Biology Methods and Protocols, 2025, **10(1)**, bpaf029
https://doi.org/10.1093/biomethods/bpaf029
Advance Access Publication Date: 12 April 2025
Teaching New Methods

Development of a sperm morphology assessment standardization training tool

Katherine R. Seymour[1,*], Jessica P. Rickard[1], Kelsey R. Pool[2], Taylor Pini[3], Simon P. de Graaf[1]

[1]School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Sydney NSW, Australia
[2]School of Agriculture and Environment, The University of Western Australia, Crawley, WA, Australia
[3]School of Veterinary Science, Faculty of Science, The University of Queensland, Gatton, QLD, Australia

*Corresponding author. School of Life and Environmental Sciences, Room 344, RMC Gunn Building, B19, Faculty of Science, The University of Sydney, Sydney 2006, NSW, Australia. E-mail: katherine.seymour@sydney.edu.au

Abstract

Training to improve the standardization of subjective assessments in biological science is crucial to improve and maintain accuracy. However, in reproductive science there is no standardized training tool available to assess sperm morphology. Sperm morphology is routinely assessed subjectively across several species and is often used as grounds to reject or retain samples for sale or insemination. As with all subjective tests, sperm morphology assessment is liable to human bias and without appropriate standardization these assessments are unreliable. This proof-of-concept study aimed to develop a standardized sperm morphology assessment training tool that can train and test students on a sperm-by-sperm basis. The following manuscript outlines the methods used to develop a training tool with the capability to account for different microscope optics, morphological classification systems, and species of spermatozoa assessed. The generation of images, their classification, organization, and integration into a web interface, along with its design and outputs, are described. Briefly, images of spermatozoa were generated by taking field of view (FOV) images at 40× magnification on DIC optics, amounting to a total of 3,600 FOV images from 72 rams (50 FOV/ram). These FOV images were cropped to only show one sperm per image using a novel machine-learning algorithm. The resulting 9,365 images were labelled by three experienced assessors, and those with 100% consensus on all labels (4821/9365) were integrated into a web interface able to provide both (i) instant feedback to users on correct/incorrect labels for training purposes, and (ii) an assessment of user proficiency. Future studies will test the effectiveness of the training tool to educate students on the application of a variety of morphology classification systems. If proven effective, it will be the first standardized method to train individuals in sperm morphology assessment and help to improve understanding of how training should be conducted.

Keywords: sperm morphology assessment; standardised training tool; reproduction; advanced semen assessment; subjective assessment

Introduction

Standardization training is an essential method of maintaining accuracy and reducing variability when performing subjective tests. Sperm morphology assessment, a key predictor of fertility and reproductive health, is a subjective test that despite suffering from human bias has no recognized standardization training method [1]. In human reproductive medicine, this issue has been explored across several studies into sperm morphology assessment variation, with uncertainty surrounding morphologist accuracy being attributed to a lack of a traceable standard to both test and train morphologists [1–3]. Existing external quality control programmes have been established in human reproductive medicine but determine accuracy by comparing population data across laboratories [4, 5]. This focuses only on the percentage of normal sperm, and accuracy is determined by observing the variation from the mean result of all participating laboratories when given the same sample (often stated as ± 2 standard deviations from the mean being an acceptable level of variation) [6]. This introduces substantial variation, as each morphologist is assessing different individual sperm, and prevents any insight into accuracy for different morphological categories even when using more complex classification systems [7]. To rectify this lack of confidence in sperm morphology assessment and improve accuracy, attempts at producing a robust training method have been made but haven't been feasible for application.

A small number of studies have investigated different training methods to improve the accuracy of assessing sperm morphology. An early investigation into training explored a classroom-based setting, wherein novice morphologists were shown images of abnormal sperm collectively and told to assess 100 sperm each as normal or abnormal using the WHO 1987, WHO 1992, and strict criteria [8]. This method yielded no significant improvement following training. Most notably this study found that in 43% of instances, novices reversed their classification of the same sperm during the second test. There was also a significant reduction in the amount of normal sperm being correctly classified post-training [8]. This classroom-based setting had potential as a training technique on a sperm-by-sperm basis but required all novices to learn at the same pace and only train using a

Received: 28 February 2025; **Revised:** 28 March 2025; **Editorial decision:** 04 April 2025; **Accepted:** 09 April 2025

simple binary classification system. Later, recommended training methods, which have no published validation metrics, involved side-by-side training with a trained assessor [6]. This was a time-consuming process for both the trainee and the assessor and relied heavily upon the assessor already being standardized. When expert morphologists are required to re-standardize, this technique loses its effectiveness, as a more qualified morphologist may not be available. This further bolsters the need for a robust training method that is accessible and has been standardized on a sperm-by-sperm basis [1]. However, as sperm morphology assessment remains a subjective technique, the issue arises of how to standardize and validate individual classifications.

The conundrum of providing validated data from subjective techniques has been explored in detail in the machine learning space, with the term 'ground truth' being coined to represent data that have been accurately classified. Supervised training, wherein a machine learning model 'learns' how to classify images from provided labelled data, is one of the key methodologies used to train models, particularly for image recognition [9]. The model will only learn off the dataset it has been provided, hence if a dataset was labelled incorrectly (e.g. due to natural human variation in subjective techniques like sperm morphology assessment), the accuracy of the model will be compromised. Numerous studies into the negative impact of poorly classified datasets on the effectiveness of machine learning models have reinforced the importance of validating subjective classifications [10]. In the medical field, this has resulted in the requirement of classifications to be first made, then validated, by an expert in the field. Machine learning has been applied to sperm morphology assessment in a handful of studies, with multiple studies using consensus classified data when generating their ground truth [11–13]. One study even noted that the prevision recall of their machine learning model could be improved by 12.6–26% when a two-person consensus strategy was used [14]. If we accept that machine learning algorithms require training datasets validated by consensus among multiple experts to achieve accuracy, it prompts the question of why human learning and classification are not held to equally stringent standards. It may be unrealistic to expect humans to outperform machine learning if they are not also trained on robust, validated data. In applications such as sperm morphology assessment, where high levels of intra- and inter-morphologist variation are well-documented, applying principles of ground truth could provide a rigorous framework for both training and evaluating human assessors accurately.

Given the limitations of previous unstandardized training methodologies for sperm morphology assessment—such as individual side-by-side training and lecture-based learning—along with the challenges associated with providing validated classifications for a subjective evaluation, the development of a training tool that addresses these factors could be of significant value to the industry. As such, this study aimed to develop an interactive sperm morphology assessment standardization training tool by establishing robust 'ground truth' in a dataset of classified sperm. The training tool was designed to provide (i) a true assessment of a user's accuracy by testing them on a sperm-by-sperm basis against expert-validated sperm morphology classifications, and (ii) a subsequent method of standardization training that was able to be performed independently and self-paced. As a key aspect of the training tool's development, this study aimed to establish what requirements are needed to create a dataset of classified ram sperm images that had ground truth to the standard expected of machine learning training data. It was hypothesized that expert morphologist consensus would be reasonably high and that consequently, sperm would not need to be labelled by multiple experts.

Materials and methods

Though formally titled the 'sperm morphology assessment standardization training tool', for brevity this title has been simplified to the 'training tool' in subsequent sections.

Data production

For the training tool to function, a dataset of clear, classified sperm images needed to be supplied. The images needed to be of high resolution, have only a single sperm per image to avoid confusion, and be classified with a high degree of confidence. To achieve this, multiple steps as outlined below, were taken.

Image collection

Semen samples from 72 rams from across NSW were sourced for the study. An Olympus BX53 microscope with differential interference contrast (DIC) and phase contrast objectives at 40× magnification was used to capture the images used in this study (Olympus Australia, Notting Hill VIC, Australia). Objectives with high numerical apertures (NA), 0.75 and 0.95 for phase contrast and DIC respectively, were chosen to maximize resolution. Along with the microscope, an Olympus DP28 camera was used to capture images using an 8.9-megapixel CMOS sensor, 25 field number (FN) at 4000px resolution. As with the objectives and microscope body, this camera was chosen specifically for use in this study. Per sire, 50 fields of view (FOV) were captured ($n = 3{,}600$), excluding particularly agglutinated samples.

Sperm morphology classification systems

To ensure the training tool was able to appropriately adapt to any sperm morphological classification system, sperm were classified into a large number of categories. This ensured that images could be sorted into the appropriate category of any classification system that used fewer categories than our dataset, e.g. the 5-category location-based classification [15] or the 8-category Australian Cattle Vets classification system [16]. For research or standardization purposes, this made the training tool highly adaptable as both a training and assessment tool. To manipulate the labelled images into multiple different classification systems, a comprehensive 30-category system was developed (Table 1).

File naming system

For both researchers and the training tool, a method of recording information about each image was developed. As the overall aim was to develop a training tool that could train users on different species, microscope optics, and classification systems, that information needed to be recorded for each sperm. To allow all relevant information for each sperm to be easily accessible, it was decided that all information for each image was to be stored in the file name.

When taking the FOV image, a simple file naming convention was used to convey the species and microscope optics being used. Using the alphabet as a code for a specific species or optics was an effective method of conveying this information. For example, the below file name could be interpreted as the following.

AA_0000013.png.
A = microscope optics used; A = phase contrast 40x, B = DIC 40x
A = species; A = ram

Table 1. 30-category morphological abnormalities classification system used to develop the labelled data for the novel sperm morphology assessment standardisation training tool. All categories used are listed with descriptions and early accounts in the literature.

Classification	Description	First described in literature
Normal	A normal sperm has an even shape and consistent size, note the smooth head and acrosome, the lack of bends or breaks in the tail and the size.	Human [17] Cattle [18]
Abaxial tail	The tail is not centred on the head, but rather attached slightly off centre.	Cattle [19]
Bent midpiece	The midpiece is sharply bent but not completely broken as seen in a broken neck nor as severe as with a midpiece reflex.	Cattle [18]
Segmental aplasia	First described as a ‘nicked’ midpiece, it is characterized by the midpiece having a gap along its length or a slightly strange texture but not enough to be considered a cytoplasmic or pseudo-droplet.	Cattle [20]
Slightly pyriform head	Head has a slightly pear shape, near the midpiece the edge of the head tucks in slightly.	Cattle [18]
Narrow head	Sperm head is narrower by around ¾ of normal size, but the length remains the same.	Human [21] Cattle [22]
Detached/decapitated head	Head is not connected to the midpiece or tail. A detached head occurs as an abnormality during developmental issues (e.g. heat stress). A decapitated head is distinguished by moving free tails being present in the ejaculate with a proximal cytoplasmic droplet.	Cattle [18]
Multiple heads	Sperm has more than one head.	Human [23] Cattle [22]
Pyriform head	The head is shaped like a teardrop and significantly tucks in at the base.	Cattle [20]
Microcephalic head	The head is at least 25% smaller than normal.	Human [23] Cattle [22]
Macrocephalic head	The head is at least 25% larger than normal, sometimes seen in conjunction with multiple tails (though there is no known incidence of this in cattle).	Human [23] Cattle [22]
Rolled head	The head has folded over itself partially or completely, often seen along the long axis of the head.	Cattle [24]
Swollen acrosome	Acrosome can be distinguished as a 'halo' above the head.	Human [21] Cattle [25]
Knobbed acrosome	Acrosome is misshaped; this could mean it is indented, beaded, or flattened.	Cattle [26]
Diadem defect	A line of dark dots along the width of the head, like a string of pearls. Coined ‘diadem’ by Blom [22].	Cattle [27]
Nuclear vacuole	There is a dark crater on the sperm head.	Cattle [18]
Teratoid	The head has been destroyed, leading to a textured surface and/or misshapen appearance.	Human [23] Cattle [20]
Missing acrosome	The acrosome is missing.	Cattle [18]
Distal midpiece reflex	The midpiece has folded over itself.	Cattle [18]
Broken neck	The head is bent almost perpendicular to the tail.	Fowl [28] Cattle [22]
Multiple midpieces	Distinguished by a lighter line down the middle of a thicker midpiece.	Cattle [20]
Pseudodroplet	First described as a ‘pseudo-swelling’ of the midpiece, it can be identified as a thickening of the midpiece which appears almost as if it were a cytoplasmic droplet. It does not appear spherical, and is longer than a cytoplasmic droplet.	Cattle [20] (first identified but not described in detail) Cattle [29] (first detailed classification)
Dag defect	The midpiece and tail are folded over onto themselves multiple times.	Cattle [30]
Corkscrew defect	The midpiece has bumps along its side as if it has been twisted.	Cattle [31]
Distal cytoplasmic droplet	A cytoplasmic droplet near the end of the tail.	Human [23] Cattle [22]
Proximal cytoplasmic droplet	A cytoplasmic droplet on the midpiece.	Human [23] Cattle [22]
Stumped tail	The tail is shorter (any length past the midpiece) than expected.	Human [23] Cattle [22]
Multiple tails	The sperm has more than one tail.	Human [23] Cattle [22]
Reflex tail	The tail has folded over itself.	Cattle [22]
Coiled tail	The tail has wrapped around itself making a coil.	Cattle [20]

_

0000013 = image ID number

Following the collection of the FOV images, additional data needed to be recorded for every individual sperm which necessitated the generation of a new file name. Per sperm, the information that needed to be recorded included:

1. If the image had been classified yet
2. Microscope optics used
3. Species
4. Image ID number
5. If the assessors found the image easy or hard to classify
6. Morphological classification

Recording species and optics allowed the data to be sorted depending on what the user needed to be trained on. The image ID number allowed each sperm to have a unique identifier, as multiple files would eventually have similar labels. The addition of a 'difficulty' label was devised for future research focused on investigating if certain abnormalities were deemed easier or harder to classify. For all the above labels, as with the FOV file names, using a simple alphabet code was sufficient. Likewise, a numeric ID number was deemed appropriate.

As the morphology of each sperm was to be classified into up to 30 different categories, a hexadecimal code was used to minimize the size of the file name. Thirty different abnormality classifications were translated into an eight-digit hexadecimal code, which would be unique for each variation of categories used. For example, the below file name could be interpreted as follows:

SAA_FFFFFFFF_XM_9999999.png
S = if the image has been classified; S = expert classified image, K = unclassified image
A = microscope optics used; A = phase contrast 40x, B = DIC 40x
A = species; A = ram

_

FFFFFFFF = morphological classification written as an 8-digit hexadecimal code

_

X = If the sperm was difficult to classify; X = easy, Y = hard
M = An extra space in case an additional piece of information needs to be recorded
9999999 = file ID number

In practice, this would be translated to encode all relevant information for an individual spermatozoon. For example, SBA_80000000_XM_0000007.png would be translated as a classified image taken using a DIC 40x objective of ram sperm, which was classified as normal and easy to identify. The training tool was built with the functionality to interpret the file name so that no additional data need be uploaded in addition to the image in order to function. The development of a comprehensive file naming convention allowed for multiple data points to be recorded per image; however, to proceed, the FOV images needed to be cropped to show only a single sperm per image.

Cropping fields of view using a machine learning algorithm

As previously mentioned, to maximize consistency in the images shown to users of the training tool, it was decided that each image should contain only one sperm and that the dimensions of the image should be of a consistent size. This required each FOV ($n = 3,600$) to have all visible sperm that weren't overlapping or too close to another sperm cropped out of the image. Initial attempts to do this by hand proved difficult and extremely time-consuming, so an automated method was explored.

Machine learning, which is a facet of the technologies that can be labelled as having artificial intelligence, has a multitude of applications and has already been applied to object recognition and labelling in other reproductive assessment contexts, such as embryo quality [32]. As such, the application of a machine learning algorithm to streamline the processing of the FOVs was deemed appropriate. A convolutional neural network that was trained using supervised learning datasets was developed. To do this, a model was created using a custom dataset of 200 manually cropped images, focusing on individual cells. This dataset was used to train a YOLOv8 object detection network, which provided precise cell detection and cropping for larger datasets. Training the custom YOLOv8 model was exceptionally fast, and its high accuracy enabled efficient and error-free processing of large image datasets. The YOLOv8 network ensured that the resulting cropped images were consistent and of high quality. To quality control, the model, a human assessor compared the raw FOV images to the resulting cropped images to ensure the model did not ignore certain abnormalities and consequently introduce bias into the dataset. Following processing by the machine learning model, 3,600 FOVs were cropped to produce 9,365 images of individual sperm. After the production and processing of the raw image data, the classification of individual sperm images could commence.

Expert morphologist classification consensus

Sperm morphology assessment is a subjective test, and consequently liable to human bias, which made providing a correctly classified sperm image challenging. To provide robust 'ground truth' data, the classifications had to be correct. To avoid variation among the classifiers, it was decided that expert morphologists would be used, with an expert being defined as an individual who was a leading reproductive researcher and assessed a minimum of 10,000 sperm in this study's target species (Ram). A preliminary study was run to determine the agreement between the expert morphologists ($n = 3$), which would then inform the 'coverage' needed when classifying images. It was hypothesized that the experts would have a relatively high agreement and that the sperm would only need to be classified by two out of the three experts which would minimize the amount of sperm each individual needed to classify.

Experts were asked to classify 800 images of ram sperm taken using DIC optics and 800 taken using phase contrast ($n = 1600$), both at 40× magnification. Sperm were classified using the 30 morphological categories as described in the section on '*Sperm morphology classification systems*'. Following classification, the data were then explored to reflect whether the consensus would have changed if the images had been labelled using a binary morphological category system (normal/abnormal). When using the normal/abnormal classification system with DIC, morphologists agreed on the label of 585 sperm out of 800 (73%) and when labelling with the 30-category system the morphologists agreed on 212 sperm out of 800 (26.5%). When using the normal/abnormal classification system with phase contrast, the morphologists agreed on the label of 641 sperm out of 800 (80%), and when labelling with the 30-category system, the morphologists agreed on 253 sperm out of 800 (31.6%)). This preliminary study revealed that expert morphologists have much higher levels of disagreement than was hypothesized, particularly when using complex classification systems. As a result of these findings, it was decided that each image would require 100% consensus between three expert morphologists to be used as 'ground truth'.

Images were labelled using an online tool (Labelbox, n.d.), commonly used for the classification of image-based data destined for machine learning applications. This allowed each expert to log onto the website independently and label images that were randomly shown to them on Labelbox. Sperm classifications that didn't have 100% consensus were shown to a different, fourth expert morphologist. During this second 'attempt' at classifying the sperm, if three of the four experts agreed on the classification, the image was included. This second step was implemented to try and increase the overall yield, as achieving 100% consensus proved to be difficult. In total, 4,821 labelled sperm using DIC optics were produced and 1,683 labelled sperm using phase contrast. The discrepancy between the size of the two optic datasets was related to choosing to focus fully on the industry gold standard of DIC early in the labelling process when time constraints became evident. After the labelling process, a dataset that had irrefutable ground truth was developed and was ready for use in the training tool.

Training tool design

The training tool detailed in this study was developed for this study. The following sections detail various aspects considered for the design of the training tool.

Capability requirements

From the outset, it was decided that the training tool would be developed to allow it to be adapted to any species, microscope optics, or classification system. The 'ground truth' dataset and its associated file naming convention (sections on '*Sperm morphology classification systems*' and '*File naming system*') were developed with this capability in mind. In addition to these requirements, the training tool also needed to have the ability to train and test users. This was implemented as two 'modes' of the training tool that users could swap between as needed. The test mode would need to provide the users with a score of accuracy that could be used to monitor their progress, while also recording the data for research purposes. The training mode would need to employ the principles of supervised training and reinforced learning. In the user context, this meant users would need to be able to classify a sperm and then receive immediate feedback as to which of the selected categories were correct or incorrect, as well as indicate the correct answer.

In a research context, the training tool would require the capability of recording data that could be accessed and downloaded by the administrator of the training tool. Along with the accuracy score, it was of interest which specific categories had been chosen by the user and if that selection was correct or incorrect. The duration a user spent classifying an image was also of interest, to explore if there was a correlation between the time taken for classification and accuracy. These data were collected for both test and training modes; however, only the test mode reported user accuracy visually within the website. All data needed to be accessible to researchers while remaining restricted from user access. Additionally, the system was required to track which user-generated each dataset. Hence, an administrator page was created for this information to be viewed and exported into a .csv file following competition of the study. The administrator page had the dual purpose of controlling all the various factors of the training tool, such as the instructional video and visual aid (explained further in the section on '*Supplementary materials*'), as well as defining what dataset could be accessed with a given access code.

Supplementary materials

Users of the training tool were further supported with two supplementary materials: a visual aid (see supplementary file 1) and an instructional video (which can be viewed by following this link: https://youtu.be/60pn1JDrs8w). The visual aid listed all possible morphological categories and provided a corresponding picture and written description. Upon signing into the training tool, users would be prompted to download the visual aid which could be used to assist with training.

The instructional video was also prompted at the same time as the visual aid and is viewed within the training tool. Users were allowed to skip the video if it had been watched before. Along with the visual aid, the video listed all possible morphological categories while showing example images. In addition to showing the categories, the video also briefly explained what sperm morphology assessment was and how to use the training tool.

Upon logging onto the training tool (and choosing the appropriate species, optics, or classification system if the option was provided), users were immediately prompted to watch the instructional video and/or download the visual aid. This ensured new users (or returning users) had the opportunity to view the video and the visual aid before attempting to assess the sperm.

Test mode

Perhaps the most critical capability of the training tool was its test mode, which allowed users to attempt to assess 100 sperm using whichever classification system was selected by the admin. Unlike the current industry approach to assessing human competency in sperm morphological classification, the training tool assessed users on a sperm-by-sperm basis rather than by their final accuracy score only. As mentioned previously, users had to correctly select all appropriate categories to be deemed 'correct' for that sperm. By testing users in this method, we could determine what specifically was causing a user to incorrectly classify an individual sperm. This allows us to recognize any abnormalities commonly misidentified or conversely easily identified. For users, this meant that their final accuracy score after completing a test was a true representation of their accuracy. If only using normal/abnormal classification, the reason that a user identified a sperm as abnormal could be incorrect even if they were right in identifying there was something wrong with the sperm. This could lead to future inaccuracies and perpetuate human bias [6]. By utilizing a sperm-by-sperm testing method with multiple morphological categories, it could be assured that users were truly able to identify abnormalities. While this test was able to identify the true accuracy of a user, it did little to provide any form of training due to a lack of feedback on a sperm-by-sperm basis. Hence, in addition to the testing mode, a training mode was developed.

Training mode

Designed to work in tandem with the testing mode, the training mode in the training tool mimicked the format of the test with one key exception. In training mode, users were presented with images for assessment in a random order, with no time limit per image and received feedback on a sperm-by-sperm basis. Feedback (incorrect/correct) was presented immediately on screen, including the provision of the correct answer. No overall accuracy score was provided after training. This instant feedback model provided users with the ability to self-teach by reflecting on their answers. As there was no time limit per image, users could spend as much time as they felt was needed before receiving their feedback. At any time, users could exit training mode and reattempt the test, or conversely, they could remain in the training mode for as long as required and continue to be shown randomly ordered sperm images. Consequently, there was no overall accuracy score provided for the test, as users were encouraged to focus on testing themselves one sperm at a time. The size of the provided dataset for the training mode could be manipulated in the admin page.

User interface with Figma

To support the functionality of the training tool, a bespoke user interface was designed to ensure users engaged meaningfully with the training materials. While the training tool needed to include all relevant components to train and test users, the user interface also needed to be considered carefully. It is recognized that users will have a negative interaction with training tools if the user interface is difficult to use, which can compromise their willingness to use the training tool [33]. During the early design process, the layout and functionality of the training tool were drafted on the design tool 'Figma' (Figma, n.d.) (Fig. 1). Figma facilitated the initial testing of the user interface and allowed for the layout of the website to be finalized before commissioning the build. The user interface was designed with the intended functionality of allowing users to select species, microscope optics, and classification systems before beginning testing or training. Users were also able to exit out of the testing and training mode to change those selections at any time. To allow for data to be correctly sorted, and to limit the use of the website without permission, an access code was included. This access code would be linked to certain datasets and could also be used to group results for export. The code could be changed at any time on the administrator page, and multiple codes could be in effect simultaneously.

Outputs

Accuracy

To differentiate this training tool from similar attempts at industry training, users needed to be assessed on a sperm-by-sperm basis, not on a population level, and the data they were tested against were validated by experts. For sperm morphology analysis, this refers to each user's accuracy being determined by each sperm being correctly labelled using all provided morphological categories, rather than just considering if the overall ratio of normal to abnormal was correct. By using the expert-validated dataset, which was explained in detail in the section on '*Expert morphologist classification consensus*', the accuracy determined during the test can be trusted to be a true representation. For the users to determine how they were progressing in training, this accuracy score was shown after each test. For research use, the

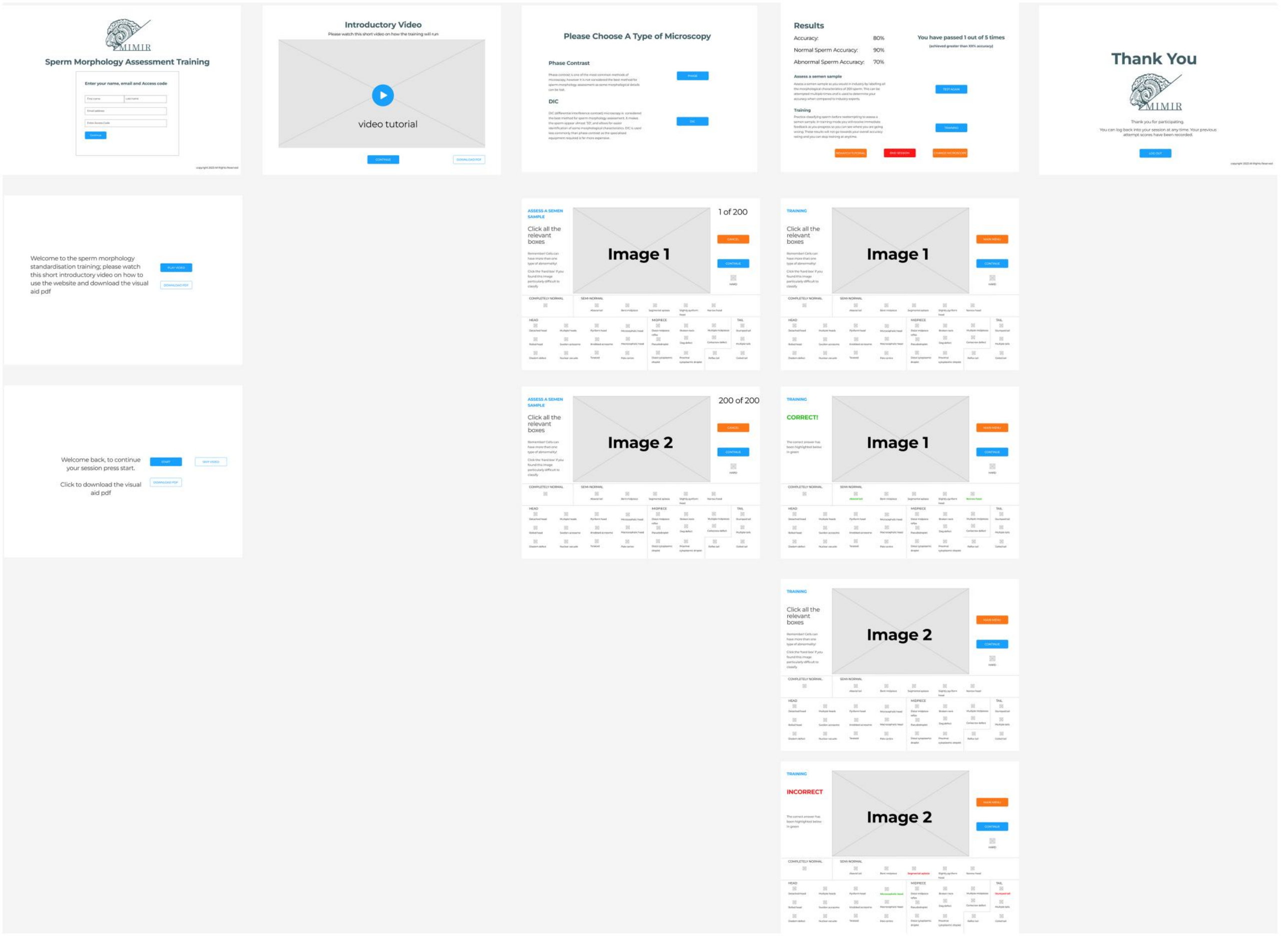
MIMIR
Sperm Morphology Assessment Training
Enter your name, email and Access code
Introductory Video
Please watch this short video on how the training will run
video tutorial
Please Choose A Type of Microscopy
Phase Contrast
DIC
Results
Accuracy: 80%
Normal Sperm Accuracy: 90%
Abnormal Sperm Accuracy: 70%
You have passed 1 out of 5 times
Assess a semen sample
Training
Thank You
MIMIR
Thank you for participating.
You can log back into your session at any time. Your previous attempt scores have been recorded.
Welcome to the sperm morphology standardisation training; please watch this short introductory video on how to use the website and download the visual aid pdf
Welcome back, to continue your session press start.
Click to download the visual aid pdf
ASSESS A SEMEN SAMPLE
Click all the relevant boxes
Image 1
1 of 200
Image 2
200 of 200
TRAINING
CORRECT!
INCORRECT
Image 1
Image 2

Figure 1. User interface for the 'sperm morphology assessment standardization training tool' generated using Figma

accuracy per sperm per category was recorded by the training tool and could be exported for further analysis. This was done by recording which categories were chosen by the user and comparing that data to what the correct classification was for that sperm (determined by the hexadecimal code in the file name as per the section on '*File naming system*' of this study). User accuracy per category was not displayed to the user, only the overall accuracy for each test.

Duration

In addition to accuracy per sperm labelled, the duration spent labelling a sperm was recorded by the training tool. These data were recorded as mm:ss.0 and were also timestamped with the real-world time and date to further validate the duration was correct. This metric, as with the sperm-by-sperm accuracy, was not accessible to the user to prevent time pressure on users to classify sperm. Duration was used for research purposes to determine if accuracy could be related to time spent classifying, or if certain abnormalities took longer to classify than others.

Conclusions

This study resulted in the successful development of a comprehensive and flexible sperm morphology assessment standardization training tool. The training tool developed in the present study enhances the class-wide sperm-by-sperm training method proposed by Davis (1995). In addition to enabling an entire class to learn simultaneously, it replaces traditional lecture-based instruction with an interactive platform that provides individualized, real-time feedback. It is hypothesized that this approach will foster a more engaging and transformative learning experience. Our tool can test and train users using spermatozoa from a variety of species, observed using a range of microscope optical modes. The tool could also be applied (where sufficient classified images of sperm are generated) across fields of human medicine, agriculture, and species conservation. More wildly, the tools' structure could be applied as a basis for the investigation of other cell types' morphological assessment. While currently configured to assess sperm morphology using 30 different morphological categories, the training tool can adjust the number of categorical classifications being used to suit any existing morphology classification system. This is due to the comprehensive file naming convention which conveys the relevant data for each specific sperm image and facilitates filtering based on that data by the training tool. By assessing expert morphologist agreement, and finding it substantially lower than hypothesized, it was concluded that every sperm morphological classification must achieve 100% consensus among the experts to be used in the training tool. This allowed the training tool to adhere to strict dataset validation principles commonly used in machine learning, i.e. ground truth. The test and train modes for the tool utilized supervised learning and reinforcement learning machine learning methodologies, respectively.

Future research will seek to validate the effectiveness of the training tool in standardizing novice morphologists using a variety of classification systems by measuring the inter-observer variability and accuracy of users. It is hypothesized that due to the robust training dataset and users being tested and trained on a sperm-by-sperm basis the use of the training tool will produce highly accurate sperm morphologists. If the training tool proves to be effective, there is the potential to consider diversifying the dataset into different species, such as bulls, to make it useful for a wider audience. Following this development, and the validation of the training tool in multiple species, we would achieve our ultimate aim of its use as a real-world classroom training tool for undergraduate and postgraduate students across the globe.

Acknowledgements

KS was the recipient of the Christian Rowe Thornett Stipend Scholarship. The authors thank M Seymour and J Samm for their valuable comments and contributions to the design of the training tool.

Author contributions

KS - Data curation, formal analysis, investigation, methodology, project administration, validation, visualisation and writing of original draft. JR - Funding acquisition, writing review and editing. KP and TP - data curation and writing review and editing. SdG - conceptualisation, data curation, funding acqusition, resources, supervision and writing review and editing.

Supplementary data

Supplementary data are available at *Biology Methods and Protocols* online.

Conflict of interest statement. The authors declare that they have no conflicts of interest.

Funding

No external funding was received for this study.

Data availability

The data and training tool are not publicly available. Following publication of further research and sufficient funding, the authors intend to make the training tool publicly available.

References

1. Tomlinson MJ. Uncertainty of measurement and clinical value of semen analysis: has standardisation through professional guidelines helped or hindered progress? *Andrology* 2016; **4**:763–70.
2. Barratt CLR, Björndahl L, Menkveld R *et al*. ESHRE special interest group for andrology basic semen analysis course: a continued focus on accuracy, quality, efficiency and clinical relevance. *Hum Reprod* 2011; **26**:3207–12.
3. Carrell DT, De Jonge CJ. The troubling state of the semen analysis. *Andrology* 2016; **4**:761–2.
4. Ahadi M, Aliakbari F, Latifi S *et al*. Evaluation of the standardization in semen analysis performance according to the WHO protocols among laboratories in Tehran, Iran. *Iran J Pathol* 2019; **14**:142–7.
5. Ombelet W, Bosmans E, Janssen M *et al*. Multicenter study on reproducibility of sperm morphology assessments. *Arch Androl* 1998; **41**:103–14.
6. Agarwal A, Sharma R, Gupta S *et al*. Sperm morphology assessment in the era of intracytoplasmic sperm injection: reliable

results require focus on standardization, quality control, and training. *World J Mens Health* 2022; **40**:347–60.
7. Riddell D, Pacey A, Whittington K. Lack of compliance by UK andrology laboratories with World Health Organization recommendations for sperm morphology assessment. *Hum Reprod* 2005; **20**:3441–5.
8. Davis RO, Gravance CG, Overstreet JW. A standardized test for visual analysis of human sperm morphology. *Fertil Steril* 1995; **63**:1058–63.
9. Saravanan R, Sujatha P (eds). A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS); 2018 14-15 June 2018.
10. Lebovitz S, Levina N, Lifshitz-Assaf H, University of Virginia. Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *Misq* 2021;**45**:1501–26.
11. Chang V, Garcia A, Hitschfeld N *et al.* Gold-standard for computer-assisted morphological sperm analysis. *Comput Biol Med* 2017; **83**:143–50.
12. Chen A, Li C, Zou S *et al.* SVIA dataset: a new dataset of microscopic videos and images for computer-aided sperm analysis. *Biocybernetics and Biomedical Engineering* 2022; **42**:204–14.
13. Shaker F, Monadjemi SA, Alirezaie J *et al.* A dictionary learning approach for human sperm heads classification. *Comput Biol Med* 2017; **91**:181–90.
14. Wong DR, Tang Z, Mew NC *et al.* Deep learning from multiple experts improves identification of amyloid neuropathologies. *Acta Neuropathol Commun* 2022; **10**:66.
15. Menkveld R. Sperm morphology assessment using strict (Tygerberg) criteria. In: Carrell DT, Aston KI (eds), *Spermatogenesis: Methods and Protocols*. Totowa, NJ: Humana Press, 2013, 39–50.
16. Perry VEA. The role of sperm morphology standards in the laboratory assessment of bull fertility in Australia. *Front Vet Sci* 2021; **8**:672058.
17. von Ebner V. Untersuchungen über den Bau der Samencanälchen und die Entwicklung der Spermatozoiden bei den Säugethieren und beim Menschen: W. Engelmann, 1871.
18. Williams WW. Technique of collecting semen for laboratory examination with a review of several diseased bulls. *Cornell Vet* 1920; **10**:87–94.
19. Aughey E, Renton JP. Abnormal spermatozoa in an Ayrshire bull. *Vet Rec* 1968; **82**:129–131.
20. Williams WW, Savage A. Methods of determining the reproductive health and fertility of bulls, a review with additional notes. *Cornell Vet* 1927; **17**:374–85.
21. Moench GL. The technic of the detailed study of seminal cytology. *Am J Obstet Gynecol* 1930; **19**:530–8.
22. Blom E. The ultrastructure of some characteristic sperm defects and a proposal for a new classification of the bull spermiogram (author's transl). *Nord Vet Med* 1973; **25**:383–91.
23. Moench GL. A consideration of some of the aspects of sterility. *Am J Obstet Gynecol* 1927; **13**:334–45.
24. Blom E, Birch-Andersen A. An 'apical body' in the Galea Capitis of the normal bull sperm. *Nature* 1961; **190**:1127–8.
25. Saacke RG, Marshall CE. Observations on the acrosomal cap of fixed and unfixed bovine spermatozoa. *J Reprod Fertil* 1968; **16**:511–4.
26. Blom E, Birch-Andersen A. The ultrastructure of a characteristic spermhead-defect in the boar: the SME-defect. *Andrologia* 1975; **7**:199–209.
27. Bane A, Nicander L. Pouch formations by invaginations of the nuclear envelope of bovine and porcine sperm as a sign of disturbed spermiogenesis. *Nord Vet Med* 1965; **17**:628–32.
28. Saeki Y. Crooked-necked spermatozoa in relation to low fertility in the artificial insemination of fowl. *Poult Sci* 1960; **39**:1354–61.
29. Blom E. A new sperm defect "pseudo-droplets" in the middle piece of the bull sperm. 1968.
30. Blom E. A new sterilizing and hereditary defect (the 'dag defect') located in the bull sperm tail. *Nature* 1966; **209**:739–40.
31. Blom E. A rare sperm abnormality: 'Corkscrew-sperms' associated with sterility in bulls. *Nature* 1959; **183**:1280–1.
32. Theilgaard Lassen J, Fly Kragh M, Rimestad J *et al.* Development and validation of deep learning based embryo selection across multiple days of transfer. *Sci Rep* 2023; **13**:4235.
33. Benyon D. *Designing User Experience*. England: Pearson, 2019.

Biology Methods and Protocols, 2025, 10, 1–8
https://doi.org/10.1093/biomethods/bpaf029
Teaching New Methods