An assessment of the subjectivity of sperm scoring

Shanan S. Tobe[1,2*], Lynn Dennany[1] and Marielle Vennemann[1,3]

1. Centre for Forensic Science, University of Strathclyde, Glasgow, UK
2. Present address: School of Biological Sciences, Flinders University, Adelaide, Australia
3. Present address: Institute of Legal Medicine, Medical School Hannover, Hannover, Germany

*shane.tobe@flinders.edu.au, +61 (0)8 8201 2132 (T), School of Biological Sciences, Flinders University, Adelaide, South Australia, Australia, 5001

Abstract

Current histological investigation of vaginal swabs after alleged sexual assault includes the scoring of spermatozoa (0, + to ++++) and the recording of visible tails. It is a method that is universally employed. Despite this method being used for 40 years, there has never been a study investigating its suitability for forensic science. Here, we investigate the reproducibility and subjectivity of sperm scoring among different investigators.

Dilutions of seminal fluid were randomly distributed onto 20 slides, stained with haematoxylin/eosin and assessed by 37 investigators, over two years. Slides were assessed for levels of spermatozoa and the presence of tails.

Each slide was scored by a minimum of 25 investigators. On no slide was there a consensus between all scores. Standard deviation remained below 1, but relative standard deviation (RSD) ranged from 6 – 105 % in a positive correlation as the average score decreased. Spermatozoa were not observed 56 times (9.6 %) and 27 investigators (73 %) did not observe spermatozoa on at least one slide. Spermatozoa with tails were observed on every slide by at least 10 examiners, but as the average score of the slide decreased, so did the observation of tails.

The current sperm scoring method is highly subjective with a particularly high % RSD in slides with low overall sperm counts. Moreover, the recording of tails does not add value to the current technique of sperm scoring. Further research might improve the objectivity of sperm scoring and the reliability of recording of tails.

Keywords: Sperm scoring; sexual assaults; spermatozoa; seminal fluid

Introduction

Semen is a body fluid that is regularly encountered in forensic casework. Identification of seminal fluid as well as the characterisation of spermatozoa is of major importance for the investigation of sexual assault cases.

Since most of the methods currently used for semen identification are presumptive in nature [1], most forensic laboratories use subsequent microscopy to confirm initial findings and classify the number of observed spermatozoa for presentation in reports and in court.

The most widely used method to classify the presence of spermatozoa after visualisation was developed in 1974 by Davies and Wilson [2]. Their method employs a series of pluses (+) to record the number of observed spermatozoa. The designations used are:

++++, many spermatozoa in every field;

+++, many or some spermatozoa in most fields;

++, some spermatozoa in some fields, easy to find;

+, hard to find, and;

0, no spermatozoa observed.

Even though the designations of the different classifications can be slightly different for each laboratory the overall method of assessment is the same. The definition of few and many is subjective and undefined, but was later amended to record whether or not tails were observed with the spermatozoa by including a “T” if tails are visible [3, 4].

Despite the subjective nature of the classification system, an international survey of 42 laboratories [5] showed that this method is universally employed by forensic science laboratories investigating allegations of sexual assault. It is therefore concerning that there are no studies either validating or investigating the limitations of this universally used technique. Further, the fact that there are no proficiency tests, in line with other forms of forensic examinations, is additionally concerning. The original study [2] on which all current work is based readily admits to missing data points and so no statistical analysis has ever been undertaken for this type of analysis.

Most subsequent work does not validate the technique but provides summaries of casework data [3, 4, 6, 7]. Such studies show the principle benefit of using microscopy as confirmatory test for the presence of semen, however the main issue with analysing case data in this manner is that the true nature of the samples is unknown [4]. There are no validated data on error rates or misidentifications.

Therefore, we feel that there is an urgent need to bridge this gap in our knowledge of the possibilities and limitations of the sperm scoring approach. Consequently, the aim of this study was to investigate the reliability of sperm scoring and recording of tails across independent investigators with the main focus on inter-individual variation in the assessment of slides containing varying concentrations of fresh human seminal fluid.

Materials and Methods

Sample preparation and staining

Pooled semen samples from anonymous donors with normal sperm counts were obtained from a fertility clinic and serial diluted to 1:5, 1:10, 1:100 and 1:1,000, which corresponds to a ++++, +++, ++ and + classification, respectively. Samples were diluted with PBS containing buccal cells and gently mixed, allowing for consistent levels of epithelial cells in all dilutions. PBS with buccal cells was used to simulate vaginal epithelial cells for the examination. Ten µL of sample were pipetted onto glass slides, heat fixed and stained using haematoxylin and eosin (HE). This created stains of approximate diameter of 6 mm (0.28 cm^2).

Five slides of each dilution were prepared and were randomly numbered from 1 thru 20. A set of reference slides were also provided as standards for the sperm scoring categories. All slides were verified to contain intact spermatozoa by the authors.

Scoring

Investigators were instructed to individually score each slide according to the Davies and Wilson [2] method and also to identify if tails were observed as per Allard [3]. All investigators hold a degree in a scientific subject, extensive experience in general microscopy and were given specific forensic training for several months before participating in this exercise. All participants had similar levels of experience and used identical laboratory equipment for this exercise to avoid inter-laboratory variation. Assessment of slides took place over two years with a total of 37 investigators, 16 in year one and 21 in year two. Each investigator scored as many slides as they could within 1.5 hours and there was no minimal requirement for number of slides to be analysed or a requirement to analyse the slides in numerical order.

Assessment

Results for each slide from all investigators over both years were combined and compared. Average, standard deviation (SD), relative standard deviation (RSD), median, maximum, minimum and Q1-Q3 range were determined for each slide. Investigators were then scored based on their results compared to the averages observed for the entire group.

Results and Discussion

Scoring system

A total of 585 examinations were performed by 37 investigators. The average number of slides analysed by each volunteer was 16. One volunteer analysed eight slides and 20 volunteers were able to assess all 20 slides within the given time. Each slide was analysed by a minimum of 25 volunteers (average 29). Slides were designed to be analysed in less time than normal casework slides by having stains that were much smaller, 0.28 cm^2, as opposed to 19.5 cm^2 for a standard microscope slide (16.9 cm^2 assuming a 1 cm end for labelling). This is approximately 1.5 % the area normally examined and allowed investigators to fully examine the slides in a shorter time. This, under equivalent timings, would equate to 2.5 weeks of 8 hour days (100 hours total) of investigation for a full slide examination.

On no slide was there a score consensus, all showed some variation in the score given (Table 1). Standard deviation remained below 1 + for all slides (0.25 – 0.81 +). The slide with the lowest SD between scores was slide 19. Over 30 comparisons slide 19 had an average grade of 3.93 +'s with a SD of 0.25 +'s. Full analysis of the results including median, maximum, minimum and Q1-Q3 range can be found in Figure 1. The scoring for each slide between different investigators was varied but remained consistent as demonstrated by the low SD. Investigators tended to agree more on classifications with the very high scoring and very low scoring slides. Standard deviation remained below ± 1 + for the entire series of slides, however, the lowest SD values were observed with the highest scoring slides (Figure 2). As the average score for the slides decreased, the % RSD therefore increased in a positive correlation. This increase in the variance of the scores the less spermatozoa present is hardly surprising because these are the most challenging to assess.

The variation between investigators illustrates the subjective nature of the classifications in the Davies and Wilson [2] scoring system; what one may classify as 'many spermatozoa in every field' another may classify as 'some spermatozoa in most fields'. Although some laboratories have modified the method to include count values for each score, such as 'more than 15 spermatozoa in all fields', this still relies on the observer identifying spermatozoa [8].

The investigators who took part in this study did not routinely score slides for spermatozoa, so the resulting trends would be typical for recently trained investigators, however findings in this study are in line with, and are less variable than, a previous inter-laboratory study looking at semen on swabs and cloth samples [9]. The current study showed an average SD of 0.62 +'s (range 0.25 – 0.81 +) and an average RSD of 43.76 % (range 6.34 – 104.88 %) whereas samples provided to active forensic laboratories in the UK and Ireland showed an average SD of 0.65 +'s (range 0.23 – 1.04 +)[1] and an average RSD of 62.37 % (range 7.69 – 152.75 %)[1] [9]. This demonstrates that our data are in line with previously published work, and provide the first instance this variation has been qualified or quantified.

Interpretation of "0" scores

Over a total of 585 examinations, a score of 0 (no spermatozoa) was recorded in 9.57 % of observations. Although this percentage appears low, it means that of the 37 investigators, 27 (73 %) did not observe spermatozoa on at least one slide, 14 (38 %) of which did so in more than one instance. Zero scores were only observed for slides with an average score less than ++, except for slide 11 which has an average score of 2.2 +'s, but for which a 0 was scored once (Table 1). This is somewhat expected as the fewer spermatozoa that are present, the more likely an examiner is to fail to identify them and is in line with previous work [9].

The relatively high proportion of investigators who were unable to identify any spermatozoa is a cause for concern. Previous reviews of case data found that despite other evidence of sexual assault (such as genital injuries) and victim information regarding ejaculation, spermatozoa were classified in less than 50 % of the samples analysed [6, 7, 10]. The identification of spermatozoa can provide an indication of where the male DNA in a sample originates and is conclusive proof of ejaculation, but is not always found in sexual assault cases. Male DNA is, however, often present in cases of sexual

[1] Generated from data in [9].

assault where spermatozoa were not found [11]. This study shows that there is an increased chance of misclassifying samples as being free of spermatozoa when there is a low number present and supports the continued genetic testing of samples even if classified as "spermatozoa free". In such difficult cases the use of immunohistochemical stainings with fluorescent labels, such as the sperm Hy-Liter system [12] could be useful to distinguish between "very few" spermatozoa and a true negative result.

Detection of intact tails

Spermatozoa with tails were observed on every slide by a minimum 10 individuals (slide 4) and a maximum of 32 (slide 13), on average tails were observed by 22.3 individuals per slide. The two highest scoring slides showed 100 % of investigators observing tails, slides 19 and 6, as did the fourth highest scoring slide, 13. The ratio between the number of investigators who observed spermatozoa with tails compared to the total number that observed spermatozoa can be found in Table 1. Although all slides were prepared at the same time and were verified to contain intact spermatozoa, up to 62 % (Slide 10) of participants failed to identify spermatozoa with tails. This is somewhat in line with previous work, where at least one participant (average 35 %) observed tails on slides examined with dilutions ranging from neat to 1:1,000, whereas for dilutions from 1:2,000-10,000 a consensus of no tails was observed [9]. This does not hold with conventional forensic understanding that time since intercourse can be indicated by the lack of tails, but agrees with an early study by Silverman and Silverman, who did not find any correlation between the time since intercourse and the proportion of spermatozoa with tails [13]. The Silverman and Silverman study [13] was not based on forensic samples, but was a clinical study. It is, however, still a strong indicator that the identification of tails may not be a suitable feature in forensic investigations to indicate time since intercourse. The variability within the observation of tails, even on higher scoring slides lends to this conclusion and has also been observed previously [9].

Conclusions

This is the first study to investigate in detail the subjectivity and applicability of both sperm scoring using the Davies and Wilson [2] method and also the recording of tails as per [3, 4]. The findings of this study are relevant for practicing forensic scientists and are of specific concern to casework situations. The currently used methodology of sperm scoring was found to be highly subjective with SD ranging from 0.25 to 0.81 (6 to 105 % RSD). Higher scoring slides showed more agreement between individual examiners than did lower scoring slides. Particularly in samples with low sperm count scores the variance is high. The high percentage of investigators (73 %) who misclassified at least one slide as being spermatozoa free is concerning because in many sexual assaults, only very few, if any, spermatozoa are being found in vaginal swabs.

According to the findings in this study, the identification of tails does not provide the significance that has been previously reported in some cases. Further investigation is recommended into this feature of sperm scoring, but we do not feel that at present it can be confidently applied to forensic casework.

There are no proficiency tests established or available for sperm scoring. Many organisations such as GEDNAP and ASCLD have proficiency testing in place for DNA (including sexual assault case simulations) and body fluid identification (including semen). Laboratories may undertake sperm scoring as part of these proficiencies, but sperm scoring is not an assessed aspect of those tests. Alternatively, laboratories may develop and employ their own testing for sperm scoring, but these tests are not independently assessed, are non-standardised, not published (either the methodology or the results) and are not an accepted or accredited proficiency test. The results of this study and that of Allard *et al*. [9], indicate that some form of formal and accepted proficiency should be developed for this type of analysis as there is for DNA, body fluids, drugs and fingerprints.

References

[1] Vennemann M, Scott G, Curran L, Bittner F, Tobe SS. Sensitivity and specificity of presumptive tests for blood, saliva and semen. Forensic science, medicine, and pathology. 2014;10:69-75.

[2] Davies A, Wilson E. The persistence of seminal constituents in the human vagina. Forensic Science. 1974;3:45-55.

[3] Allard JE. The collection of data from findings in cases of sexual assault and the significance of spermatozoa on vaginal, anal and oral swabs. Science and Justice. 1997;37:99-108.

[4] Willott G, Allard J. Spermatozoa—their persistence after sexual intercourse. Forensic science international. 1982;19:135-54.

[5] Fourney RM, DesRoches AN, Buckle JL. Chapter 14 - Bilogical Evidence and Forensic DNA Profiling. In: Nic Daeid N, Houck MM, editors. Interpol's Forensic Science Review. London: CRC Press; 2007. p. 591-662.

[6] Grossin C, Sibille I, Lorin de la Grandmaison G, Banasr A, Brion F, Durigon M. Analysis of 418 cases of sexual assault. Forensic Science International. 2003;131:125-30.

[7] Riggs N, Houry D, Long G, Markovchick V, Feldhaus KM. Analysis of 1,076 cases of sexual assault. Annals of emergency medicine. 2000;35:358-62.

[8] Dziak R, Parker L, Collins V, Johnston S. Providing evidence based opinions on time since intercourse (TSI) based on body fluid testing results of internal samples. Canadian Society of Forensic Science Journal. 2011;44:59-69.

[9] Allard JE, Baird A, Davidson G, Jones S, Lewis J, McKenna L, et al. A comparison of methods used in the UK and Ireland for the extraction and detection of semen on swabs and cloth samples. Science & Justice. 2007;47:160-7.

[10] Young WW, Bracken AC, Goddard MA, Matheson S. Sexual assault: review of a national model protocol for forensic and medical evaluation. Obstetrics & Gynecology. 1992;80:878-83.

[11] Sibille I, Duverneuil C, Lorin De La Grandmaison G, Guerrouache K, Teissiere F, Durigon M, et al. Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Science International. 2002;125:212-6.

[12] De Moors A, Georgalis T, Armstrong G, Modler J, Frégeau CJ. Sperm Hy-Liter™: An effective tool for the detection of spermatozoa in sexual assault exhibits. Forensic Science International: Genetics. 2013.

[13] Silverman EM, Silverman AG. Persistence of spermatozoa in the lower genital tracts of women. JAMA. 1978;240:1875-7.

Figure 1: The results of the sperm scoring analysis. Median values are shown as the bold line, the Q1-Q3 range is shown as the bar, and max and min results are shown by the error bars. Slide number is provided on the x-axis with the number of individual analyses for that slide in parentheses. The y-axis scale corresponds to the + system [2], where 0 = 0, 1 = +, 2 = ++, 3 = +++ and 4 = ++++.

Figure 2: The average score (red – left axis), standard deviation (green – left axis) and the percent relative standard deviation (% RSD, purple – right axis) plotted from minimum to maximum % RSD. Slides are ordered from highest to lowest % RSD.

Table 1: The slide number; average score; standard deviation; percentage of observed slides with tails over the total observations (% tails); number of 0 scores; the percentage of observed slides with tails over the total observations greater than 0 (% tails positive count), and; the number of observations per slide. The table has been sorted by average score. Slides that had a 0 score have been highlighted.