Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice

CECILY MORRISON, KIT HUCKVALE, BOB CORISH, RICHARD BANKS, MARTIN GRAYSON, JONAS DORN, ABIGAIL SELLEN, and SIÂN LINDLEY, Microsoft Research

Sophisticated ubiquitous sensing systems are being used to measure motor ability in clinical settings. Intended to augment clinical decision-making, the interpretability of the machine-learning measurements underneath becomes critical to their use. We explore how visualization can support the interpretability of machine-learning measures through the case of Assess MS, a system to support the clinical assessment of Multiple Sclerosis. A substantial design challenge is to make visible the algorithm's decision-making process in a way that allows clinicians to integrate the algorithm's result into their own decision process. To this end, we present a series of design iterations that probe the challenges in supporting interpretability in a real-world system. The key contribution of this article is to illustrate that simply making visible the algorithmic decision-making process is not helpful in supporting clinicians in their own decision-making process. It disregards that people and algorithms make decisions in different ways. Instead, we propose that visualisation can provide context to algorithmic decision-making, rendering observable a range of internal workings of the algorithm from data quality issues to the web of relationships generated in the machine-learning process.

CCS Concepts: • **Human-centered computing** → *Empirical studies in visualization*;

Additional Key Words and Phrases: Human-centred machine learning, visualization, health, in-the-wild study

ACM Reference format:
Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Siân Lindley. 2018. Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 12 (July 2018), 28 pages.
https://doi.org/10.1145/3181670

1 INTRODUCTION

The recent availability of ubiquitous sensing has made the opportunity to sense movement, and by extension measure motor ability, a realistic possibility in clinical settings. For example, systems are being built to assess standing balance in older adults using gyroscopic and force sensors [51], detect and respond to freezing gait *in situ* for Parkinson's patients using accelerometry [35], and analyse gait in multiple sclerosis patients [15]. As these systems become more sophisticated, they

The reviewing of this article was managed by special issue associate editors Marco Gillies and Rebecca Fiebrink.

Authors' addresses: C. Morrison, K. Huckvale, B. Corish, R. Banks, M. Grayson, A. Sellen, and S. Lindley, Microsoft Research Cambridge, 21 Station Road, Cambridge, CB1 2FB; emails: {cecilym, t-chhuck, rcorish, rbanks, mgrayson, asellen, sianl}@microsoft.com; J. Dorn, Novartis Pharma AG, Digital Development, Novartis Campus, WSJ-027.02.74, 4056 Basel Switzerland; email: Jonas.dorn@novartis.com.

https://doi.org/10.1145/3181670

are using machine-learning algorithms to interpret the sensed data and propose a measurement or action. Systems intended to augment clinical thinking, the interpretability of the machine-learning measurements becomes critical to their use.

We use the example of Assess MS to explore how we can enable clinicians to use the results of machine-learning measurements in their clinical decision-making. Assess MS is a depth-sensing computer vision system to support the clinical assessment of motor ability in Multiple Sclerosis (MS) [26, 39]. Using novel machine-learning algorithms, it aims to provide a more consistent and fine-grained measure of motor ability than currently possible through neurological examination. The system captures patients doing specified assessment movements with a Kinect under the supervision of a health professional. The depth data is then processed, and the machine-learning algorithm returns a classification of motor ability based on the well-known clinical scale, the Expanded Disability Status Scale (EDSS) [29].

A substantial design challenge is to make visible the algorithm's decision-making process in a way that allows clinicians to integrate the algorithm's result into their own decision process. Assuming that it is enough to utilise numbers that correlate with a known clinical framework, overlooks a large literature on information in medicine that illustrates the myriad of ways in which data is part of a larger negotiated diagnostic process that requires interpretation of measures provided [21]. Measurement and context are elided when a clinician carries out an examination, but when the measurement of motor ability is done by an algorithm and its interpretation by a clinician, an understanding of how the algorithm arrived at its measurement is needed to support this negotiated interpretation.

Interactive computer vision systems, as exemplified by Assess MS, pose particular challenges to the interpretability of machine-learning algorithms. While a growing corpus of work focuses on using visual analytics to support the interpretation of machine-learning results in decision-making contexts (e.g., Reference [50]), most of this work is inspired by networks, or complex data sets. They offer little guidance in terms of image data. In our case, we need to relate the algorithmic decision process to the *temporal body*. Indeed, previous research has highlighted that computer decision support on images of the body is very different than working with data points [18], offering few concrete starting points.

In this article, we present a series of design iterations in which we iteratively explore potential visualisations for expressing the predicted machine-learning measurement of motor ability provided by Assess MS for a clinical audience. The key finding of this article is to illustrate that simply making visible the algorithmic decision-making process is not helpful in supporting clinicians in their own decision-making process. It disregards that people and algorithms make decisions in different ways. Instead, we propose that visualisation can provide context to algorithmic decision-making, rendering observable a range of internal workings of the algorithm from data quality issues to the web of relationships generated in the machine-learning process.

The specific contributions that this article makes are:

- "Making visible" the algorithm was too naïve an approach to be useful for clinical purposes as people and machines "think" differently.
- Intelligibility may be achieved through providing context to algorithmic decision-making, such as data quality.
- We highlight a potential new role of curation, in which someone with hybrid computational and medical expertise is able to inspect a machine-learning model for accuracy, something a clinician is not skilled, or needs, to do.
- We demonstrate that machine-learning algorithms cannot be designed separately from the design of the application itself, questioning the "box" metaphor often associated with machine-learning algorithms.

2 RELATED LITERATURE

2.1 Machine Learning in Motor Assessment

The application of machine learning to MS is most commonly applied to the automated identification of MS lesions within panels of magnetic resonance images (MRI). This large literature includes examples that: prove the technical capabilities of specific machine-learning algorithms (e.g., Reference [14]); respond to articulated workflow problems, such as human performance variability (e.g., Reference [49]; or address a clearly articulated clinical problem, such as transition to full MS from the precursor syndrome (e.g., Reference [7]). There is also a small literature on the application of machine learning specifically to movement analysis in MS [1, 41]. The overarching focus of this literature is algorithmic development and validation; considerations of how such systems might be used in clinical practice are restricted to the framing of the technical problem.

The advent of ubiquitous sensing technologies has made the opportunity to measure motor ability in clinical practice a real possibility [45]. For example, gyroscope data has been used for machine-learning tremor classification in Parkinson's disease [9]. This study takes the same approach as Assess MS using labelled data based on a relevant clinical standardized rating scale. Computer vision has also been used to assess finger tapping in people with Parkinson's, by which features were computed that estimated speed, amplitude, rhythm and fatigue in tapping and then used to train an algorithm to predict a symptom severity score [24]. Most recently, depth-sensing has been used to develop gait-indices for MS patients that align to the clinical standardized rating scale in a clinic setting [15].

While all of these studies were carried out in clinics, there is no discussion how clinicians would use such a system in clinical practice. It seems assumed that by producing clinically meaningful scores no further explanation of how those scores were reached is needed. This notion is at odds with social science literature in this area as discussed below.

2.2 Interpreting Image Data in Medicine

Images of the body are now commonplace in medicine. They are a tool that allows clinicians to "see" inside the body and perceive what they cannot see with their own eyes. This augmented "seeing" can be used for a range of purposes, from diagnosing disease to guiding surgical procedures (see Reference [37] for overview). The useful information that may be obtained from medical images, however, is not self-evident [36]. It relies on skills that clinicians learn through extensive apprenticeship. Clinicians establish a professional vision that allows them to see pertinent information in an image [17]. A medical image is not an objective view of the body [21], but a foggy window onto the body whose contents must be interpreted within the frame of a larger clinical picture and technical limitations of the imaging method.

The most relevant study of machine learning to interpret image data is an ethnography of a mammography screening service carried out in the context of a clinical trial of a computer-aided detection tool for breast cancer. The tool was intended to support radiologists by using a machine-learning algorithm to draw attention to specific parts of the image through prompts. This attention cue was hoped to counteract the effects of variability in concentration and make the visual search pattern more systematic in a human reader. The intention of this support was to enable a switch from having two human readers of the mammography images to a computer-reader and a human-reader [18]. The study raised two salient issues: the use of contextual information in image interpretation and the ways readers take account of other reader's opinions.

The ethnography showed that mammography images are interpreted alongside other documents, such as previous images or the patient's record. The authors suggest that the decision of whether a patient has breast cancer is "achieved through the coherent marshalling of ensembles of evidence." They point out that readers not only are skilled in marshalling these ensembles, but they know how the information was produced. As a result, they can make decisions about how much they can rely on pieces of information. An algorithm, in contrast, does not have the availability or capability to integrate a wider set of data in an ad hoc manner. Nor can the algorithm's classification be utilised in such an ensemble, if the generation process of its output is not understood.

Even so, human readers felt the need to generate accounts of the computer's reading, and ensure their own thinking addressed the prompts offered. This could lead to elaborate (and potentially incorrect) assessments of the computer's reading, for example, suggesting ways that cancer might be seen in the image when not present. This behaviour, however, mimics the way that human readers would interact with their colleagues. They would develop a conversation around different opinions, and negotiate a final decision. This and the previous finding described highlight the importance of making visible the decision-process of machine-learning algorithms so that they can be incorporated into an interpretive process that relies heavily on colleagues and known information practices.

This study suggests that some level of interpretability of the machine-learning assessment of images is needed to allow clinicians: to utilise the classification results within a wider set of data and knowledge; and to enable explanatory accounts that support collaborative interpretation with other human colleagues. To usefully include sensor-based assessment in clinical practice, it will be essential to help a clinician understand the decision-making process of the algorithm in such a way that they can incorporate the result into their own thinking. It is not enough to provide a clinically meaningful classification. In the next section, we consider research on interpretability of machine learning in both clinical and non-clinical settings.

2.3 Interpretability of Machine-learning Algorithms

There is some work in the machine-learning literature that focuses explicitly on the need for intelligibility of classification of machine-learning models, particularly in healthcare, e.g., Reference [6]. This literature demonstrates that there is still a significant trade-off in performance between models that are considered intelligible (e.g., generalized additive models) and those that are considered most powerful (e.g., deep neural nets). This is in part because of the univariate nature of generalized additive models. Some research is closing this gap, but still unable to deal with large-dimensional data [32]. This trade-off is very pertinent to Assess MS as the small amounts of medical data being used require more powerful models to gain the calibre of results necessary for medicine at the potential expense of the interpretability of the model.

The human-computer interaction literature has focused on interpretability of machine learning mainly in the context of interactive machine learning [2]. Interactive machine learning focuses on supporting users to incorporate their own judgements into the model through an iterative training approach of providing training samples or labels to the model and seeing the resulting change in the model. This has been delineated in applications such as email classification [47], image search [13], and alarm triage [3]. These works are all predicated on the idea that transparency, or a "white box" model increases users' abilities to interact with systems based on machine-learning models [19].

Studies in this literature have addressed the problem of how best to provide explanations that achieve this "white box." Why or why not questions have been one of the key mechanisms explored for successfully providing feedback to users [28, 31]. Principles for these explanations are laid out in Reference [27]: Be iterative, be sound, be complete, but do not overwhelm. The authors suggest that information should be provided in way that allows people to gradually build their mental models, striking a balance of accuracy of explanation with the amount of information provide.

In a real-world study, Reference [25] found that explanation fostered trust when an algorithm violated expectation, but too much information eroded trust.

Work on the explanation of context-aware systems have expanded what might be included in an algorithmic explanation. Along with Why and Why not questions, Reference [30] adds What, What if, How to, Inputs, Outputs, and Certainty. The last three highlight aspects specific to sensor behaviour and include notions of certainty. What has been sensed? What has been computed? How certain are the computations? These general types of information to provide are a helpful starting point. They can be nuanced by findings that show that preference to how a system frames its answers is not uniform [47] and that errors may be judged based on the likelihood of human error in the same situation [12].

We can see how explanatory accounts play out in medicine more specifically by looking at clinical decision support systems for diagnosis. Reference [5] looks at the over and under reliance on such systems. They show that fuller explanations substantially increased trust (to the point of over-reliance), where measures of certainty did not. These authors also found that health professionals trusted the system most when it provided reasoning similar to their own. In this case, that was achieved through providing the data used for decision-making. Other participants wanted explanation to go beyond the data to the pathology. Confidence intervals, however, were hard to interpret without explicit guidance. Showing the confidence intervals of all possible predictions was more useful.

The HCI literature on algorithm interpretation is heavily focused on the laboratory environment. In this article, we want to examine how we might apply some of this learning to a real-world system. To date, computer decision support systems that use machine learning have struggled to gain purchase in practice because of unresolved questions about diagnostic accuracy, safety and feasibility of use in busy clinical settings [40]. Yet, examples such as prospective genetic fingermarking in cancer research, which use visualisation of very uncertain results to support clinical decision-making [16] show that machine learning can play a useful role in medicine if contextualized properly.

The work in this section is primarily textual or dealing with categories that are reasonably discreet. A challenge of the Assess MS is to apply these ideas to image data, and specifically image data of the body. Visualization is one way that this might be done.

2.4 Visualization

Visualization is one way to support users in exploring and interpreting machine-learning algorithms that they are working with. Mane et al. [34] provide a key example in the health domain, proposing a visualisation that aggregates a patient's history with predicted future path for different medication scenarios. The prominent design feature of this system was to provide predictions that could be interpreted in the context of a patient's data, without offering an explicit "decision." The article suggests that by using visualisation rather than a numeric output, context can be provided in a way that allows the health professional to reason about the prediction. This example has guided us in the role visualization could usefully take in communicating the Assess MS results, but we need to look elsewhere to gain inspiration as to how we might apply this to concept to image data.

The HCI literature has explored how visualization can be used to convey classification boundaries to help users refine models about concepts that may not be discrete, such as the weather. In ManiMatrix, the classification boundaries are provided in square pie-chart type graphs as well as through color-coded confusion matrices [22]. EnsembleMatrix is similar for multiclass machine-learning algorithms, using heat map visualisations among others to reveal commonly confused classifications in the algorithm [48]. While the focus of this work is to enable interactive training, it does highlight the importance of visualising classification boundaries to spot outliers or near boundary decisions.

Other systems focus on facilitating the integration of domain expertise. G-PARE, a visual analytic tool for comparing two uncertain graphs, is a particularly good example. Each uncertain graph is produced by a machine-learning algorithm that outputs probabilities over node labels. It provides several different views that allow users to obtain a global overview of the algorithms output, as well as focused views that show subsets of nodes of interest. Users can follow cascades of misclassifications by comparing the algorithm's outcome with the ground truth [44]. This example draws attention to the importance of views of different granularity.

Elzen [10] focuses on how domain experts can explore and adjust decision trees directly. This author provides an overview of a range of visual examples of decision trees, characterising them as either node-link diagrams, icicle plots, or a combination of these two. The most relevant finding of this work is that current visualisations do not integrate the tree visualization (the structure of the decision tree) with the data visualization (the visualization of class distributions). This distinction in the role of visualisations may be unhelpful for clinical-decision making, which may need different perspectives on the measurements and at how they were arrived.

The literature does not provide direct examples of visualization of machine-learning algorithms that compute over images of the body. Descriptions of rehabilitation systems, however, offer some perspective on visualisation of sensed health data. In Reference [4], a knee rehabilitation system provides real-time visualization of knee flexion for each repetition with a summary bar chart of knee flexion across exercises and the change in flexion over time. In Reference [11], new graph visualisations are being developed for specific design foci, such as providing detailed step data to cue social support. Auditory approaches, representing movement with a wave of sound, have also been successfully tried to support the management of chronic pain [46]. None of these examples explicitly reference an image of the body.

3 ASSESS MS

3.1 Multiple Sclerosis

Multiple Sclerosis (MS) is a chronic inflammatory disease of the central nervous system, which causes a variety of symptoms, either in combination or alone. These may include numbness or paralysis, tremor, cognitive difficulties, vision-loss, and reduction in motor strength. While symptoms are very diverse, stereotypical symptoms include: intention tremor, the shaking of the hand when intending to touch a target (e.g., the nose); ataxia, the wobble of head or torso when balancing; and impaired walking requiring aids for balance and to address paralysis. The common focus on these symptoms when portraying MS reflects that motor ability loss, as opposed to sensory or cognitive ability loss, is one of the key non-invasive indicators of disease progression.

The disease course is most frequently characterised by relapses in which the affected person experiences neurological symptoms followed by extended periods of remission in which symptoms may improve. Over time, the disease can enter into a progressive phase in which a steady deterioration occurs. About 15% of MS patients have on-going deterioration from disease onset [20]. More tangibly, some patients can lose their lives from the disease within a period of years while others can live their entire lives affected only by minor sensory loss. The unpredictability of the disease course is challenging for patients and clinicians alike in making treatment decisions, making the ability to track MS particularly useful.

The condition is currently assessed with a standardized rating instrument based on clinical examination, the EDSS [29]. Patients are asked to perform a range of functional exercises, including stretching out one arm to the side and then touching the nose (Finger Nose Test) or walking on a

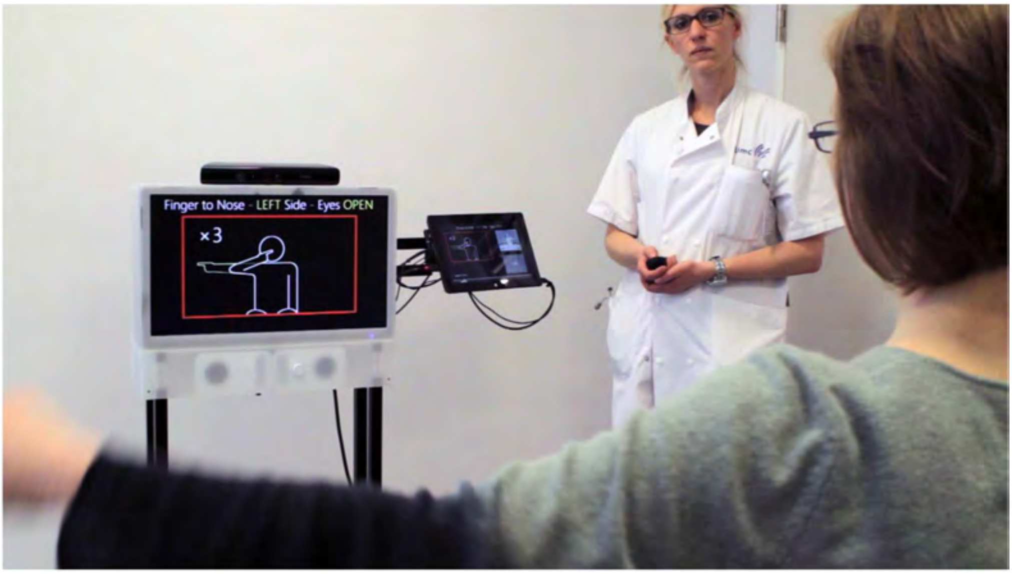


Fig. 1. The Assess MS prototype used by a health professional to capture depth and RGB videos of specific assessment movements performed by patients in a clinical setting with a Kinect.

pretend tight rope (Tightrope Walking). These exercises cover seven functional groups from sensory abilities (e.g., numbness) to motor and cognitive abilities. Each exercise is given a sub-score, often on an ordinal scale from 0 (no disability) to 4 (no function); sub-scores are then summarized into Functional System scores and, together with the ability to walk, are combined to create an ordinal score, from EDSS 0 to EDSS 10.

Although the EDSS is a widely-used and accepted outcome measure, it suffers from low intra- and inter-rater reliability making disease tracking difficult [8].

3.2 Assess MS System

Assess MS is a system to support the clinical assessment of Multiple Sclerosis (MS) using depth-sensing computer vision. Its aim is to provide a consistent quantified metric of motor ability for patients with MS to enable more reliable tracking of disease progression than currently possible. The system has four parts: (1) a prototype device for capturing depth and RGB videos of specific assessment movements performed by patients; (2) an interface for clinicians to label those patient videos with the relevant ordinal EDSS sub-score as well as rank patients; (3) a supervised machine-learning pipeline used to classify the severity of motor dysfunction, predicting the EDSS ordinal sub-score for each assessment movement; and, finally, (4) a visualization interface to present results to clinicians, the centrepiece of this article.

A prototype device, shown in Figure 1, is used by a health professional to capture depth and RGB videos of specific assessment movements performed by patients in a clinical setting with a Kinect; see Reference [38]. The original movement protocol focuses on a set of 11 movements covering upper body, trunk, and lower body motor ability thought to be clinically indicative of disease progression. The visualizations in this article draw mainly from the EDSS Finger Nose test. In this movement, the patient stretches their hand out to the side and touches their nose three times before placing the hand in their lap. This is done on both the left and right and with eyes open and closed. Tests that show ataxia, the wobble seen when holding both arms out to the side or drawing squares in the air in front of the patient are also used.

The videos are then labelled by clinicians with the appropriate ordinal EDSS sub-score as well as ranked in relationship to each other to provide the training and test data for algorithm development and validation; see Reference [43]. For the purposes of the visualizations presented in this article, a supervised machine-learning approach has been taken to predict the clinician-provided EDSS sub-score labels from 0 to 4, providing integer classifications for each movement. While algorithmic development remains on-going, an early version of the algorithm has been published in Reference [26]. Details critical for understanding the visualizations are presented below.

Customized randomized Forests and novel ensembles of randomized Support Vector Machines are used to discriminate landmarks in the depth videos that contribute to the classification. Unlike other applications that assess motor ability [33], the skeleton provided by the Kinect SDK is not utilised. Rather data reduction is achieved through the calculation of optical flow throughout the video. The choice of decision trees as opposed to other methods was made ostensibly with the ability to explain the results. In theory, each evaluative step in the branching decision tree can be examined, and a descriptive account given of the decision-making process.

The landmarks chosen are random spatial-temporal cubes in the depth video that contain the optical flow information. In other words, the "features" utilised in this case are numeric values that approximate (through a series of abstractions) the number of changes in direction of either X- or Y-oriented movement within a three-dimensional cube of space-time. Classification is performed by in a series of branching rules, each of which compares either a single feature, or the result of an arithmetic function involving two features, to some arbitrary numeric threshold. The result of this calculation determines either the next rule that will be applied or, in the final step, assigns a categorical label. The features, evaluation rules and their ordering are all learned from data during a process of training.

The Assess MS prototype is currently being used to collect depth data for training and validation of the machine-learning algorithms being developed. The Assess MS system has not been deployed as a predictive tool in clinical consultations. As such, the visualizations proposed here are in preparation for planned future use in clinical consultations.

4 VERIFYING ALGORITHMIC MODELS

Our first design exploration was to present a set of visualisations developed for verifying the machine-learning algorithm to our clinical colleagues. While never intended for clinical consumption, it provided a starting point to what aspects of visualisation may or may not be useful. Shown in Figure 2, it consisted of a heat map of the spatial-temporal cube features used by the algorithm in its decision process. The yellow boxes are those cube features most important in the decision-tree, fading out to red as least important. The spatial-temporal cubes are aggregated across patients and flatten in time onto a single image. A random frame from a single patient video was used as the background to show how a body may relate to the spatial temporal cubes. The approach is similar to the heat maps used in Reference [48], differing in that they are imposed on the body rather than confusion matrix.

This visualisation was shown to five of our clinical collaborators at a two-day quarterly team meeting. This set of visualisations served its intended purpose in ensuring that the computer was making decisions based on areas of the image that would be expected, such as the nose area in the Finger Nose test. It also showed some more surprising results, such as emphasising trunk movement in the "Drawing Squares" movement rather than hand movement as expected by the clinicians. In retrospect, this too was understandable, as there was too much unintended variation in the hand movement to be a substantial feature, but the ataxic symptoms of the trunk (wobble) could still be noted. Most importantly, it could be verified that the machine was not relying upon unexpected features, such as height.

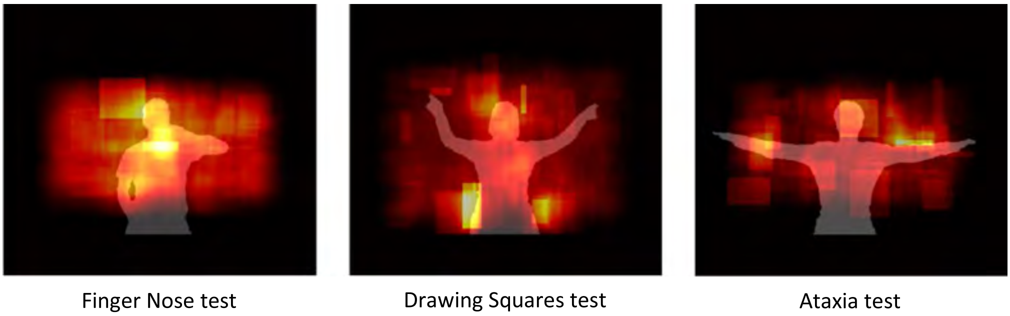


Fig. 2. Heat map of the spatial-temporal cube features used by the algorithm in its decision-making. The yellow boxes are those cube features most important in the algorithmic decision-making fading out to light red as least important.

The clinicians found it very difficult to relate to this set of visualisations even after multiple explanations. Two distinct issues arose, the relationship of: (1) data to the body; and (2) data to time. The relationship between the spatial-temporal cubes and the body is not obvious in this visualisation probe. Many of the cubes are off the body, such as next to the waist. From an algorithmic point of view this makes sense; these boxes would be expected to contain movement when the person had ataxia symptoms in comparison to no movement when the person was healthy. The interior of the body is less likely to show differences between patients and healthy volunteers. However, when a health professional assesses a patient, they look at the body and not where it might go in space. This is a substantial disconnect that made it difficult for the health professionals to link the algorithm's decision process to their own.

The few representations that were on-body, such as the nose area in the Finger Nose Test, did not provide enough granularity to add to the doctor's general knowledge. In this case, for example, every doctor knows that symptoms will appear around the nose. The clinicians asked many questions trying to understand how the visualisation might further be inspected to give them a view into the specific characteristics of those symptoms. One clinician said, "Does the height of the box represent the amplitude of the tremor?" Another asked, "Would the specific configurations of the boxes be different for each patient so that I could learn to distinguish patterns?" These questions demonstrated that the clinicians were looking for something that they could interpret, but this visualisation did not provide it.

A further challenge for clinicians in interpreting this visualisation was the cohort view represented on the image of a single image. While this is important from a machine-learning point of view to test the validity of the model generated, it cannot provide the individual differences that clinicians look at when assessing a patient in person. The confusion generated by overlaying the visualisation onto a single patient made apparent this disconnect. Clinicians attempted to relate the boxes to that patient's body and could not think about a whole cohort at once. That said, three-dimensional scatter plots used earlier in the project to represent cohort data were well received, as the clinicians used the distance between patients to understand their relationship to a cohort.

The representation of time, or rather lack of it, was also problematic. Not only did clinicians want to relate the boxes to the body, but to the body in time. For example, one clinician asked "I would not expect to see a box at the elbow. Is that at the beginning or the end?" Another clinician said, "Does the strength of the colour represent the speed of the movement?" Time, however, can be challenging to represent, particularly in the cohort view applied here. The presentation of the visualisation to our clinical team members, while never intending to be for their clinical consumption, was nonetheless, very informative. It highlighted the *relationship of data to the body*

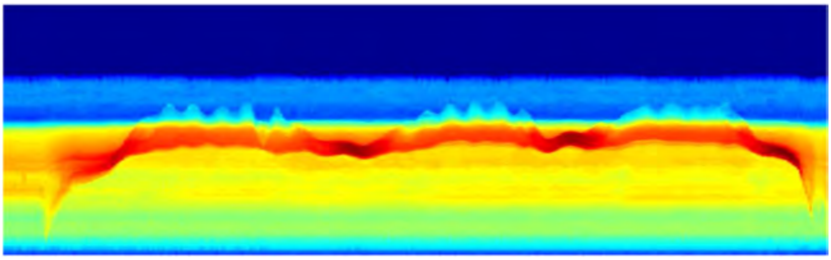


Fig. 3. An annotated version of the axis-aligned projection representation presented in Figure 3. Clinicians could identify three repetitions of the finger nose test demarcated by the grey boxes and the tremor demarcated by the black boxes.

and *time* as important aspects of any visualisation that we would need to provide to help clinicians integrate algorithmic decision-making into their own thinking.

5 VISUALISING ALGORITHMIC DECISION-MAKING

Our second design exploration focused on making the algorithmic decision-making process visible to the clinician. Building upon what we had learned in showing clinicians visualisations for machine-learning verification, we focused on notions of representing time and body. Our initial design approach was to overlay algorithmic features onto a visualisation that the health professionals found compelling. The work presented in this section builds up this picture through discussing: visualisation choice, exploring temporal representation, and exploring bodily representation.

5.1 Visualisation Choice

To choose an appropriate base representation of the patient video, we informally interviewed nine neurologists and asked them to discuss three potential visualisations. Shown in Figure 3, these abstracted time and the body in different ways. These are mapped onto the horizontal and vertical axes, respectively, in Figure 3.

- **Visualisation 1** is a video (shown as a static image here) that shows the outline of the person's body with the heatmap of the spatial-temporal cubes (as in the visualisation probe) that cross the body outline through time. This view addresses the relationship of data to body and time, but does not attempt to abstract either.
- **Visualisation 2** uses an axis-aligned projection technique that captures movement in one dimension only, encoding time and movement into a single image. Each frame of a depth video is reduced to a 1 by n pixel image, where n represents the height of the image, and the intensity of each pixel summarises the largest distance in a particular row in the image. This visualisation provides an abstraction away from the body, but elements of the movement are still visible in the wave form.
- **Visualisation 3** provides a graph of the produced machine-learning classification that relates cerebellar dysfunction and pyramidal dysfunction, two neurological systems. Time of an individual movement is not shown, but change over time is shown through multiple points on the graph. This visualisation abstracts the body and takes a view of time separate from the movements themselves.

Time: The clinicians preferred being able to overview an entire video in a single image, as in Visualisation 2. It made comparison easier, whereas video had to be held in memory to do the same. Their interest was also piqued by the idea of tracking a patient over time and relating two

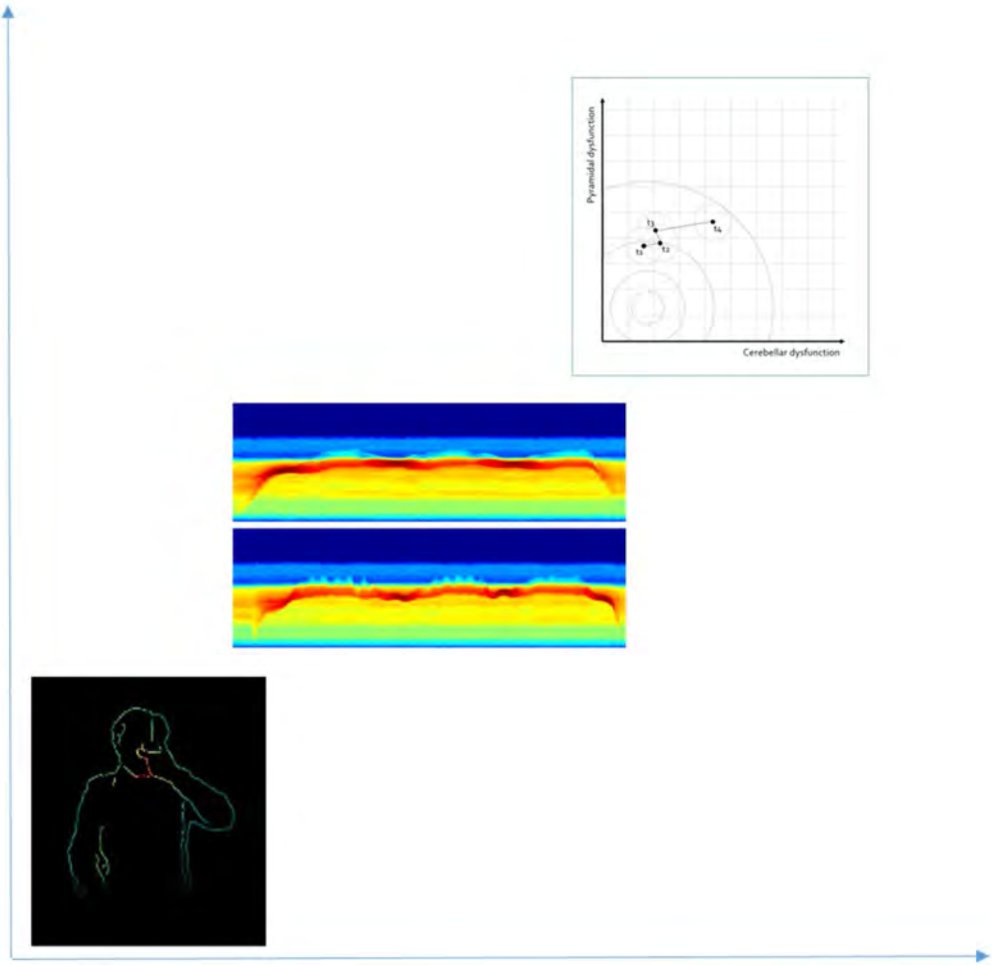


Fig. 4. Visualizations used to elicit discussion with health professionals. They abstract time (x-axis) and the body (y-axis) in different ways: Visualisation 1 (lower left), Visualisation 2 (middle), Visualisation 3 (upper right). Interview findings suggested.

neurological systems together. While the clinicians found this quite novel, they found it hard to imagine how this potential new way of looking at MS would be of benefit.

Body: The clinicians felt no need to see the body in the videos. If anything, they felt this unhelpful as it did not augment their existing view of the patient beyond the clinical examination. Visualisation 2, however, was of interest, because it highlighted specific aspects of the movement (see Figure 4 for annotated visualization). For example, the clinicians could see the three repetitions of the arm going in and out in the Finger Nose test as demarcated by the three grey contiguous boxes over the wavy red section of the image. They could also see the tremor demarcated by the black boxes around the spikey light blue waves. This alternative form of seeing the tremor, even if without a measurement of amplitude, provided a starting point to pattern-match and compare to other patients. The clinicians did not want to entirely abstract away from the body into a graph, as they were more comfortable with images that related to clinical decision-making as they currently knew it.

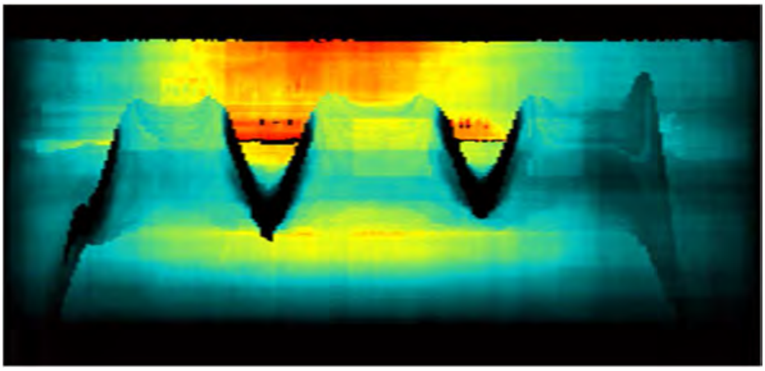


Fig. 5. Feature-overlaid axis-aligned projections of a patient performing three repetitions of a Finger Nose movement. It uses the heat-map of features used in the machine decision-making (described in Figure 2) in a temporal representation.

The findings of the interviews led us to conclude that we should utilise Visualisation 2 as the starting point for portraying the algorithm's decision process. It worked well as an intermediate image that neither portrayed the body, nor abstracted it to a data point. The ability to see time in a single image was also critical for clinicians to develop the ability to compare visualisations among patients. The aim in the exploration presented in the next section was to find a way to map the features in the machine-learning decision-making process onto this visualisation.

5.2 Exploring Time

Figure 5 illustrates the axis-aligned projected image overlaid with a heat map of the features most important in the machine classification for a single patient. It is similar to Figure 2 but with the incorporation of time provided by the axis-aligned projection. Three horizontal bands of feature sampling, smeared over the middle portion of the movement are seen. These correspond to the upper head, lower face/shoulder, and elbow regions during arm flexion. Of these, only the lower face fits easily into the clinical conceptual explanation–since it is here that a finger would approach during the flexion portion of a Finger Nose test. It is unclear, for example, why the top of the head should be important in the machine classification. Unfortunately, the temporal alignment of features is not clearly associated with different phases of movement making a temporal representation not useful.

The result is that neither the relationship of machine-learning model to body or time is elucidated in this first attempt. The mapping of features onto this visualisation was much more challenging than we expected. While this is frustrating from the point of view of the original purpose of the visualisation to explain to a clinical audience how the classification is proceeding, it has proven useful in uncovering how algorithm design choices might affect the scope for visualisation. Although feature space-time cubes were dimensioned stochastically, bounds were placed on each dimension. These were both spatial (a roughly 100 pixel border around the edge of the image was excluded from sampling) and temporal (the temporal size of a cube could be no less than 20% and no greater than 40% of the total movement duration).

These heuristic choices made by the algorithm designers ended up having a number of context-specific consequences. First, certain parts of the movement that a human might consider important for decision-making are not included at all. The defined dark regions visible in the arm region in Figure 5 exist because the hand extends spatially into the arbitrary margin from which no sampling took place. Second, the minimum temporal bound tends to prevent features being isolatable by

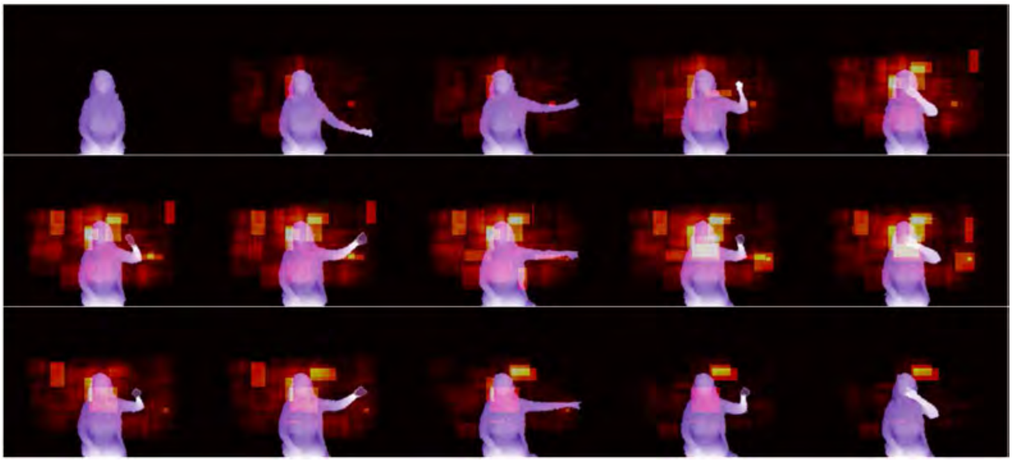


Fig. 6. Heat map of spatial-temporal cube features (similar to Figure 2) represented over time through providing multiple images. Time is read left to right and then top to bottom. Each row is a repetition of the Finger Nose movement. This representation highlighted that not all aspects of the machine decision-making are relevant or in line with clinical decision-making.

specific phases of movement in a way that would assist clinical explication, for example arm flexion (where intention tremor would be most marked) versus extension. This is because 20% of the video duration actually approximates the duration of a complete cycle of movement. Third, that the maximum duration (40% of video duration) encompasses up to two cycles might also tend to mask patterns of sampling (e.g., the same area sampled in each repetition).

The representation produced in Figure 5 highlights the possibility that there is a minimal penalty for features that smear across multiple movements—this is an emergent consequence of the nature of the movement (cyclical, mirrored) and the choice of representation that just sums acceleration changes over time. Corroborating this, we compared the frequency distribution of random possible feature durations that were generated at the start of training and on which all features were based with the frequency distribution of those features that were actually incorporated into trees. If duration was important for classification, given the representation, then you might expect the frequency distribution to be altered to enrich selection of a certain duration. There was no such effect, implying that temporality as understood in clinical terms in not incorporated into classification.

5.3 Exploring the Body

As it was difficult to see anything about the body in the axis-aligned projection representation once features were mapped, we returned to a body image view initially used by the machine-learning team but added time through providing multiple images and only showing a single person's data. Shown in Figure 6, each image, reading from left to right and then top to bottom, provides a heat map of important features (similar to Figure 2) of equal time slices in the video. This visualisation raises several issues. Like the verification visualisations, many of the boxes are not on the body. Indeed, they are on both sides of the body, even though the movement is only performed on one side. This is an artefact of the training process, which uses Finger Nose tests on both left and right sides to increase the amount of data.

This example shows that not all of the decision-making aspects of the machine-learning algorithm are relevant to the clinician. Some of the boxes are about the algorithm figuring out which side the movement is on, rather than about the movement itself. We also see that the third repetition is not being used for classification. This most likely stems from the fact that patient videos are not well aligned by the third repetition due to variation in speed, making them less useful for classification from a machine point of view. However, this is certainly not the case from a person's perspective. That the algorithm makes decisions that are artefacts of the data, but not relevant to clinicians, poses a challenge for visualization, which needs to remove these artefacts before clinical presentation.

A further issue stems from sampling visible at the top of the head, which is unexpected from a clinical point of view. One possibility is that this is a confounding feature based on head tremor, which is associated with multiple sclerosis but which, according to the rational clinical account, should not play a role in assessing the performance of a Finger Nose test. Both head tremor and other forms of motor performance (e.g., intention tremor) would reasonably be expected to be covariant with disease progression. Uncovering this kind of issue is important to be able to make claims about the construct validity of the underlying test method and to guard against unexpected performance in patients with non-standard disease presentations (e.g., minimal head tremor but substantial upper limb-related dysfunction) who might be poorly represented in training data.

6 CONTEXTUALIZING ALGORITHMIC DECISION-MAKING

Our third design exploration created a visualisation application based on our learnings from the previous two iterations. We learned from the second iteration that the decision-making process of the algorithm and the clinician can be quite different. We showed that some of that decision-making process was irrelevant (e.g., deciding whether a movement was left side or right), and other bits not well associated to movement (e.g., long features). *This demonstrated that supporting a clinician's decision-making using the Assess MS results by simply making visible the algorithmic decision-making is unhelpful.* In this section, we consider ways to contextualise the algorithm classification for clinical use as an alternative mechanism to support interpretability.

We chose to take an approach articulated as meta-models of machine cognition [42]. The argument here, proposed in the context of developing systems for end-user programming, is that most aspects of the machine's model will be unintelligible, even if relevant. However, there are certain aspects of the model that will be particularly important to a person interacting with it. It is these aspects of the model that we must make visible rather than concentrating on the inner workings of the algorithm. While not an exhaustive list, the author highlights the following.

- Confidence: How sure is the model that a given output is correct?
- Command: How well does the model know the domain?
- Complexity: Did the model do a simple or complex thing to arrive at the output?

While the Assess MS application is not directly interactive in the same way, the notion of a meta-model that provides visibility into key aspects of the model that can support clinical-decision making, in contrast to attempting to communicate the process itself, is something that we decided to instantiate. We did this in two steps. First, we held a workshop with our clinical partners to understand the information that they might potentially use in their decision-making. Second, we designed and built a visualisation application that portrays aspects of context we thought would be most useful.

6.1 Workshop

6.1.1 Method. We utilised the user-centred design technique, Design Studio (Kelani), to encourage our clinical collaborators to rapidly sketch their ideas as a way to discuss and clarify requirements. Our 12 participants used a set of visuals that capture different views of the body and time to "sketch" out their ideal interface for receiving results from Assess MS. In groups of four, individuals first created their own sketch, followed by the negotiation of a small group "sketch." To facilitate discussion, the groups were divided by seniority and were provided with the same clinical scenario, which included the examination being done by a nurse and the results interpreted by a doctor. Each group had a note-taker who recorded the session. Our focus was as much on the discussion generated as the artefacts produced in the hour long session.

The visual elements, shown in Figure 7, were chosen to convey distinct visualisation choices. The first visualisation set is an RGB video of the patient performing the movement or two videos provided side-by-side to enable comparison between different time points. The second set show abstracted movement patterns of a patient, portrayed both on and off the body image. In the third, two different graph representations are provided that show change over time. The final set are why/why not explanations of the machine-learning result in both a textual and visual form. We also gave our participants the choice of designing for a tablet screen or a large monitor to see the role of size and portability in their choices.

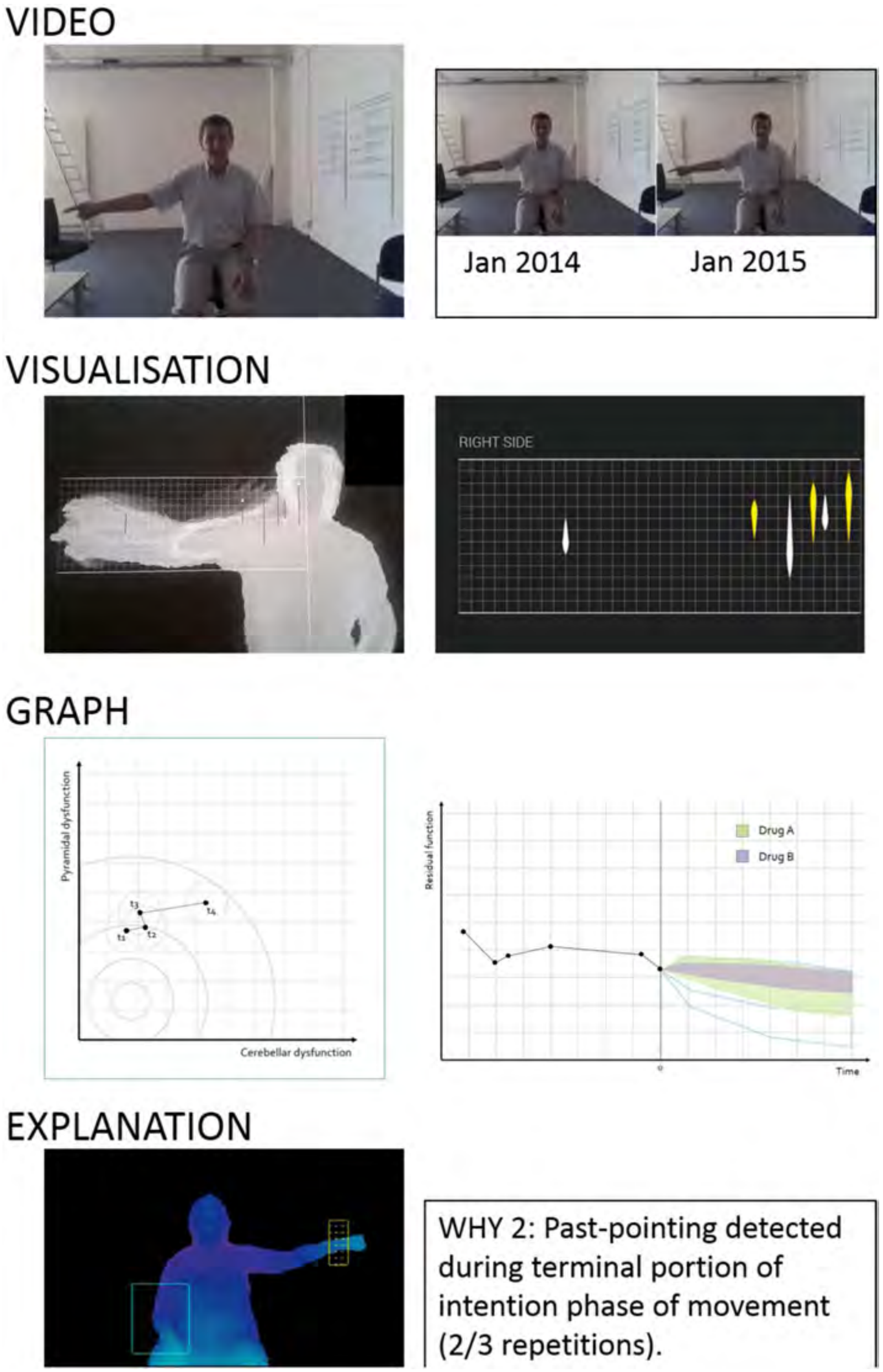
Each representation was chosen for a specific reason. Video interpretation relies on clinicians' existing skills. Abstractions of movement enable comparison and pattern-matching. Graphs can capture the statistical relationships created between patients while doing the machine learning, offering insight not previously available. Last, we provided why/why not explanations as one of the key approaches used in other domains to explain machine-learning algorithms to users. We chose both textual and visual explanations as we have examples of textual approaches in other domains, but our problem domain relies on the visual.

The "sketches" created by the participants were labelled with their names and collated into the working groups and then collected. Analysis of the workshop consisted of two parts. The choices of the visual elements on each sketch were tabulated. Notes were made on position and on seniority of clinician. These results were then contextualized with commentary recorded during the workshop sessions. The lead researcher worked with the note-takers to understand and discuss key themes in each group. Findings were presented to and discussed among the project team.

6.1.2 Findings. Video of the patient remained the most salient visual tool for the clinicians and appeared in all sketches. In part this was a matter of trust: "We need to have the raw data, because we [the doctors] are better than the computer." Trust could also be akin to understanding, as the system produces finer-grained measures in future: "If they have a 2.5, then how does it look in reality… The video increases confidence in what it means." But there were also real concerns around how to best integrate clinical knowledge with the machine: "Just to see. Is there something you know about the patient that changes the way you see?" The clinicians particularly liked the comparative videos that allowed them to see the patient at multiple time points as it augmented rather than ignored their own abilities.

Other visual elements also were employed to support sense-making, including abstract representation of movement and why/why not explanations. Most participants had at least one of these, but these did not figure prominently in the group sketches. One senior clinician argued that: *"When I clearly see a difference, and it says no difference, I want to know why not."* More junior doctors were content with a number and did not want any expression of uncertainty. While there was a general sense that clinicians wanted to know what the algorithm was doing, this had lower priority than elements that they could already easily relate to, such as the videos. It is not clear whether the abstract representations were not important enough to make it into the final sketches or whether they are just too far from current clinical practice.

Graphing change over time was also key for many of the health professionals. While there was no agreement on the exact representation or even what such a graph would tell them, there was a clear desire to gain a prognostic view of the patient in relationship to the patient cohort. The clinicians spent a lot of time debating how uncertainty of a machine-learning algorithm should be



WHY 2: Past-pointing detected during terminal portion of intention phase of movement (2/3 repetitions).

Fig. 7. Distinct visualization choices used during the Design Studio workshop with the Assess MS team.

presented in a temporal view. Junior clinicians did not want an uncertainty portrayed. They said: *"It's too much information." "In the end, we just want to have a number."* Senior clinicians were more open to seeing levels of algorithmic certainty. One said: *"I would like to have a graph of the values at different time points and their variability... The most interesting thing would not be the interpretation, but the facts over time so that I could write my interpretation."*

All clinicians agreed that any temporal view, such as a graph, should include treatment decisions, something they have been accustomed to with the use of digital clinical information systems.

There was a clear preference for video, over all of the other visual elements presented. Video enables the health professionals to use their highly tuned professional vision, extending it to include temporal change and higher granularity. Yet, there seemed to be an openness to options that give them alternative information about the patient over time and in a cohort as well as potentially more abstract views on the movement. The unfamiliarity with these visual methods made it difficult for them to more active participate in their design. This suggested that visualizations created by designers are needed to be used in every day practice of the clinicians for some period before a more solid view could be taken on how exactly what information would augment their decision-making process.

6.2 Application Design

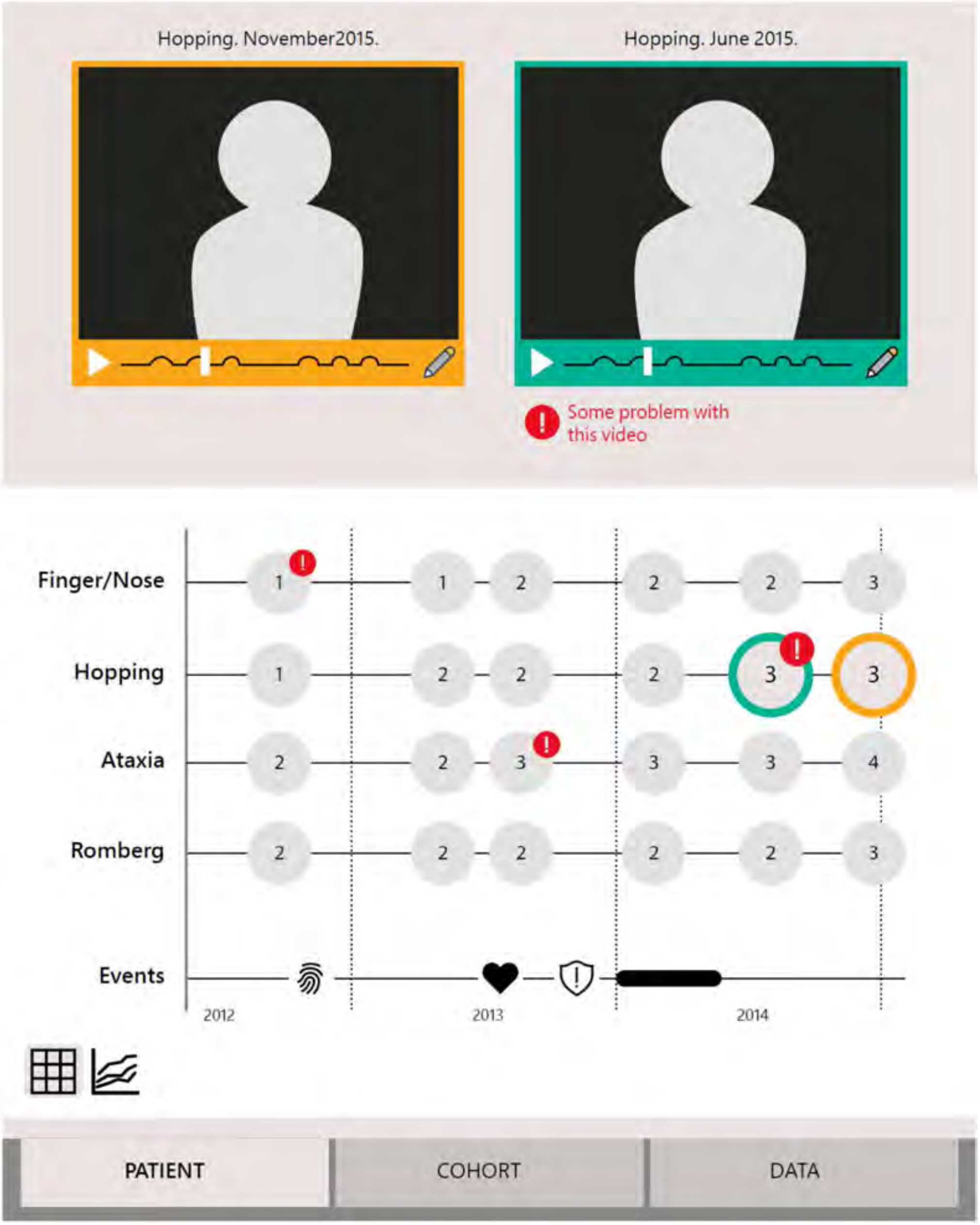
Our final step was to build an application for visualising results from Assess MS as an exploration of how we embody a meta-model of the machine learning alongside the preferences clinicians had expressed in regard to the visualization options presented to them. It explicitly considers which aspects of the machine-learning model need to be made visible to support clinical decision-making. We draw upon the three categories proposed in the original meta-models of cognition paper discussed above—*confidence, command, and complexity*—and we consider what those might look like in the Assess MS context. This was built as a Windows 10 application and contains three main screens:

Patient Overview, Algorithmic Interpretation, and Cohort Comparison. We describe each of these below. We finish the section with a theoretical evaluation done in conjunction with our clinical team as the application was used only with simulated data.

6.2.1 Patient Overview Screen. The patient overview screen, as shown in Figure 8(a), displays the predicted measurements given for each assessment movement over multiple time points. The movement is identified on the left followed by a string of scores in grey circles, which are spaced to indicate temporality of assessment. One can scroll back and forth as well as zoom in/out to change the temporal view. The last line on this graph includes critical events, such as a medication change or relapse, that align to particular measurements. An alternative graph view that highlights change of score in multiple tests is also available (see Figure 8(b)).

Clicking on any time point brings up the associated video and this can be compared to a video from any other time point. These are shown at the top of the screen. Underneath each video is a static abstract representation of the movement (the wavy line to the right of the video play button). It has been included to enable an at-a-glance comparison between movement profiles in the videos, something highlighted as critical to pattern-matching approaches utilised in clinical decision-making in earlier design iterations. By scrubbing through the video, one can track the movement seen in the video at the specific time point in the static representation indicated by the white bar to enable direct comparison between the two static representations.

The static representation would need to be developed specifically for each movement. In particular, it needs to include a specific visual language that provides temporal context to the static representation. For example, in the Finger Nose movement, lines could be inserted over the static

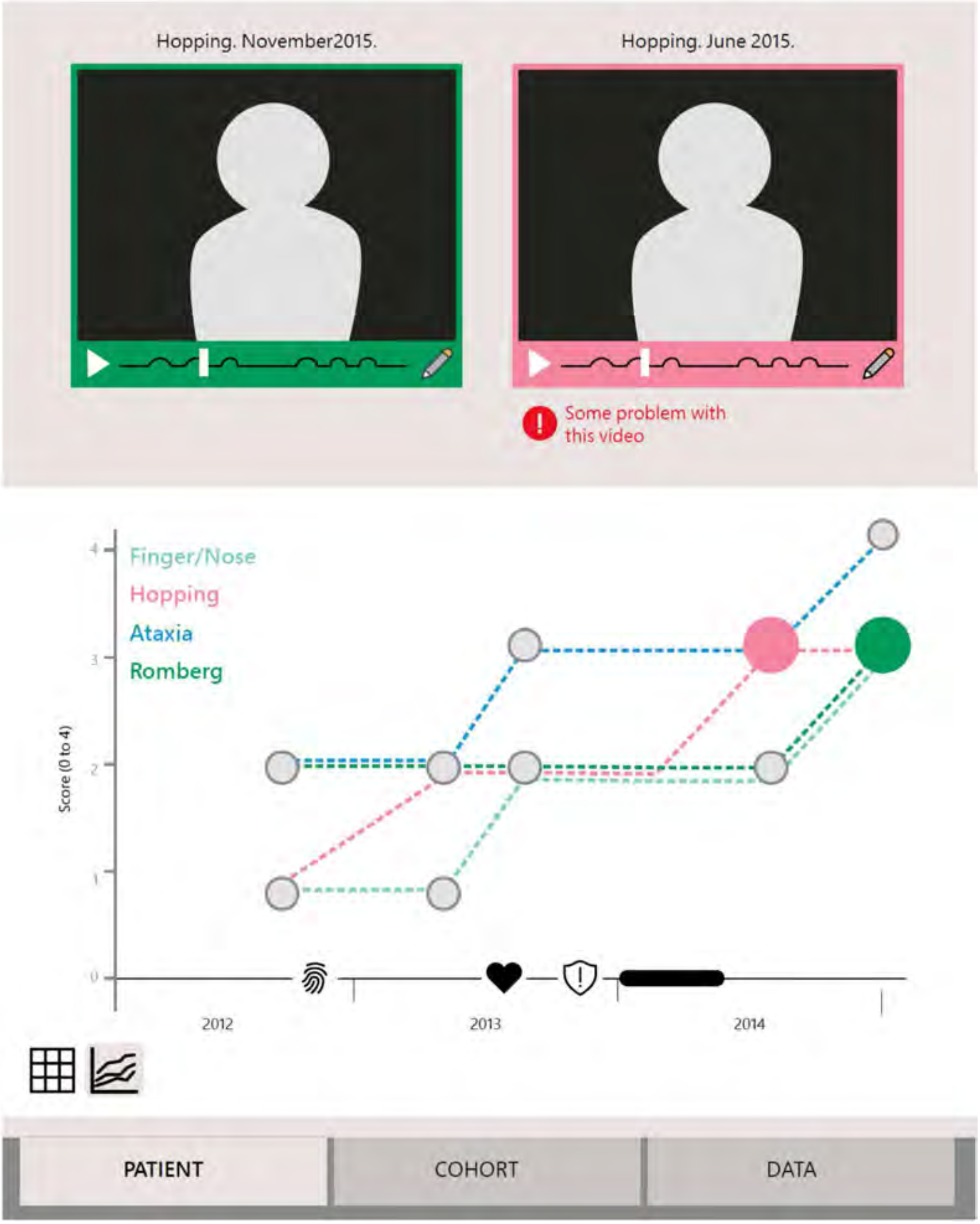


Some problem with this video

Finger/Nose	1	1	2	2	2	3
Hopping	1	2	2	2	3	3
Ataxia	2	2	3	3	3	4
Romberg	2	2	2	2	2	3

Fig. 8a. Overview screen visualization application for clinicians.

representation to demarcate points in the movement for which the arm is fully extended and when the nose is touched. This enables the clinician to see the temporal elements of tremor as the hand approaches the target for example, making comparison of these static representations possible even if not temporally aligned. With time, we would expect clinicians to learn the association between this representation and common patterns in patients, drawing on their honed skills as pattern matchers.



Some problem with this video

Fig. 8b. Alternate overview screen of visualization application for clinicians.

A key aspect of this screen is to indicate potential problems with data quality that suggest a clinician should use caution when relying upon the numeric result. This indicator is shown in the form of a red exclamation mark on the data point. Clicking on the video brings up a general description of the data quality issue. More details can be brought up by clicking on the exclamation point, which brings up a page that lists the specific problems. This page also includes a greater depth of information of how the score was reached. For example, it will contain the individual scores of each Finger Nose movement, which are assessed separately and then amalgamated. The data quality indicator is activated for three reasons: (1) difficulties pre-processing the data, (2) the comparative training data is small, (3) or the algorithmic confidence is low.

6.2.2 Algorithmic Interpretation Screen. The algorithmic interpretation screen places the patient within the training data, as shown in Figure 9. Placement on the x-axis is determined by EDSS sub-score in conjunction with the rank data of patients. Utilising the rank that contributes to the refinement of the EDSS sub-score (see Reference [43]), the screen is the ability to illustrate where a patient is in the spectrum for a given score. Are they just a 2 or nearly a 3? This is an indicator of change as a patient may have just shifted over the line from 1 to 2 or their level of motor ability may have changed dramatically with the same score change. As in the patient overview screen, videos of patients in the cohort can be compared.

The y-axis represents a similarity distance metric to other patient nodes in the training data. The way this distance metric is determined depends to some extend on the algorithm. The current Assess MS algorithm is a variation of a decision forest. In this case, we can calculate similarity between all nodes by looking at how many times patients end up in the same leaf node across the entire forest. This metric can provide a sense of why a particular score may have been given. For example, patients in the Finger Nose movement, may exhibit signs of dysmetria (cannot accurately touch the nose) or tremor (when approaching the nose). These receive the same score, but require different treatment options. They will appear as different clusters in this visualisation, even though they have gotten the same score.

There are some limitations to this screen that we discovered when trying to build it. As the training data is cross-sectional (static), it is only possible to compare how a patient progresses in relationship to the training data. It is not possible to compare how motor function changes in one patient versus another. It is also only possible to review the result of one test at a time. For this reason, there is a play button that allows one to track a patient through the training set over time for the particular movement chosen. That said, we feel that the approach of trying to exploit the relationships that form between data points during the machine-learning process has much to offer to the understanding of context. The next screen, cohort comparison, is a speculation on how to visualize this progression if longitudinal patient data were available.

6.2.3 Cohort Comparison Screen. The cohort comparison screen, as shown in Figure 10, assumes that at some point in future it will be possible to incorporate data from other patients and not just those who took part in forming the training database. This is particularly important as the training database has only cross-sectional data, reducing the opportunities of understanding temporality that could come with longitudinal data. On this screen, the trajectory of one of a patient's scores is shown in the context of other patient trajectories of the same score. At the top, instead of showing video, a graph of predicted measurements from all assessment movements of a particular patient (and a comparison patient) are shown. This provides a temporal view of change over time that video does not.

We might imagine that this screen enables a clinician to find patients with a similar trajectory and review medication or other critical events. This might need to be done over several pages to look at different scores. Indeed, they may find a whole cohort of similar patients, revealing interesting sub-types of the disease. While such possibilities require substantial new infrastructure for sharing medical data that address privacy concerns, it does help us imagine how the relationships established in a machine-learning system might fruitfully contribute to decision-making in ways other than just providing a consistent score. Such uses of machine learning are already at play in areas such as cancer genomic treatment.

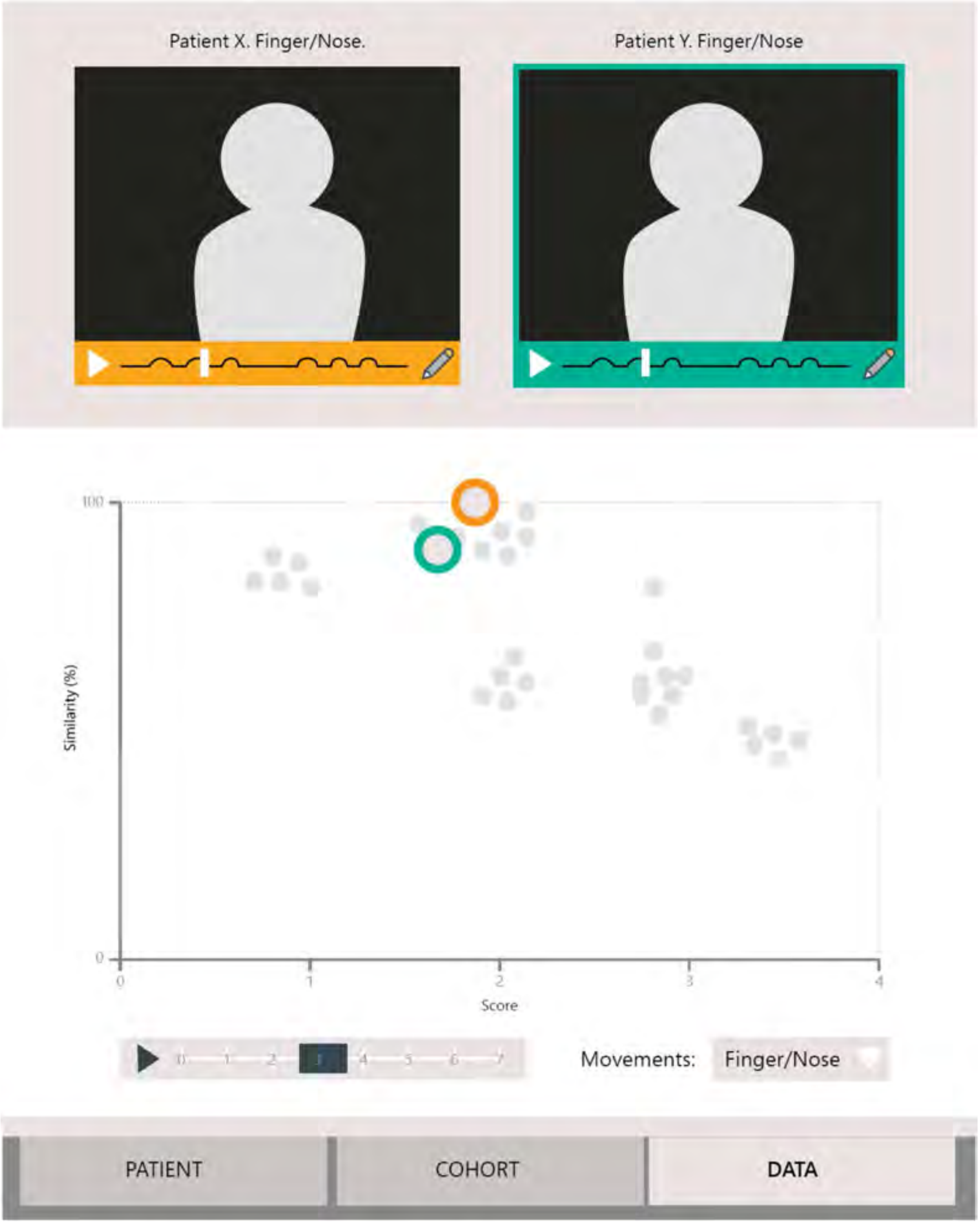
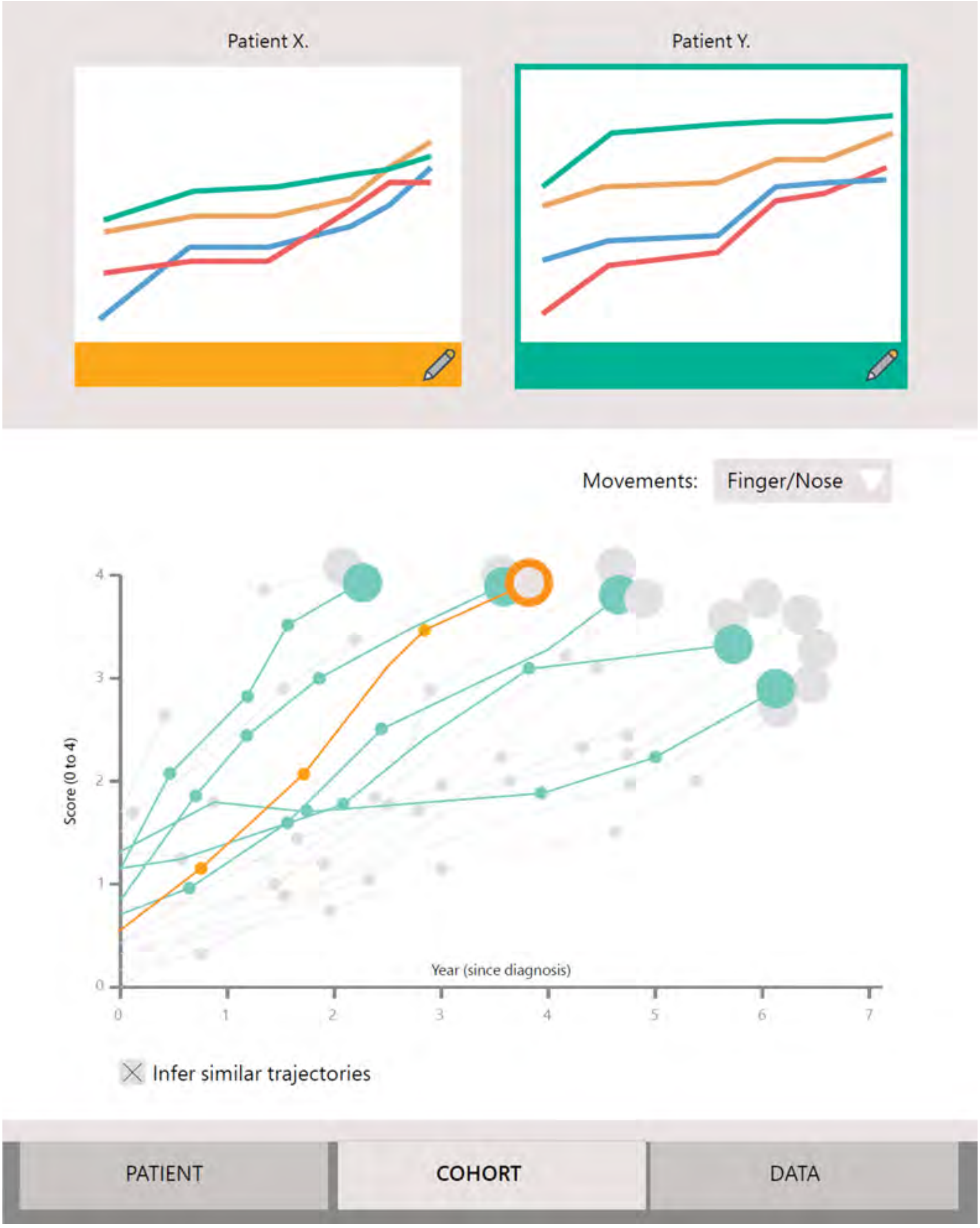


Fig. 9. Algorithmic interpretation screen of visualization application for clinicians.

6.3 Application Analysis

6.3.1 Design Analysis. The main aim of the design explorations embodied in this application design was to understand how we might make visible particularly relevant aspects of the machine-learning model without attempting to explain the entire model. We drew inspiration from the concept of meta-models of machine learning presented by Reference [42]. This article presents three questions that suggest different approaches to revealing the model, which are repeated here.



Movements: Finger/Nose

☒ Infer similar trajectories

Fig. 10. Cohort comparison screen of visualization application for clinicians.

- Confidence: How sure is the model that a given output is correct?
- Command: How well does the model know the domain?
- Complexity: Did the model do a simple or complex thing to arrive at the output?

In the following paragraphs we draw out how we achieved these in this application, highlighting the design lessons for visualising relevant aspects of a machine-learning model.

On the Patient Overview screen (Figure 8), issues with data quality are highlighted as a way to communicate *confidence* and *command*. Difficulties in pre-processing data are can lead to an incorrect or low confidence machine classification. Other issues include: the patient being unlike those seen in the training set; movements performed incorrectly by the patient; or potentially, the patient is too disabled to complete the movement as the system expects. Highlighting these indicate issues of *confidence* of the model. In places where training data may be sparse, as with those with high levels of disability, *command* could be low. We did not elicit questions from the third question: *complexity*. In all our cases, the complexity of decision-making was the same.

While an exclamation in a red circle may not seem like visualizing the algorithm, the example here suggests that in minimal design language we can make visible important aspects of the machine-learning model that are interpretable and potentially directly relevant to clinical decision-making.

On the Algorithmic Interpretation screen (Figure 9), confidence and command are visually represented in placing the patient in relationship to the training data set. How *confident* the classification of a predicted measurement of 1 can be interpreted by how close the circle is to the 1 mark or the 2 mark. Similarly, the number or tightness of a cluster indicates *command*. Interestingly, this visualisation is not technically revealing how the algorithm is working but does provide additional context for clinical interpretation. It makes tangible what is otherwise either hidden (the ranking as opposed to the classification of patients) or disguised in words (the training set is not well populated). We would argue that this visualisation provides a more tangible way to assess how the algorithm is working.

The Algorithmic Interpretation screen is perhaps not fully explained by the questions in the meta-model of cognition approach. It further embodies an approach of trying to exploit the relationships that form between data points during the machine-learning process. The Cohort Comparison screen (Figure 10) is an even stronger attempt at such. This design exploration has suggested another perspective on visibility; one in which we can provide another view of the patient through exploiting relationships that form in the machine processing of the data without necessarily needing to interpret them. We could think of visibility as adding a dimension to the ways that clinicians think about and explore patient data rather than explaining algorithmic decision-making.

6.3.2 Domain Expert Evaluation. We were keen to get feedback on whether the approach of only presenting relevant aspects of the model, in contrast to the approach taken in iterations 1 and 2 in which we attempted to elucidate the whole model, would be compelling to clinicians in their decision-making. Unfortunately, it was not possible to evaluate the application with patient data in a clinical setting as the Assess MS system has not yet been deployed as a complete system and won't be deployed until a clinical validation study is complete and appropriate FDA approvals have been sought. This is many years out. As such, we have considered how we might elicit useful feedback from our clinical team without them having the experience of such a system in practice.

The first step was to gain a domain expert evaluation by presenting the design back to the 12 people who participated in the workshop. This was done with data simulated from cross-sectional actor data to give a sense of what the application might do if longitudinal patient data was available. First, an introduction was done on a projected screen; then, in small groups, team members had an opportunity to try the app on a tablet computer; finally, one researcher facilitated a focus group. Each group was reminded of the patient scenario used in the first workshop and it was suggested that they walk through what data they might look at in this case. Our clinical colleagues found it immensely difficult to imagine what the flow of such a situation would be without having actual data. Instead, they focused on specific features and described how they would use these.

Everyone liked the video comparison, the feature that was also the most popular in the design workshop. The ability to see the past was a new opportunity to clinicians that made them feel that they could do a better job in their current assessment. Comparison is much easier than classification for a person. Those more senior were particularly taken with the idea of the cohort comparison screen and that they might be able to determine sub-types of the disease. There remained a scepticism among some that such patterns were actually in the data. While there was often talk of need to share the visualisations with patients and that their key role would be for the patient, no concrete examples were given as to what this might look like.

The facilitator brought people's attention to the specific design considerations that were intended to bring aspects of the model relevant to decision-making. First, it was asked what people thought the red exclamation points meant as nobody had commented on them. Dutifully explored, one clinician said, “It's good to realise that the machine can be wrong too.” Although through conversation our clinical colleagues came to understand why those indicators might be useful, they were happy to take the warning at face value and simply watch the video to make their own assessment. The focus on the videos was so strong that little emphasis was put on the predicted measurements at all, despite them being the original reason for the system.

The facilitator then oriented the discussion towards the Algorithmic Interpretation screen. The clinicians were immediately attracted to the ability to see whether a patient was in between a classification. They also liked the ability to play time as a video to see when big jumps occurred. As with the video, this offered them something that they could not do on their own. Some felt that the similarity metric was useful but would be more useful if it contained patients that they knew to give clues as to what treatments might be most beneficial. This suggestion further illustrates a strong desire of clinicians to augment their current skills through the data. One clinician captured this sentiment in a particularly revealing way, “If this application causes me to ask better questions of myself when I'm making a decision, then it has done a good job.”

This workshop was in many ways revealing, if unexpectedly so. It was very challenging for the clinicians, despite their involvement in the project, to imagine a future in which a machine-learning algorithm predicted movement ability. That said, it was clear that having features in the application that built on their current skills, such as video and more fine-grained scores (even if visual rather than actual classifications) gave the clinicians an entry point to the system that would enable them to become more familiar with the possibilities of what a machine-learning system might offer their clinical practice in future, such as understanding of where a patient fits within a cohort or sub-type.

We might expect from clinical reactions that it could take some time to fully appreciate why they may want to understand how the models are working. Perhaps a predicted movement classification that disagreed with their own might prompt further inquiry into the workings of the machine. What we did see is that the data reliability indicator was broadly understood and did not distract as some of the earlier design iterations did. We would suggest that this is a positive start. The visual nature of the Algorithmic Interpretation screen seemed to draw attention more readily than the data quality indicators, but potentially it communicated less discrete information.

Perhaps the most compelling insight that we had as researchers was to think differently about what design is trying to achieve in making visible relevant parts of the model. While we had given up the notion of making all aspects of the model visible, we were still very focused on making relevant aspects of the model visible within the larger context of other useful features in the application. Commentary from our clinical colleagues helped us reframe the potential role of design. They noted that encouraging clinicians to ask questions of the data, not just be able to interpret machine predictions was equally valuable. This fits nicely with some of the explorations that we did to utilise data produced in the machine-learning process (as opposed to the end result).

Visibility, while it certainly must contain relevant aspects of the model, can also be thought of more broadly as stimulating thinking through the artefacts of the machine process.

7 DISCUSSION

This article presents a series of design iterations that explore the best way to communicate visually sensed data for measuring motor ability for Multiple Sclerosis with Assess MS. The first iteration probes the relationship between the temporal body and data. The second iteration aims to make visible the algorithmic decision-making process. Found unhelpful, the third iteration offers a way to provide context to the algorithmic decision-making process instead without explicitly revealing algorithmic decision-making. These design explorations have tested a number of implicit assumptions that we, and we would expect many others, have about how we might productively support people to work in concert with applications that rely on machine learning. *In particular, we examine what it might mean to "make visible" an algorithm in the spirit of white box algorithms.*

The literature is clear that some level of interpretability of machine-learning algorithms benefit interactive applications. We responded to this by attempting to reveal the decision-making process. We showed that the algorithm made decisions that were irrelevant to users, such as whether a movement was being done with the left or the right hand. We also showed that how it made its decision depended on the data in ways that differed from how a person might make a decision. For example, in one instance the decision-making data happened mainly on the second of three repetitions of movements because, we suspect, video alignment is best at this point. A person would use the additive understanding acquired with each repetition. *This design exploration illustrated that just "making visible" the algorithm was too naïve an approach to be useful for clinical purposes as people and machines "think" differently.*

Our design explorations did suggest, however, that there are times when it is necessary to check precisely the algorithmic decision-process to determine the construct validity of a clinical measure. In the case of Assess MS, for example, we showed that the head was playing a key part in decisions about upper body dysmetria. While head tremor is frequently co-present with symptoms of disymetria, it is not necessarily indicative of it. *This finding raises the point that in some domains, health being one, that there may be a need for curator–a person with both computing and domain knowledge that can inspect the models produced by machine learning.* In cases like Assess MS, in which the features are not human interpretable, appropriate visualisations will be needed to support the curator role that differ from those used by clinicians.

That an algorithm could not just be "made visible," posed a challenge as to how visualisation could enable the use of algorithmic decisions in the clinical reasoning process. We decided to provide context instead. This was first instantiated as an indication of the data quality and second as the relationships between data points that revealed important aspects of the algorithmic decision. Our proposed interface, for example, showed that visualisation made obvious whether scores changes were a small jump over the boundary line or a big change. *We might say that the relationships embodied in the statistical processes of machine learning is the medium of machine "thought" that clinicians need access to, rather than the process of classification.*

Our interactions with our clinical colleagues continued to emphasise the importance of building on their existing visual pattern-matching skills, with a focus to augment clinician's current capabilities. This is best exemplified by the decision to maintain video in the final design iteration rather than replace it with a static, comparable visualisation. Instead, we linked the two to support the benefits of both. *We then focused on augmenting current clinical skill*s through providing videos from different time points. This uses the clinicians' skills for assessing movement, while providing them a resource not previously available. As systems like Assess MS move into widespread deployment, it will enable researchers to move from these ideas to evaluable statements.

Finally, our study addressed the “box” metaphor associated with machine-learning algorithms, which suggests that these algorithms are self-contained and can be inserted into applications without need of opening. Our design explorations illustrate that a machine-learning algorithm embedded in an interactive application is anything but a self-contained box. In the case of Assess MS, the features were chosen to gain the best level of machine accuracy, a substantial challenge when the movement differences are subtle and the level of noise in the data high. That said, these features, spatial-temporal cubes of pixels in a depth video, are not human interpretable, making visualisation highly problematic. So too did seemingly mundane choices, such as feature parameters, which lacking a penalty on length made it impossible to pinpoint aspects of the movement that showed disability. *This finding demonstrates that machine-learning algorithms cannot be designed separately from the design of the application itself.*

8 CONCLUSION

Ubiquitous sensing applications driven by machine-learning algorithms are becoming increasingly common in our world. This is particularly true in medicine, where there have been large numbers of explorations into their usage for sensing motor and cognitive ability. The now interactive nature of such applications means it is essential, particularly in domains such as medicine, that we find mechanisms to enable domain experts to assess the results of a machine-learning algorithm to integrate it into their wider decision-making process.

Addressing the interplay between human experience and machine-learning algorithm is tricky. Work has been done to try and create machine-learning algorithms that have results that are human interpretable [6]. Yet, it is difficult to get the full power of machine learning with such approaches, particularly given the advances in deep learning. Our work suggests that visualisation might help bridge this gap. In the case of Assess MS, we show that this cannot be done naively by making visible the algorithmic process itself as human and machine decision-making differ. Rather, a nuanced view of the context of the machine-learning result can be an alternative.

REFERENCES

[1] Murad Alaqtash, Thompson Sarkodie-Gyan, Huiying Yu, Olac Fuentes, Richard Brower, and Amr Abdelgawad. 2011. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. In *Proceedings of the 2011 Conference on Engineering in Medicine and Biology*. 453–457.

[2] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2011. Effective end-user interaction with machine learning. In *Proceedings of the 2011 Conference on Artificial Intelligence*. 1529–1532.

[3] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. 2011. CueT: Human-guided fast and accurate network alarm triage. In *Proceedings of the 2011 Conference on Human Factors in Computing Systems*. 157.

[4] Mobolaji Ayoade and Lynne Baillie. 2014. A novel knee rehabilitation system for the home. In *Proceedings of the 2014 Conference on Human Factors in Computing Systems*. 2521–2530.

[5] Adrian Bussone, Simone Stumpf, and Dympna O’Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics*. 160–169.

[6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare. In *Proceedings of the 2015 Conference on Knowledge Discovery and Data Mining*. 1721–1730.

[7] O. Ciccarelli, P. P. Kwok, V. Wottschel, D. Chard, M. L Stromillo, N. De Stefano, A. J. Thompson, D. H. Miller, and D. H. Alexander. 2012. Predicting clinical conversion to multiple sclerosis in patients with clinically isolated syndrome using machine learning techniques. *Multiple Sclerosis J.* 18 (Oct. 2012), 30–31.

[8] Jeffrey A. Cohen, Stephen C. Reingold, Chris H. Polman, and Jerry S. Wolinsky. 2012. Disability outcome measures in multiple sclerosis clinical trials: Current status and future prospects. *Lancet Neurol.* 11, 5 (May 2012), 467–476.

[9] N. D. Darnall, C. K. Donovan, S. Aktar, H.-Y. Tseng, P. Barthelmess, P. R. Cohen, and D. C. Lin. 2012. Application of machine learning and numerical analysis to classify tremor in patients affected with essential tremor or Parkinson’s disease. *Gerontechnology* 10, 4 (June 2012), 208–219.

[10] Stef van den Elzen and Jarke J. van Wijk. 2011. BaobabView: Interactive construction and analysis of decision trees. In *Proceedings of the 2011 Conference on Visual Analytics Science and Technology*. 151–160.

[11] Daniel A. Epstein, Alan Borning, and James Fogarty. 2013. Fine-grained sharing of sensed physical activity. In *Proceedings of the 2013 Conference on Pervasive and Ubiquitous Computing*. 489.

[12] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the 2011 Conference on Human Factors in Computing Systems*. 147. DOI:https://doi.org/10.1145/1978942.1978965

[13] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive concept learning in image search. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems*. 29.

[14] Ezequiel Geremia, Olivier Clatz, Bjoern H. Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage* 57, 2 (July 2011), 378–390.

[15] Farnood Gholami, Daria A. Trojan, József Kövecses, Wassim M. Haddad, and Behnood Gholami. 2016. A microsoft kinect-based point-of-care gait assessment framework for multiple sclerosis patients. *IEEE Journal of Biomedical and Health Informatics* 21, 5 (2016), 1376–1385.

[16] Benjamin A. Goldstein, Alan E. Hubbard, Adele Cutler, and Lisa F. Barcellos. 2010. An application of random forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics* 11 (Jan. 2010), 49.

[17] Charles Goodwin. 1994. Professional Vision. *Amer. Anthropol.* 96, 3 (Sept. 1994), 606–633. DOI:https://doi.org/10.1525/aa.1994.96.3.02a00100

[18] Mark Hartswood, Rob Procter, Mark Rouncefield, Roger Slack, James Soutter, and Alex Voss. 2003. "Reparing" the machine: A case study of the valuation of computer-aided detection tools in breast screening. In *Proceedings of the 2003 ECSCW Conference on Computer Support Cooperative Work.*

[19] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 CSCW Conference on Computer Supported Cooperative Work*. 241–250. DOI:https://doi.org/10.1145/358916.358995

[20] Christian P. Kamm, Bernard M. Uitdehaag, and Chris H. Polman. 2014. Multiple sclerosis: Current knowledge and future outlook. *Eur. Neurol.* 72, 3–4 (Jan. 2014), 132–141.

[21] Bonnie Kaplan. 1995. Objectification and negotiation in interpreting clinical images: Implications for computer-based patient records. *Artific. Intell. Med.* 7, 5 (Oct. 1995), 439–454.

[22] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. 2010. Interactive optimization for steering machine classification. In *Proceedings of the 2010 Conference on Human Factors in Computing Systems*. 1343–1352. DOI:https://doi.org/10.1145/1753326.1753529

[23] N. Kelani. Design Studio Method. Retrieved from http://www.bigspaceship.com/design-studio/.

[24] Taha Khan, Dag Nyholm, Jerker Westin, and Mark Dougherty. 2014. A computer vision framework for finger-tapping evaluation in parkinson's disease. *Artific. Intell. Med.* 60, 1 (Jan. 2014), 27–40.

[25] René F. Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems*. 2390–2395.

[26] Peter Kontschieder, Jonas F. Dorn, Cecily Morrison, Robert Corish, Darko Zikic, Abigail K. Sellen, Marcus D'Souza, Christian P. Kamm, Jessica Burggraaff, Prejas Tewarie, Thomas Vogel, Michela Azzarito, Peter Chin, Frank Dahlke, Chris H. Polman, Ludwig Kappos, Bernard Uitdehaag, and Antonio Criminisi. 2014. Quantifying progression of multiple sclerosis via classification of depth videos. In *Proceedings of the 2014 Conference on Medical Image Computing and Computer Assisted Intervention.*

[27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 2015 Conference on Intelligent User Interfaces*. 126–137.

[28] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Andrew Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Trans. Interact. Intell. Syst.* 1, 1 (Oct. 2011), 1–31.

[29] John F. Kurtzke. 1983. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology* 33, 11 (Nov. 1983), 1444.

[30] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 2009 Conference on Ubiquitous Computing*. 195–204.

[31] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 2009 Conference on Human Factors in Computing Systems*. 2119.

[32] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. *Proceedings of the 2013 Conference on Knowledge Discovery and Data Mining*. 623. DOI:https://doi.org/10.1145/2487575.2487579

[33] Linda P. Lowes, Lindsay N. Alfano, Brent A. Yetter, Lise Worthen-Chaudhari, William Hinchman, Jordan Savage, Patrick Samona, Kevin M. Flanigan, and Jerry R. Mendell. 2013. Proof of concept of the ability of the kinect to quantify upper extremity function in dystrophinopathy. *PLOS Curr.* 5 (Jan. 2013). DOI: https://doi.org/10.1371/currents.md.9ab5d872bbb944c6035c9f9bfd314ee2

[34] Ketan K. Mane, Chris Bizon, Charles Schmitt, Phillips Owen, Bruce Burchett, Ricardo Pietrobon, and Kenneth Gersing. 2012. VisualDecisionLinc: A visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. *J. Biomed. Info.* 45, 1 (2012), 101–106. DOI: https://doi.org/10.1016/j.jbi.2011.09.003

[35] Sinziana Mazilu, Ulf Blanke, Michael Hardegger, Gerhard Tröster, Eran Gazit, and Jeffrey M. Hausdorff. 2014. GaitAssist: A daily-life support and training system for parkinson's disease patients with freezing of gait. In *Proceedings of the 2014 Conference on Human Factors in Computing Systems.* 2531–2540. DOI: https://doi.org/10.1145/2556288.2557278

[36] Helena M. Mentis, Amine Chellali, and Steven Schwaitzberg. 2014. Learning to see the body. In *Proceedings of the 2014 Conference on Human Factors in Computing Systems.* 2113–2122.

[37] Helena M. Mentis and Alex S. Taylor. 2013. Imaging the body: Embodied vision in minimally invasive surgery. In *Proceedings of the 2013 CHI Conference on Human Factors in Computing Systems.* 1479–1488. DOI: https://doi.org/10.1145/2466110.2466197

[38] Cecily Morrison, Marcus D'Souza, Kit Huckvale, Jonas F. Dorn, Jessica Burggraaff, Christian Philipp Kamm, Saskia Marie Steinheimer, Peter Kontschieder, Antonio Criminisi, Bernard Uitdehaag, Frank Dahlke, Ludwig Kappos, and Abigail Sellen. 2015. Usability and acceptability of ASSESS MS: A system to support the assessment of motor dysfunction in Multiple Sclerosis using depth-sensing computer vision. *JMIR Human Fact.* 2, 1 (2015).

[39] Cecily Morrison, Kit Huckvale, Bob Corish, Jonas Dorn, Kenton O'Hara, ASSESS MS Team, Antonio Criminisi, and Abigail Sellen. 2016. Assessing multiple sclerosis with kinect: Designing computer vision systems for real-world use. *Human-Comput. Interact.* 31, 3–4 (2016), 191–226.

[40] Mark A. Musen, Blackford Middleton, and Robert A. Greenes. 2014. *Biomedical Informatics.* Springer, London.

[41] Sébastien Pierard, Remy Phan Ba, Valérie Delvaux, Pierre Maquet, and Marc Van Droogenbroeck. 2013. GAIMS: A powerful gait analysis system satisfying the constraints of clinical routine. *Multiple Sclerosis J.* 19, S1 (Oct. 2013), 359.

[42] Advait Sarkar. 2015. Confidence, command, complexity: Metamodels for structured interaction with machine intelligence. In *Proceedings of the 2015 Conference of the Psychology and Programming Interest Group.*

[43] Advait Sarkar, Cecily Morrison, Jonas F. Dorn, Rishi Bedi, Jacques Steinheimer, Saskia Boisvert, Jessica Burggraaff, Peter D'Souza, Marcus Kontschieder, Sebastian Rota Bulò, and Lorcan Walsh. 2016. Setwise comparison: Consistent, scalable, continuum labels for computer vision. In *Proceedings of the 2016 Conference on Human Factors in Computing Systems.* 261–271.

[44] Hossam Sharara, Awalin Sopan, Galileo Namata, Lise Getoor, and Lisa Singh. 2011. G-PARE: A visual analytic tool for comparative analysis of uncertain graphs. In *Proceedings of the 2011 Conference on Visual Analytics Science and Technology.* 61–70.

[45] Sheldon R. Simon. 2004. Quantification of human motion: Gait analysis-benefits and limitations to its application to clinical problems. *J. Biomech.* 37, 12 (December 2004), 1869–1880. DOI: https://doi.org/10.1016/j.jbiomech.2004.02.047

[46] Aneesha Singh, Annina Klapper, Jinni Jia, Antonio Fidalgo, Ana Tajadura-Jiménez, Natalie Kanakam, Nadia Bianchi-Berthouze, and Amanda Williams. 2014. Motivating people with chronic pain to do physical activity. In *Proceedings of the 2014 Conference on Human Factors in Computing Systems.* 2803–2812.

[47] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Human-Comput. Studies* 67, 8 (Aug. 2009), 639–662.

[48] Justin Talbot, Bongshin Lee, Ashish Kapoor, and Desney S. Tan. 2009. EnsembleMatrix: Interactive visualization to support machine learning with multiple. In *Proceedings of the 2009 Conference on Human Factors in Computing Systems.* 1283.

[49] Leonard Verhey, Colm Elliott, Helen Branson, Cristina Philpott, Manohar Shroff, Tal Arbel, Brenda Banwell, and Douglas Arnold. 2014. Rate of agreement for manual and automated techniques for determination of new T2 lesions in children with multiple sclerosis and acute demyelination (P2.242). *Neurology* 82, 10 Suppl. (Apr. 2014), P2.242.

[50] Vedrana Vidulin, Marko Bohanec, and Matjaž Gams. 2014. Combining human analysis and machine data mining to obtain credible data relations. *Info. Sci.* 288 (Dec. 2014), 254–278.

[51] William Young, Stuart Ferguson, Sébastien Brault, and Cathy Craig. 2011. Assessing and training standing balance in older adults: A novel approach using the "Nintendo Wii" Balance Board. *Gait Post.* 33, 2 (Feb. 2011), 303–305. DOI: https://doi.org/10.1016/j.gaitpost.2010.10.089

Received December 2016; revised November 2017; accepted December 2017