Intra- and inter-laboratory variability in the assessment of sperm morphology by strict criteria: impact of semen preparation, staining techniques and manual versus computerized analysis

Gerardo Barroso¹, Ramazan Mercan¹, Kemal Ozgur¹, Mahmood Morshedî³, Paul Kolm², Kevin Coetzee³, Thinus Kruger³ and Sergio Oehninger¹,⁴

¹Department of Obstetrics and Gynecology, ²Bio-statistics, The Jones Institute for Reproductive Medicine, Eastern Virginia Medical School, 601 Colley Avenue, Norfolk, Virginia, 23507, USA and ³Reproductive Biology Unit, Department of Obstetrics and Gynecology, Tygerberg Hospital, University of Stellenbosch, Capetown, South Africa
⁴To whom correspondence should be addressed

We designed prospective studies to compare manual and computerized analysis of sperm morphology by strict criteria using different semen processing and staining techniques. A total of 54 semen samples were studied; slides were prepared from each subject from liquefied semen and after washing, and stained with Diff-Quik® or Papanicolaou. An intra-laboratory, blind performance was performed manually (two observers) and using a computerized analyzer (two readings). This demonstrated a very good correlation between manual analysis of liquefied and washed samples with both staining techniques [intraclass coefficient (ICC) = 0.93 and 0.83]. Greater agreement was observed between computerized readings (washed samples) of Diff-Quik® (ICC = 0.93) than of Papanicolaou-stained slides (ICC = 0.66). An excellent intra-laboratory correlation was observed for within-computer readings (ICC = 0.93). There was moderate agreement between inter-laboratory computer readings (two centres, ICC = 0.72). Although there was lower inter-laboratory agreement for manual and manual versus computer readings, overall results of all manual and computer analyses showed good agreement (ICC = 0.73). Diff-Quik® staining is reliable for both manual (liquefied) and computer (washed) analysis of strict sperm morphology. Intra- and inter-computer analyses using this method reached satisfactory levels of agreement. There is still high inter-laboratory variability for the manual method.

Key words: inter-laboratory variability/intra-laboratory variability/normal sperm morphology/staining/strict criteria

Introduction
The assessment of sperm morphology, an integral component of the basic semen evaluation, has been demonstrated to aid the clinician in the management of male infertility (Oehninger and Kruger, 1995). It is clear that the evaluation of a single sperm feature or function may not provide enough power for the prediction of fertilization or implantation outcome. This reflects the complexity and multiplicity of events leading to sperm–oocyte interaction and conception. Notwithstanding these considerations, sperm morphology assessed by strict criteria (Kruger et al., 1986) has been shown by multiple authors to have a high predictive value not only for the outcome of advanced assisted reproductive technologies [in-vitro fertilization (IVF) and gamete intra-Fallopian transfer (GIFT)] but also for those of intrauterine insemination and in-vivo reproduction (Kruger et al., 1988; Oehninger et al., 1988; Hinting et al., 1990; Enginsu et al., 1991; Kobayashi et al., 1991; Check et al., 1992; Eggert-Kruse et al., 1993; Grow et al., 1994; Ombelet et al., 1994; Toner et al., 1994; Grow and Oehninger, 1995; Coetzee et al., 1998; Duran et al., 1998). A few studies, however, have not supported those findings (Coates et al., 1992; Matorras et al., 1995; Karabinus and Gelety, 1997).

The examination of sperm morphology by more standardized and stringent criteria has enhanced objectivity and decreased intra-laboratory variability (Menkveld et al., 1990). The World Health Organization (WHO, 1992) has also recommended that stricter criteria should be applied when assessing the morphological normality of the spermatozoon. This has led to the establishment of lower threshold levels for normality (WHO, 1992).

Despite these efforts, a large inter-laboratory variation still exists in the examination of sperm morphology (Comhaire et al., 1994; Ombelet et al., 1997a, 1998). Adequate technician training is of paramount significance to achieve consistent results within a given laboratory. Even when strict criteria (Kruger et al., 1986; Menkveld et al., 1990) are utilized, inter-laboratory variation is probably the result of various factors including: (i) different semen and smear preparation techniques, (ii) differences in interpretation, and (iii) technician experience (Coetzee et al., 1999).

Semen assessment by automated means has the potential to avoid the biases, subjectivity and consequent lack of intra- and inter-laboratory reproducibility inherent in the visual methodologies. Although computerized systems have been available for a few years, problems in their development have not yet allowed for their routine use (Wang et al., 1991; Davis et al., 1992; Davis and Gravance, 1994; Kruger et al., 1995; Hofmann et al., 1996). Use of the ‘signature’ method (Kruger et al., 1993, 1995, 1996) has enabled a satisfactory level of agreement between manual (visual, optical microscopic) and computerized (incorporating strict criteria into the Hamilton Thorne Research IVOS dimensions sperm analyser) method of evaluation.

We have designed prospective, blind studies to compare
manual and computerized analysis of sperm morphology by strict criteria using different semen processing and staining techniques. Our specific aims were (i) to perform an intra-laboratory comparison of morphological examination of liquefied versus washed semen, Papanicolaou versus Diff-Quik® staining techniques, and manual versus computerized analyses; and (ii) to carry out an inter-laboratory (two centres) assessment of sperm morphology comparing liquefied versus washed semen and manual versus computerized analysis. The main objective of these studies was to confirm the accuracy and reproducibility of the existing methods used to judge this important sperm feature through the application of strict criteria.

Materials and methods
The studies received the approval of the Institutional Review Boards at Eastern Virginia Medical School and University of Stellenbosch. A total of 54 ejaculates from eight fertile men (donors currently enrolled in the artificial insemination programme) and 46 subfertile patients (members of infertile couples presently undergoing evaluation) were randomly selected and examined. Semen samples were obtained by masturbation after 3–5 days of sexual abstinence.

Examination of liquefied semen
Following liquefaction, a basic determination of sperm concentration and percentage of progressive motility was performed using an automated semen analyser (Cell Soft®; Cryoresources, NY, USA) and (inter-laboratory manual assessment variability, ICC = 0.57). In addition, a very high degree of agreement between the results obtained by manual analysis of liquefied and washed semen samples (ICC = 0.93). In addition, a very high level of agreement was observed for the computer analysis of washed samples (within-computer variability, ICC = 0.93). The inter-laboratory comparisons (Norfolk and Tygerberg laboratories) demonstrated a relatively low level of agreement between the manual readings obtained from the two centres (inter-laboratory manual assessment variability, ICC = 0.57).
Table I. Indices of agreement for Diff-Quik® stained slides in Norfolk (N) and Tygerberg (T) analysed by computer (C) or manually (M) (n = 54)

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean ± SD</th>
<th>Min.²</th>
<th>Max.²</th>
<th>Mean difference</th>
<th>95% CI for ICC difference</th>
<th>ICC ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-M liquefied-1</td>
<td>8.78 ± 3.65</td>
<td>2</td>
<td>17</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>N-M liquefied-2</td>
<td>8.63 ± 3.52</td>
<td>2</td>
<td>15</td>
<td>0.15</td>
<td>-5.69:5.97</td>
<td>0.82</td>
</tr>
<tr>
<td>N-M washed</td>
<td>9.41 ± 3.64</td>
<td>2</td>
<td>15</td>
<td>-0.22</td>
<td>-3.88:3.42</td>
<td>0.93</td>
</tr>
<tr>
<td>N-C washed-1</td>
<td>15.83 ± 8.59</td>
<td>2</td>
<td>37</td>
<td>-0.22</td>
<td>-3.88:3.42</td>
<td>0.93</td>
</tr>
<tr>
<td>N-C washed-2</td>
<td>15.80 ± 8.04</td>
<td>2</td>
<td>35</td>
<td>0.04</td>
<td>-8.40:8.48</td>
<td>0.93</td>
</tr>
<tr>
<td>N-M liquefied</td>
<td>9.29 ± 3.23</td>
<td>2</td>
<td>17</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>T-M liquefied</td>
<td>16.15 ± 7.27</td>
<td>2</td>
<td>30</td>
<td>-6.62</td>
<td>-18.91:5.20</td>
<td>0.57</td>
</tr>
<tr>
<td>N-C washed</td>
<td>15.62 ± 8.03</td>
<td>2</td>
<td>34</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>T-C washed</td>
<td>18.53 ± 8.27</td>
<td>2</td>
<td>41</td>
<td>-2.01</td>
<td>-17.00:12.98</td>
<td>0.72</td>
</tr>
<tr>
<td>N-M liquefied</td>
<td>9.18 ± 3.45</td>
<td>2</td>
<td>17</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>N-C washed</td>
<td>15.81 ± 8.05</td>
<td>2</td>
<td>34</td>
<td>-6.64</td>
<td>-21.09:7.82</td>
<td>0.48</td>
</tr>
<tr>
<td>T-M liquefied</td>
<td>16.11 ± 7.35</td>
<td>2</td>
<td>30</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>T-C washed</td>
<td>18.74 ± 8.35</td>
<td>2</td>
<td>42</td>
<td>-2.64</td>
<td>-18.20:12.93</td>
<td>0.66</td>
</tr>
<tr>
<td>N-M washed</td>
<td>9.41 ± 3.64</td>
<td>2</td>
<td>15</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>N-M washed</td>
<td>9.18 ± 3.45</td>
<td>2</td>
<td>17</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>N-C washed</td>
<td>15.81 ± 8.05</td>
<td>2</td>
<td>34</td>
<td>0.15</td>
<td>-5.46:5.75</td>
<td>0.82</td>
</tr>
<tr>
<td>All manual liquefied</td>
<td>11.37 ± 4.49</td>
<td>2.2</td>
<td>22</td>
<td>-6.62</td>
<td>-20.74:12.93</td>
<td>0.72</td>
</tr>
<tr>
<td>All computer washed</td>
<td>16.42 ± 7.45</td>
<td>2.5</td>
<td>37.3</td>
<td>-6.62</td>
<td>-16.38:6.27</td>
<td>0.73</td>
</tr>
</tbody>
</table>

CI = confidence interval; ICC = intraclass correlation coefficient. ²Range of normal morphology values observed for all slides evaluated per category.

On the other hand, a superior degree of agreement was observed for the computerized readings obtained from the two centres (inter-laboratory computer assessment variability, ICC = 0.72).

For both centres, there was only moderate agreement between manual (liquefied semen) and computerized measurements (washed semen) (Norfolk, ICC = 0.48; Tygerberg, ICC = 0.66). This was also the case for results for washed semen assessed manually and by computer (Norfolk, ICC = 0.51). Overall, however, there was good agreement between all manual measurements of liquefied semen and all computerized analyses of washed samples performed in both centres (ICC = 0.73).

Table II presents the results of the comparisons of Papanicolaou and Diff-Quik® staining for liquefied and washed samples, manual and computer assessments, performed at Norfolk (n = 21). There was a high degree of agreement between results from manual analysis of liquefied and washed samples for Papanicolaou-stained samples (ICC = 0.83). There was only moderate agreement for results from the manual analysis of Diff-Quik® and Papanicolaou-stained liquefied samples (ICC = 0.69) or washed samples (ICC = 0.66). There was moderate agreement for the computerized readings of Papanicolaou-stained slides (ICC = 0.66) and lower agreement for the computer results of Papanicolaou and Diff-Quik® stained slides (ICC = 0.51).

Discussion

These studies aimed to compare and optimize the methods and techniques presently used to assess sperm morphology using strict criteria. Liquefied semen is the source of spermatozoa generally used to examine the percentage with normal morphology. Semen washing, on the other hand, has been recommended to ensure a high-quality slide with negligible background staining in order to enhance the accuracy of the computer readings (Kruger et al., 1996; Menkveld et al., 1997). The two more widely utilized stains are the classic...
Papanicolaou and the more recently introduced Diff-Quik® method (Coetzee et al., 1998). Results of the intra-laboratory evaluation (Norfolk) demonstrated a very high degree of agreement for manual analysis between liquefied and washed samples for both staining techniques. The level of agreement was lower, however, when the two staining techniques were compared in the liquefied and in the washed samples. An excellent correlation for morphology assessment has been reported using Papanicolaou and Diff-Quik® stains in both unwashed and washed samples (Menkveld et al., 1997).

For the computer readings (washed samples), a very high degree of agreement was observed for the Diff-Quik® stain, substantially higher than was observed for Papanicolaou stain. Slight variations have been reported (Coetzee et al., 1997; Menkveld et al., 1997) in the percentage with normal morphology assessed using Diff-Quik® and Papanicolaou staining methods. It was concluded that the washed Diff-Quik® smears provided an optimal preparation method for computerized sperm morphology evaluation, comparing favourably with manual examinations.

The inter-laboratory comparisons (two centres) revealed, at best, only a moderate level of agreement between the manual readings obtained. Even when the same semen processing (liquefied sample), staining technique (Diff-Quik®) and classification system (strict criteria) are used, there are still differences in sperm morphology assessments between laboratories. This fact has been elegantly documented (Ombelet et al., 1997b, 1998) by a cluster analysis of the methodology and results of sperm morphology assessment in 86 centres using only one method of slide preparation and evaluation. We concur with the quality control proposal put forward by those authors. These strategies should include: (i) the advancement of reliable computerized methods; (ii) standardization of the methodology for sperm morphology assessment; (iii) geographical centralization of the assessment of sperm morphology; and (iv) reference population studies for all laboratories (Ombelet et al., 1998).

Here, a very low intra-computer variability (for one centre) was observed. Also, a relatively good level of agreement was found for the computerized readings obtained at the two centres. Overall, there was moderate agreement between all results from manual assessments of liquefied semen and computerized analysis of washed samples. Therefore, it is apparent that trained andrologists and computer analysis can determine sperm morphology similarly over a wide range of percentage normal forms (Hofmann et al., 1996; Kruger et al., 1996). Although improvements related to speed and overall assessment of anomalies (different types and locations within the cells) are warranted, the use of the computer seems to offer clear advantages over the manual method in terms of improved accuracy, repeatability and objective evaluation. An adequate staining resulting in good contrast, improved cell focusing and high repeatability on a cell by cell and slide by slide basis will permit a more universal and daily clinical use of this equipment (Kruger, 1995). In a recently performed multi-centre study (Coetzee et al., 1999) the magnitude of the variance produced by computer readings reached the same level as the manual evaluation and a <10% coefficient of variation was obtained if the correct quality control measures were implemented. The computerized assessment of strict morphology has therefore the potential to become the standard method to examine this basic sperm parameter.

We conclude that Diff-Quik® staining is a fast, simple and reliable method for both manual (liquefied) and computer (washed) examinations of strict sperm morphology. Intra- and inter-computer readings of washed/Diff-Quik®-stained samples demonstrated satisfactory levels of agreement and consequently offer a very promising means for the examination of this sperm feature. The high inter-laboratory variability observed for the manual method requires immediate implementation of corrective strategies.

References


Received on February 9, 1999; accepted on April 29, 1999