REPRODUCIBILITY AND RELIABILITY OF REPEATED SEMEN ANALYSES IN MALE PARTNERS OF SUBFERTILE COUPLES

Esther Leushuis
Jan Willem van der Steeg
Pieternel Steures
Sjoerd Repping
Patrick M.M. Bossuyt
Marinus A. Blankenstein
Ben Willem J. Mol
Fulco van der Veen
Peter G.A. Hompes

FERTILITY AND STERILITY 2010 DEC;94(7):2631-5
ABSTRACT

OBJECTIVE
To determine the precise degree of variability that is represented by the reproducibility and reliability of semen analysis. The general assumption is that semen analyses need to be repeated because of a high degree of within-individual variability. However, the precise degree of variability is not well established in male partners of subfertile couples.

DESIGN
Retrospective cohort study.

SETTING
Two university hospitals in the Netherlands, which routinely perform two semen analyses in the male partner of subfertile couples.

PATIENTS
Male partners of subfertile couples.

MAIN OUTCOME MEASURES
We assessed the test-retest reproducibility, by calculating the coefficient of variation ($CV_w$) for five semen parameters. The $CV_w$ expresses, on a relative scale, the degree of closeness of repeated measurements taken in the same subject. We also estimated the reliability of these semen parameters, in terms of the intraclass correlation coefficient, which express the ratio of the between-subject variability over the total variability.

RESULTS
We analysed the data of 5,240 men and found that the $CV_w$ of all semen parameters ranged from 28% to 34%. The intraclass correlation coefficients of these semen parameters were moderate to high: volume: 0.70; concentration: 0.89; motility: 0.58; morphology: 0.60; total motile count: 0.73.

CONCLUSIONS
This study affirmed the presumed large within-subject variability and the limited reproducibility of semen analyses in subfertile men. Whether this degree of variability within men justifies one or more repetitions of the semen analysis in view of consequences for clinical management should be the topic of future studies. Until then it seems reasonable to perform two semen analyses.
INTRODUCTION

Semen analysis is the cornerstone of the laboratory evaluation of the male partner of subfertile couples.\(^1\)\(^2\) It is well known that semen analyses of the same man vary from one ejaculate to the other.\(^3\)\(^-\)\(^7\) The results of repeated semen analyses can vary in principle because of three factors: (1) preanalytic influences, such as duration of abstinence or seasonality; (2) analytic variation in the method of analysis and/or the evaluator; and (3) inherent biologic variability.\(^5\)\(^8\)\(^-\)\(^10\)

The majority of reports on the reproducibility and reliability of the semen analysis involve only healthy men, but it may well be that the variability in a subfertile population differs from that in the general population.\(^6\) This is an important distinction, because the subfertile population is the target group for performing a semen analysis during the fertility workup. One of the earliest reports on the quantification of the variability of semen parameters within men from a subfertile population did not take agreement beyond chance into account.\(^10\) Another study reported high variability of semen parameters within one man in terms of coefficients of variation, but these samples were collected specifically for IUIs. The authors of the latter study also reported that the fluctuations in the samples of one man were substantially smaller than the fluctuations between men.\(^9\)

In an attempt to achieve a more reliable approximation of the true values of individual semen parameters, several guidelines recommend to repeat the semen analysis once or twice in the basic fertility workup.\(^3\)\(^,\)\(^11\)\(^,\)\(^12\) Yet, the first step in deciding whether or not to repeat the semen analysis in the basic fertility workup would be to assess the actual degree of within-subject variability that is represented by the reproducibility and reliability in male partners of subfertile couples. This was the aim of the study reported in this article, which was based on two routinely performed semen analyses in a large cohort of male partners of subfertile couples.

MATERIALS AND METHODS

Patients
We searched the consecutive electronic records of male partners of subfertile couples at two university hospitals in the Netherlands (Vrije Universiteit medical center [VUmc] and Academic Medical Center [AMC], Amsterdam). For one hospital (AMC) records were available of men visiting the hospital between January 1998 and 2008. For the other hospital (VUmc) data were available for men between January 2004 and August 2008. In all men, two semen analyses had been performed routinely as part of the basic fertility workup. Morphology was registered in only one center (AMC). Because the data were collected during standard patient treatment and were handled anonymously in our study, no Institutional Review Board approval was required according to Dutch law.

Semen analysis
Semen analyses were performed according to the World Health Organization (WHO)
Men were advised to maintain 3 days of sexual abstinence before semen analysis. The interval of the time of abstinence was recorded when the sample was handed in. Semen samples were collected at the hospital or at home by masturbation directly into a 50-mL polyethylene jar.

Semen samples were analyzed within 1 hour after ejaculation. After liquefaction, semen volume was measured in a graded tube with 0.1-mL accuracy. Sperm concentration was counted and motility was assessed in a Makler counting chamber at a magnification of x200. Sperm concentration, motility and morphology assessment were performed by trained laboratory technicians. Total motile count (TMC) (prewash) was calculated as semen volume multiplied by sperm concentration multiplied by the percentage of sperm with progressive motility (WHO grade A).³

**Data analysis**

We calculated the mean difference between the two analyses for all semen parameters and the associated 95% confidence intervals (CI), to see whether there was a systematic difference between measurements. We first calculated the reproducibility of semen parameters by evaluating the degree of within-subject variability between the two analyses. We did so by calculating within-subjects coefficient of variation (CVₔ) for each parameter. The CVₔ expresses the within-subjects SD relative to the subject’s true score.¹³ It expresses, on a relative scale, the degree of closeness of repeated measurements taken in the same subject, on different occasions under the same conditions. CVₔ values close to zero indicate better reproducibility and less variability within subjects. With a CVₔ of 0, repeated measures are always identical and no within-subject variability exists. As not all semen parameters are normally distributed, we applied a natural logarithmic transformation to sperm concentration and TMSC and a square root transformation to sperm morphology. In addition to the CVₔ, we calculated the probability (p) that the two semen analyses of one man differ by a factor (k) or more given a certain CVₔ, to provide a more clinically applicable statistic (Figure 1).¹⁴

*Figure 1. This formula (1) calculated the probability that two semen analyses differ by a factor k or more for a specified value of the CV. For example (2), the frequency of two semen analyses that differ by 10% or more (so that k = 1.1) with a CV of 30% is 45.*

\[
p(k) = 2\Phi\left\{\frac{-\log_e (k)}{\sqrt{2\log_e (CV^2 + 1)}}\right\}
\]

\[
p(1.1) = 2\Phi\left\{\frac{-\log_e (1.1)}{\sqrt{2\log_e [(30/100)^2 + 1]}}\right\} = 0.45
\]
We also estimated the implications of the within-subject variability for the reliability of semen analysis parameters, the extent to which these parameters can be used to distinguish between men. For this purpose, we calculated the intraclass correlation coefficient for each parameter (ICC). The ICC expresses the proportion of the total variability that can be attributed to the genuine variability between men, the remainder being within-subject fluctuations and random error. This means that, where the CVW looks at within-subject fluctuations only, the ICC looks at the consequences of these fluctuations—limiting reproducibility—for distinguishing between men. The ICC takes values in the range 0 to 1 with values close to 1 indicating that the contribution of the within-subject fluctuations is relatively small compared with the genuine differences between men. With a perfect ICC of 1 all differences in semen parameters between men are genuine, and a repeated semen analysis will be identical to the first semen analysis of that man. With an ICC of 0, all differences between two semen analyses within a man are due to within-man fluctuations or error. The ICC is interpreted similar to \( \kappa \), with values >0.80 indicating outstanding reliability, and values between 0.60 to 0.79 presenting substantial reliability.

The agreement between repeated semen analyses was also presented graphically with use of scatter plots. The cutoff values according to the WHO reference values were displayed in the plots with vertical lines on the reference value for the first semen analysis and a horizontal line on the second semen analysis. The number of concordant or discordant pairs was reported in percentages of the total cohort. In addition to obtaining measures of within-subject fluctuations, we examined a potential determinant of these. The effect of the duration of abstinence on the different semen parameters was assessed by a repeated-measurements procedure with use of mixed-effects models. Calculations were performed with use of SPSS 16.0 program (SPSS Inc., Chicago, IL) and SAS System 9.1 (SAS Institute, Cary, NC).

RESULTS

We analysed the data of 5,240 men with two routinely performed semen analyses: 2,941 from the AMC and 2,299 from the VUmc. The results for the first and second semen analysis and the average differences between the two samples are summarized in Table 1. The mean male age at the first semen analysis was 36.3 years (5\textsuperscript{th} to 95\textsuperscript{th} percentile: 27 to 47 years). The mean number of days of abstinence for the first semen analysis was 4 days (5\textsuperscript{th} to 95\textsuperscript{th} percentile: 2 to 7 days). The median number of days between the two semen analyses was 19 days (5\textsuperscript{th} to 95\textsuperscript{th} percentile: 13 to 292 days). There was a significant but small difference between the semen analyses for semen volume, and sperm motility (all grades). No differences between the two semen analyses were found for the other semen parameters.

The CVW were comparable and high for all semen parameters, ranging from 28% to 34% (Table 2). The probability to detect a difference of 10% or more between two semen analyses for the same man was 0.45, given the corresponding CVW of 30%. There was more variability in the corresponding ICC. The ICC for sperm concentration was high (ICC: 0.89), indicating excellent reliability despite its limited reproducibility.
Table 1. Characteristics of the first and second semen analysis.

| Semen parameters | First semen analysis | | | | Second semen analysis | | | | Systematic differences between the two semen analyses | |
|------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|                  | n   | mean | 5th | 95th | n   | mean | 5th | 95th | n   | mean | 5th | 95th | n   | mean | 5th | 95th |
| Volume (mL)      | 5,240 | 3.3  | 1.0 | 6.2  | 5,240 | 3.3  | 1.0 | 6.2  | 5,240 | -0.04 | -0.07 | 0 |
| Concentration (10^6/mL) (median) | 5,240 | 38.0 | 4   | 71  | 5,240 | 38.7 | 5.2 | 71  | 5,240 | -0.72 | 0.22 | 1.22 |
| Motility (%)     |                   |                   |                   |                   |                   |
| WHO grade A      | 5,050 | 38.0 | 4   | 71  | 5,040 | 38.7 | 5.2 | 71  | 5,042 | -2.76 | -3.10 | -2.42 |
| WHO grade B      | 5,040 | 15.3 | 2.9 | 33.3 | 5,010 | 12.6 | 0.3 | 31  | 5,005 | -2.10 | -2.62 | -1.57 |
| WHO grade A+B    | 5,050 | 53.2 | 18  | 84.7 | 5,040 | 51.2 | 17  | 80  | 5,042 | -2.10 | -2.62 | -1.57 |
| Morphology (% normal)\(^a\) | 2,730 | 35.0 | 9   | 60  | 2,710 | 35.4 | 9   | 61  | 2,707 | 0.26  | -0.28 | 0.80  |
| TMC\(^b\) (10^6) (median) | 4,870 | 53.2 | 0.83 | 277.3 | 4,870 | 54.4 | 0.94 | 271 | 4,870 | -2.13 | -4.67 | 0.41  |

\(^a\)According to WHO 1999.
\(^b\)Calculated as volume * concentration * motility/100.

Table 2. Intraclass correlation coefficients and CV\(_w\) of semen parameters in two semen samples for each man (N=5,240).

<table>
<thead>
<tr>
<th>Semen parameter</th>
<th>CV(_w)</th>
<th>95% confidence interval</th>
<th>ICC</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lower limit</td>
<td>upper limit</td>
<td>lower limit</td>
<td>upper limit</td>
</tr>
<tr>
<td>Semen volume</td>
<td>28%</td>
<td>27%</td>
<td>28%</td>
<td>0.70</td>
</tr>
<tr>
<td>Sperm concentration</td>
<td>29%</td>
<td>30%</td>
<td>28%</td>
<td>0.89</td>
</tr>
<tr>
<td>Sperm motility</td>
<td>34%</td>
<td>33%</td>
<td>34%</td>
<td>0.58</td>
</tr>
<tr>
<td>Sperm morphology</td>
<td>29%</td>
<td>28%</td>
<td>30%</td>
<td>0.60</td>
</tr>
<tr>
<td>Total motile count</td>
<td>30%</td>
<td>31%</td>
<td>29%</td>
<td>0.73</td>
</tr>
</tbody>
</table>
In contrast, the reliability of sperm motility and morphology was moderate, represented by ICCs of 0.59 and 0.60, respectively. The ICC of semen volume was 0.70 (Table 2). The percentage of concordant results between the two semen analyses according to the WHO reference values between the first and second semen analysis ranged from 61% to 70% and the percentage with discordant results ranged from 6% to 21% (Figure 2).

The duration of abstinence had an effect on the results of all semen parameters (all P values < .003), except on morphology (F value <0.001, P value .99) (Table 3). Increasing duration of abstinence was associated with decreasing sperm motility and morphology, but with increasing semen volume, sperm concentration and TMSC.

**Table 3. The effect of the duration of abstinence and seasonality on the within-subject variability.**

<table>
<thead>
<tr>
<th>Semen parameter</th>
<th>β</th>
<th>df</th>
<th>F</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semen volume</td>
<td>0.081</td>
<td>2,898</td>
<td>92.0</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sperm concentration</td>
<td>0.035</td>
<td>2,898</td>
<td>8.94</td>
<td>0.003</td>
</tr>
<tr>
<td>Sperm motility</td>
<td>-0.393</td>
<td>2,789</td>
<td>13.9</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sperm morphology</td>
<td>-0.001</td>
<td>1,367</td>
<td>0.00</td>
<td>0.990</td>
</tr>
<tr>
<td>Total motile count</td>
<td>0.074</td>
<td>2,707</td>
<td>49.6</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

* A positive β represents an increase in the variability of the semen parameters within men with increasing days of 2 abstinence or seasonality. A negative β represents a negative association.

**DISCUSSION**

In this study, we described the within-individual variability of semen analyses in male partners of subfertile couples in a large retrospective cohort. This study affirmed the presumed large within-subject variability of semen analyses in subfertile men, represented by the limited reproducibility. The low reliability of sperm motility and morphology indicated that the within-subject variability is high relative to the between-subject variability, making it difficult to classify men on the basis of a single measurement. In contrast, the high reliability of sperm concentration provided better classification potential based on a single measurement. The reliability of semen volume was moderate.

The strength of this study is that we analysed an unselected large cohort of male partners of subfertile couples that had two semen analyses routinely performed. Numerous studies have evaluated the within-subject variability of various semen parameters, but no study has assessed this variability in >5,000 male partners of subfertile couples.
Figure 2(a). Scatterplot of the results of the first semen analysis versus the second for semen volume.

Figure 2(b). Scatterplot of the results of the first semen analysis versus the second for sperm concentration.
Figure 2(c). Scatterplot of the results of the first semen analysis versus the second for sperm motility.

Figure 2(d). Scatterplot of the results of the first semen analysis versus the second for sperm morphology.
The two studies that previously reported CV\textsubscript{W} on semen parameters, respectively of men from an IUI program and a retrospective analysis of 74 infertile men and 65 normal men producing five or more ejaculates each, reported CV\textsubscript{W} similar to the ones obtained in our study, arriving at similar conclusions about the high within-subject variability. These CV\textsubscript{W} were 27.1% and 29.6% for semen volume, compared with 28% in our study. Results were also similar for sperm motility with CV\textsubscript{W} of respectively 34.4% and 34.2%, versus 34% in our study. The results on sperm concentration were more contrasting, with CV\textsubscript{W} of 54.2% and 56.8% in the two other studies, compared with 29% in our study. Sperm morphology was not analysed in these two other studies.

These previous studies reported ICCs that were higher than in our study: 0.84 and 0.92 for semen volume, compared to 0.70 in our study; 0.78 and 0.88 for sperm motility, compared with 0.58 in our study; and 0.92 for both studies, compared with 0.89 in our study. The high ICC in these studies might be explained by a more heterogeneous study population compared with our study population. The same semen parameter will have a higher ICC in a heterogeneous population compared with a homogenous population, because the proportion of differences between men will be larger relative to the overall variability.

In our study, all semen parameters, except sperm morphology, were affected significantly by the duration of abstinence. A number of previous studies have investigated the effect of abstinence and seasonality on semen parameters. Although there are reports in the literature on the effect of abstinence on the variability of the semen analysis, the studies differ greatly in aim, study design and population, and the results differ accordingly. The results of our study indicate that it is preferable to standardize duration of abstinence to lower the within-subject variability caused by this confounder and to increase reproducibility and reliability.

Although this study affirms the presumed large variability of semen analyses in male partners of subfertile couples, a problem that could be remedied by repeating the semen analysis, it is not entirely clear how the two semen analyses results, when different, should be combined and interpreted. The ultimate trial to evaluate this would be to correlate combinations of results with capacity to predict spontaneous pregnancy. Until then it seems reasonable to perform two semen analyses.
REFERENCES