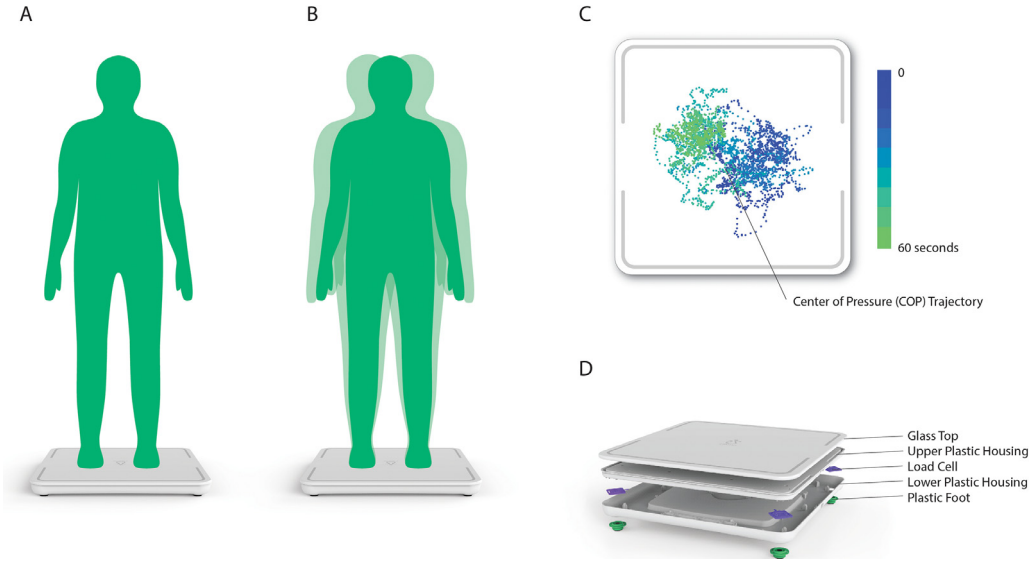

Fig. 2. Reproduced (with permission) from the work by Forth et al. [18]. A) Person stands still on device for 60 s for postural stability assessment. B) During the assessment, the person's body will make small movements, which are captured in COP measurements. C) Plot of COP measurements collected by the scale during the postural stability assessment. D) Location of load cell sensors inside the device.

Fall risk assessment has been a popular focus in recent fall-related research [14], with important applications in the wellbeing of older adults [15] as well as in the prevention of occupational falls [16]. Others have discussed sensor-based fall risk assessments, including the validation methods of the assessments [14], but did not discuss the impact of sensor accuracy on the accuracy of fallrisk predictions. Others still have discussed machine-learning based fall detection and fall prevention methods, covering the types of sensors used and the accuracy of algorithmic outputs [15], but did not address the concern of sensor accuracy and its algorithmic impact. The question of the impact of sensor accuracy on sensor or machine-learning -based fall risk assessments remains a gap in the literature.

This work presents a simulation-based method developed to assess the impact of sensor noise on algorithmic outputs from a medical device used to assess fall risk, the ZIBRIO Stability Scale. Results of the simulations were used to determine sensor measurement accuracy requirements for production of the device. While this method was developed for one specific application, it may be adapted to other devices where sensor-based signals inform a machine learning algorithm.

The ZIBRIO Stability Scale (formerly called the ZIBRIO SmartScale) was designed to provide an accessible method of measuring human standing balance and postural stability through the study of posturography [17]. The Stability Scale uses the Briocore algorithm (Forth & Lieberman-Aiden, 2019) to calculate postural stability scores (PS scores), which are reported as integers ranging from 1 (worst balance) to 10 Brios (best balance). The Briocore algorithm uses 60 s of center-of-pressure (COP) data (3600 samples) collected during a postural stability test to calculate the PS score. The COP data is collected through four load cell sensors that are located at the corners of the device platform. The device is illustrated in Fig. 2 [18]. In a previous study, the accuracy of the Stability Scale's COP measurements was found to be within 0.5 mm of measures taken from a "gold-standard" laboratory force plate [17].

Materials and methods

A pool of 30 postural stability test datasets was compiled for the analysis. The pool was comprised of 3 randomly selected datasets for each PS score (1 to 10 Brios, inclusive). All datasets were gathered from a database of postural stability tests of human subjects, with the exception of the scores of 10 Brios, which were produced by running the postural stability assessment with a static calibration load (mass of 45.5 kg) placed atop the Stability Scale. Datasets from human subjects were collected in the study described in Forth et al. 2020 [18] (experimental protocol was approved by the Westerns IRB (#20171926 and #20172324) and the University of Texas Health Science IRB (HSC-MS-16–0019)). During testing, one human subject's data set which scored 2 Brios was observed to be more sensitive to sensor noise than others. To better capture the observed variability, two additional data sets were added to the pool of "2 Brios" scores. All five datasets with scores of 2 Brios were included in the analysis to account for variability captured by the unexpected result, bringing the total number of datasets analyzed to 32.

To evaluate the impact of sensor noise on PS score calculation, we ran a series of simulations in which increasing levels of random sensor noise, or "perturbations," were iteratively added to COP measurements in each test dataset. Forty levels of perturbation magnitude were selected, ranging from 0.25 mm to 10 mm in increments of 0.25 mm. A single dataset contains 3600 COP measurements. For each COP sample in a dataset, a perturbation was randomly selected within each magnitude range (e.g. between −0.25 and +0.25 mm) and was added to the x component (COPx), y component (COPy), or both x and y components. The perturbations were performed in four different conditions: perturbing COPx only (Condition 1), perturbing COPy only (Condition 2), perturbing COPx and COPy by the same amount (Condition 3), and perturbing COPx and COPy by different amounts (Condition 4). The conditions are depicted visually in Fig. 3. All perturbation calculations and

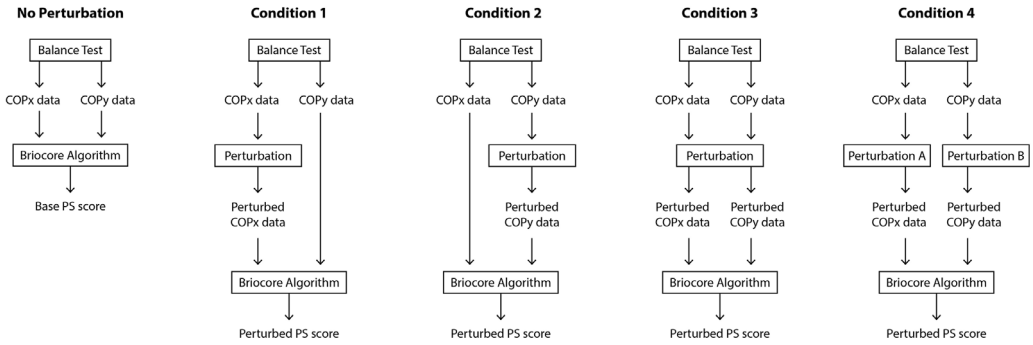

Fig. 3. Visualization of PS score perturbation process compared to base PS score calculation process ("No Perturbation" condition).

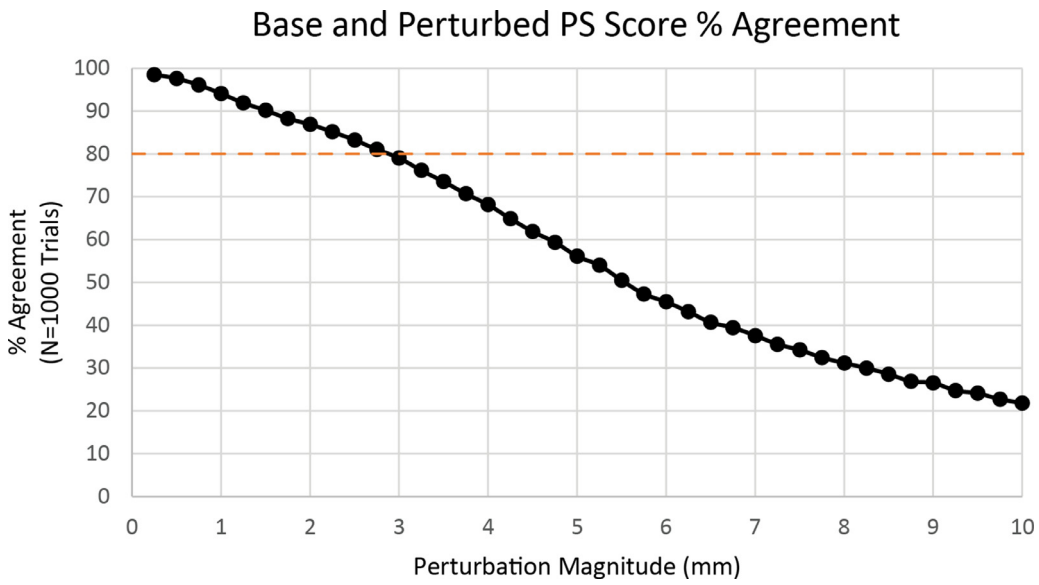

Fig. 4. Percent agreement between base PS score and perturbed PS score.

the evaluation of their impact upon the PS scores were performed in Matlab v. R2017a (The Math Works Inc. 2017).

An example process for perturbation of a dataset was as follows: A dataset was imported into Matlab and the COPx and COPy trajectories were calculated from the raw load cell sensor data. For "Condition 1″ (COPx only), for the perturbation magnitude of 0.25 mm, a random number between −0.25 and 0.25 was added to each COPx value in the dataset (3600 samples). This produced a new set of COP values, with the COPx values "perturbed" and the COPy values unchanged. The Briocore algorithm was then applied to this modified set of COP values to produce a new PS score. This process was then repeated for the other 39 perturbation magnitudes; therefore, 40 different output PS scores were generated for each level of added noise. This process was repeated among the other three perturbation conditions.

Once this was complete, each initial dataset had been perturbed at 40 noise magnitudes under four conditions, resulting in a total of 160 distinct simulations. Each simulation was repeated 1000 times to characterize the behavior of the perturbed score outcome at each perturbation range. Since random noise was added during each simulation, the results of each trial were expected to vary slightly (Monte Carlo simulation) and produce a distribution of results. The 160 distinct simulations were performed 1000 times on each of the 32 balance test data sets.

The new PS scores resulting from the simulations were compared with the base PS score, which was the result of the Briocore algorithm calculation on the original, unperturbed dataset. Out of all the perturbation conditions, Condition 4, perturbing COPx and COPy by different values, was found to have the largest impact on PS score outcomes. This analysis was prioritized as the worst-case scenario noise profile. In the presented results, 32 initial postural stability test datasets were perturbed in COPx and COPy by different values at each of the 40 perturbation magnitudes 1000 times. To quantify the impact of noise upon PS score, the percent agreement between perturbed PS score and the base PS score across the 1000 trials was calculated at each perturbation magnitude. In cases of disagreement, the directionality of disagreement between the perturbed and base score (whether the PS score was overestimated or underestimated relative to the base score) was also assessed.

In practice, an incorrectly high PS score would overestimate an individual's postural stability, while a lower PS score would underestimate postural stability. The PS score has been associated with a fall risk prediction [18], and from a conservative standpoint, overestimating fall risk is preferred. Wrongfully overestimating postural stability was considered to represent unacceptable PS score sensitivity to sensor measurement error, therefore the percentage of overestimated PS scores was also calculated at each perturbation level.

Results

A perturbation magnitude of +/- 3 mm was selected as the highest acceptable error level in the COP measurements. At +/- 3 mm of error, the average agreement between the base PS score and the perturbed PS score was 79.01+/−17% (Fig. 4). In general, agreement between the base PS score and perturbed PS scores decreased as the perturbation magnitude increased. At perturbations magnitudes below +/- 3 mm, score agreement was at or above 80%, and agreement steadily decreased to 20%

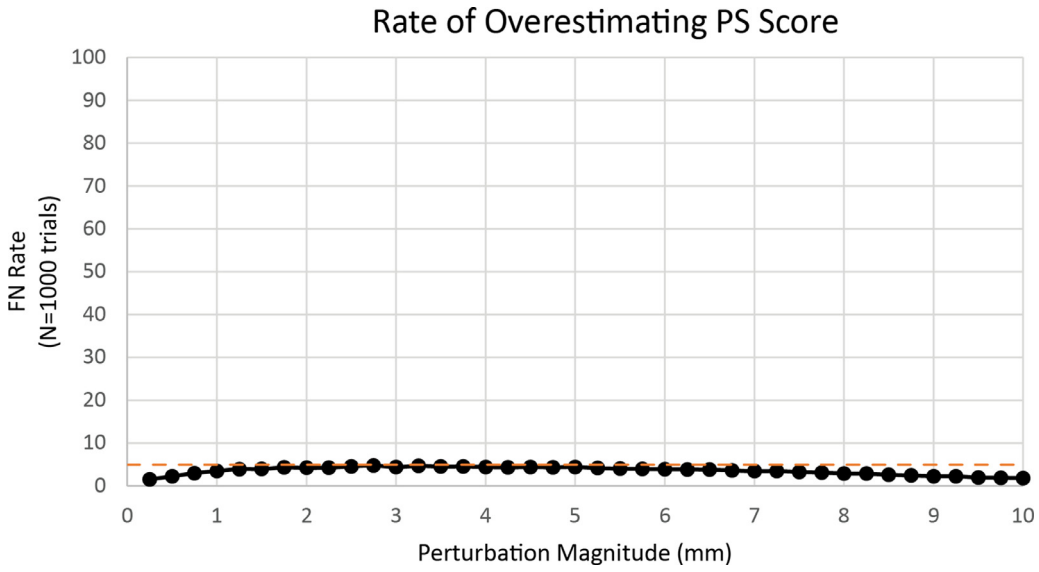

Fig. 5. Rate of overestimating PS score.

at +/- 10 mm perturbation magnitude. At +/- 3 mm, the percentage of perturbed PS scores that was either equal or lesser to the base PS score was 95.54+/−8% (Fig. 5). As a general rule, the outcome of the simulations showed that the errors in sensor measurement would serve to cause an underestimation of PS score about 95% of the time, even in cases where greater than +/- 3 mm of error was simulated.

Two additional datasets were included to evaluate the sensitivity of a PS score of "2 Brios" after one of the three randomly selected datasets, "dataset 2C", responded to perturbations in a manner that was inconsistent with simulations. For example, the percent agreement between base and perturbed score at +/- 0.25 mm (lowest perturbation magnitude) was 46% at where all other simulations had percent agreements of approximately 100% at the same level The two additional datasets followed the expected pattern near 100% agreement at the lowest perturbation magnitude, however, all five datasets" were included in the final analysis. Investigation into dataset 2C revealed that the unrounded PS score equalled 2.4666, representing a unique case, highly susceptible to having the rounding direction changed if the score increased by as little as 0.0334 points.

Discussion

In light of the increasing number of medical devices that generate algorithm-based scores, it is important for device manufacturers to ensure that the sensors used in such devices have adequate resolution so as to not compromise the accuracy of the algorithm's outputs. Medical sensors are sometimes used to generate an output in the same unit of measure as what is reported on the specification sheet from the manufacturer. For example, a load cell used in a weighing scale application will have a known weight accuracy resolution, which can inform the resulting accuracy of the overall weight scale in a relatively straightforward manner. However, when sensor measurements are being used as input for an algorithmic score, the impact of the sensor accuracy on the final score output is not as straightforward. In order to address this problem, we devised a method to simulate sensor noise at varying levels and evaluate how this would affect the resulting algorithmic score output.

Based on the results of our simulations, +/- 3 mm was selected as the tolerable error for COP measurements in the ZIBRIO Stability Scale device. The rationale for this selection was a balance of several factors: firstly, to ensure a high level of stringency in the accuracy of the PS score; secondly, to ensure a large margin for error relative to the measured MAE, as +/- 3 mm is approximately 5x the MAE when compared with a laboratory force plate [17]. At +/- 3 mm perturbation magnitude, 95.54% of the simulated PS scores were either equal to or lower than the original base score, providing confidence that a user will receive a representative and nominally conservative estimate of their fall risk. Approximately 80% of the perturbed PS scores (79.01%) were equivalent to the base PS score at this level of noise, therefore, +/- 3 mm was determined to be an acceptable requirement.

From an engineering standpoint, +/- 3 mm accuracy is possible to achieve with small, lightweight, low-cost sensors which could be used in the production of our small, portable posturography devices. In testing, both the current generation and previous generation of Stability Scale prototypes had maximum absolute error below the +/- 3 mm requirement (when compared to a gold standard laboratory force plate), and can be considered adequate in calculating COP for the purposes of assessing postural stability through the Briocore algorithm.

In terms of sensor performance, realizing an error of +/- 3 mm during COP measurement is estimated to require approximately 1.1 kg of load to be erroneously captured (0.28 kg at each sensor) in each direction (x and y). This is well outside the expected error limits of the low-cost 50 kg GML671 load cells which are used in the device; the load cells have a comprehensive error rate of 0.1% of full scale (approximately 0.05 kg). Using the COP equation described in Bartlett et al. 2019 [17], 3 mm of error in the x or y direction, when measuring a 65 kg load (637.43 N), would represent 1.09 kg of measurement error across all 4 sensors, or 0.27 kg of measurement error in a single sensor.

Although the results of our analysis suggest that the sensors in the 4th generation Stability Scale are adequate to achieve the desired measurement accuracy and noise tolerance, it is important to note the following limitations. The analysis was performed on a test pool of only 32 postural stability tests. The simulations run during algorithm sensitivity analyses were intended to account for unexpected changes to COP measurement accuracy, which could arise from mechanical issues, electrical noise, or sensor flaws or damage. However, this simulation method did not capture the likelihood of incorrectly calculating or reporting PS scores due to software issues. Therefore, further software verification activities would be needed to ensure PS score accuracy during regular use.

When medical sensors are used to generate algorithmic scores or outputs, sensor accuracy requirements are not always straightforward. Sensors are unlikely to give perfect readings at all times when used in real-world conditions, and device manufacturers must determine acceptable

specifications for sensor accuracy. Here we have described a simulation-based method to determine such accuracy requirements in light of the resulting impact on algorithmic score outputs. While we have described a replicable method to determine how sensor inaccuracy will impact algorithmic score outputs, resulting judgements for sensor accuracy tolerance will vary in other applications. In medical device applications, manufacturers must determine sensor accuracy requirements based on the clinical and ethical implications, paying special attention to the consequences of incorrect predictions upon patient care. Clear evaluations of algorithm noise susceptibility will be necessary for future adoption of algorithm-derived healthcare insights.

Ethics statement

Data from human participants was collected in studies described in Forth et al. [18] which were approved by Westerns IRB (#20171926 and #20172324) and University of Texas Health Science IRB (HSC-MS-16-0019).

Acknowledgements

N/A.

Declaration of Interest Statement

ZIBRIO, Inc., a privately held company, provided financial support for this study. No external funding was provided for this study. Authors are employees of ZIBRIO, Inc.

References

[1] R. Bhardwaj, A.R. Nambiar, D. Dutta, A study of machine learning in healthcare, in: In: 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC); 2017, Turin: IEEE, 2017, pp. 236–241. http://ieeexplore.ieee.org/document/8029924/. Available from.

[2] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, et al., Artificial intelligence in healthcare: past, present and future, Stroke Vasc. Neurol. 2 (4) (2017) 230–243.

[3] G. Rong, A. Mendez, E. Bou Assi, B. Zhao, M. Sawan, Artificial intelligence in healthcare: review and prediction case studies, Engineering 6 (3) (2020) 291–301.

[4] A. Bohr, K. Memarzadeh, in: The Rise of Artificial Intelligence in Healthcare applications. In: Artificial Intelligence in Healthcare [Internet], Elsevier, 2020, pp. 25–60.

[5] K.-.H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018 Oct) 719–731.

[6] D.S. Char, M.D. Abràmoff, C. Feudtner, Identifying ethical considerations for machine learning healthcare applications, Am. J. Bioeth. 20 (11) (2020 Nov) 7–17.

[7] I.Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, M Ghassemi, Ethical machine learning in healthcare, Annu. Rev. Biomed. Data Sci. 4 (1) (2021 Jul 20) 123–144.

[8] R.M. Williams, S. Smarr, D. Prioleau, J.E. Gilbert, Oh No, Not Another Trolley! On the need for a co-liberative consciousness in CS pedagogy, IEEE Trans. Technol. Soc. 3 (1) (2022 Mar) 67–74.

[9] S. Fazelpour, D. Danks, Algorithmic bias: senses, sources, solutions, Philos. Compass [Internet] 16 (8) (2021 Aug).

[10] G.S. Nelson, Bias in artificial intelligence, N. C. Med. J. 80 (4) (2019 Jul) 220–222.

[11] A. Koene, L. Dowthwaite, S. Seth, in: IEEE P7003TM Standard for Algorithmic Bias Considerations, In: ACM/IEEE International Workshop on Software Fairness, 2018, pp. 38–41.

[12] J.K. Paulus, D.M. Kent, Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities, Npj Digit. Med. 3 (1) (2020 Dec) 99.

[13] M.W. Sjoding, R.P. Dickson, T.J. Iwashyna, S.E. Gay, T.S. Valley, Racial bias in pulse oximetry measurement, N. Engl. J. Med. 383 (25) (2020 Dec) 2479–2480.

[14] Rafael N. Ferreira, Nuno Ferrete Ribeiro, Cristina P. Santos, “Fall risk assessment using wearable sensors: a narrative review.”, Sensors 22 (3) (2022) 984, doi:10.3390/s22030984.

[15] Sara Usmani, Abdul Saboor, Muhammad Haris, Muneeb A. Khan, Heemin Park, “Latest research trends in fall detection and prevention using machine learning: a systematic review.”, Sensors 21 (15) (2021) 5134, doi:10.3390/s21155134.

[16] Kodithuwakku Arachchige, N.K. Sachini, Harish Chander, Adam C. Knight, Reuben F Burch V, Daniel W. Carruth, “Occupational falls: interventions for fall detection, prevention and safety promotion.”, Theor. Issues Ergonomics Sci. 22 (5) (2021) 603–618, doi:10.1080/1463922X.2020.1836528.

[17] K.A. Bartlett, K.E. Forth, C.S. Layne, S. Madansingh, Validating a low-cost, consumer force-measuring platform as an accessible alternative for measuring postural sway, J. Biomech. 90 (2019 Jun) 138–142.

[18] K.E. Forth, K.L. Wirfel, S.D. Adams, N.J. Rianon, E. Lieberman Aiden, S.I. Madansingh, A postural assessment utilizing machine learning prospectively identifies older adults at a high risk of falling, Front. Med. 7 (2020 Dec) 591517.