Published in final edited form as:

J Urol. 2013 August ; 190(2): 683–688. doi:10.1016/j.juro.2013.02.030.

Reproducibility and Reliability of Semen Analyses in Youths at Risk for Infertility

Matthew S. Christman¹, Kate H. Kraft², Gregory E. Tasian³, Stephen A. Zderic³, and Thomas F. Kolon³

¹Naval Medical Center San Diego, San Diego, CA

²C. S. Mott Children's Hospital/University of Michigan, Ann Arbor, MI

³The Children's Hospital of Philadelphia/University of Pennsylvania, Philadelphia, PA

Abstract

Purpose—There is little normative data on semen analyses in youths at risk for, but who do not present with, infertility. Standard practice amongst infertility specialists includes evaluation of two separate semen samples given the degree of within-subject variability. We hypothesize that males transitioning from pediatric to adult care who are at risk for infertility would similarly have this variability.

Materials and Methods—A retrospective review of patients with a history of cryptorchidism or varicocele who submitted two semen samples for evaluation of fertility potential was conducted. The within-subject coefficient of variation (CV_w) and intraclass correlation coefficient (ICC) were calculated for each semen parameter to evaluate reproducibility and reliability, respectively.

Results—A total of 79 subjects were studied. Mean age $\pm$ SD was 18.8 ± 1.2 years (range 17.8 – 24.7). The CV_w was high for each semen parameter, ranging from 36% for volume and motility to 82% for total motile count (TMC). ICC for a single SA ranged from 0.55 for motility to 0.88 for total count. ICC for TMC was 0.78 [95% CI, 0.67 – 0.85], consistent with substantial reliability.

Conclusions—Although within-patient variability of individual SA parameters was found, overall there was substantial agreement between consecutive semen analyses in this population at risk for infertility -- particularly regarding TMC, which is the most important determinant of fertility from a SA. Therefore, it is possible to appropriately classify some young men based on the result of a single measurement as they transition from pediatric to adult care.

Keywords

reliability; reproducibility; semen analysis; adolescent

Introduction

Cryptorchidism and varicoceles are diagnoses that commonly afflict the pediatric and adolescent male patient. Associations of these disease processes with infertility have been described.¹⁻³ As patients near adulthood, their care may be transitioned to an adult urology practice. In particular, issues related to subfertility may become even more pertinent as the individual matures and looks to start a family.

Corresponding Author: Matthew S. Christman, MD, Department of Urology, Naval Medical Center San Diego, 34800 Bob Wilson Drive, San Diego, CA 92134, (619) 532-7200 (office); (619) 532-7234 (FAX), matthew.christman@med.navy.mil.

Disclosures: The views expressed in this manuscript are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the United States Government.

Semen analysis (SA) is a standard part of the evaluation of an infertile man. Due to the inherent variability of this test within an individual, guidelines have generally recommended two SA for characterization of the individual.⁴ Ample data are available regarding SA in the normal adult population.⁵⁻¹⁰ Data from subfertile adult populations is less robust; and even less is known regarding SA in a peri-adolescent cohort.¹¹⁻¹⁴ To our knowledge, there are no prior studies providing normative data or assessing the variability of SA in youths transitioning from pediatric to adult medical care. The WHO defines “youths” as those individuals aged 15–24 years.

SA in this age population has been used both for clinical decision making and for research purposes to further understanding within the field of pediatric andrology. Having a more thorough understanding of SA will ultimately enable us to provide better care with an improved understanding of how to treat conditions such as cryptorchidism and varicocele. We aimed to determine the reliability and reproducibility of SA in the transitioning male population at risk for subfertility. In other words, we aimed to test the test (i.e., SA). We hoped to answer the question of whether a single or multiple SA in male youths was/were necessary to adequately characterize patients. We hypothesized that this cohort would demonstrate relatively large variability of the SA, as is seen in the adult population.

Materials and Methods

An institutional review board approved retrospective cohort study of patients with a history of surgically corrected cryptorchidism (unilateral or bilateral) or non-operated varicocele was conducted. In our practice, SA is regarded as part of the intermediate and long-term follow-up of each of these diagnoses. Requests to collect a SA from these patients is routinely made around the age of 18 years, once they have reached a Tanner V stage of sexual development. To be included in the study, patients must have submitted two semen samples for evaluation of fertility potential. Patients were excluded if a SA was missing data for an individual parameter, if they underwent surgical intervention during the interval between the first and second SA, or if the abstinence period preceding collection was known to be less than 3 days. In cases where patients had 3 SA, the first 2 consecutive samples were analyzed.

SA data were extracted for semen volume (mL), semen density (million sperm/mL), total sperm count (million sperm), and sperm motility (% motile sperm). From these parameters a value for the total motile count (TMC) was calculated for each sample ($TMC = \text{volume} \times \text{density} \times \% \text{ motility}$). Descriptive statistics of each sample were then calculated. Concordance rates were calculated based on the number of times that the samples agreed with each other (i.e., both normal or both abnormal). To calculate these rates, definitions of normal/abnormal were dichotomized according to WHO standards for volume, density, count, and motility.¹⁵ TMC was defined as abnormal for values less than 20 million/ejaculate; this cutoff has been used to define individuals as candidates for spontaneous pregnancy in infertility clinics.¹⁶

To gauge the precision of a SA in this population we calculated the within-subject coefficient of variation (CV_w), which is the square root of the within-patient mean squares divided by the grand mean, expressed as a percentage.¹² This value illustrates reproducibility as a reflection of the relative closeness of a repeated measure in the same patient. Essentially, it is a standardized form of the standard deviation that has utility in situations where the deviation is proportional to the mean, thus allowing comparison of variability.¹⁷ A Bland-Altman plot was constructed and tested with Spearman correlation to determine whether the deviation of an individual semen parameter was proportional to (dependent on) the mean throughout the entire study population. As Bland and Altman

described, if the deviation is not independent of the mean, CV_w should be used to express measurement error rather than standard deviation.¹⁸ Use of CV_w also allows a relative comparison of the variability between different parameters, such as motility vs. count, which is not possible with standard deviation. If each repeated SA for an individual patient is identical, the CV_w would be zero.

The intraclass correlation coefficient (ICC) was calculated to describe the accuracy of SA as a test. This value helps to demonstrate how reliably the test can differentiate between individuals by expressing the variability within individual patients relative to the variability of the entire population being tested. ICC was calculated according to the method of Shrout and Fleiss.¹⁹ Using a one-way random effects ANOVA model, this value was computed both for the individual measures [ICC(1,1)] and for the average of the two samples [ICC(1,2)]. To remain consistent with the assumptions of this model, a skewness and kurtosis test was performed for each semen parameter to test the assumption of normality. Transformation of non-parametric data was then performed as needed to provide normally distributed data that also met the assumption of homoscedasticity for the ANOVA model. ICC values approaching one would suggest that the within-patient differences are small relative to the actual differences between individuals within the population.

An ideal test would have a low CV_w (indicating reproducibility) and a high ICC (indicating reliability). An upper limit of a two-fold difference ($CV_w = 100\%$) – for serologic tests -- is generally accepted regarding variability.¹⁷ Similarly, benchmarks for interpretation of the degree of agreement have been developed; statistics of agreement with values from 0.41 to 0.60, 0.61 to 0.80, and > 0.80 have been described as having moderate, substantial, and almost perfect agreement respectively.²⁰ Statistical analysis was performed using SPSS version 20 (IBM, Armonk, New York) and STATA version 12.0 (StataCorp, College Station, Texas).

Results

A total of 86 subjects were eligible for the study by having submitted two semen samples. Seven patients were excluded: one was missing motility data from the SA, one had an abstinence period of one day, and five had undergone varicocelectomy between submissions of the two samples. Mean age of the group ($n=79$) was 18.8 ± 1.2 years (median 18.3, range 17.8-24.7). The average time between semen analyses was 1.3 ± 1.6 years (median 0.5).

Table1 describes the individual semen parameters from each SA. Figure1 illustrates the relationship between the two consecutive SAs for volume, density, motility, and TMC. Concordance rates ranged from 77% to 92% for each semen parameter (Table2). Sperm count demonstrated the highest concordance; motility, the lowest.

There was no correlation between the individual patient deviation and mean for both volume and motility ($\rho=0.101$, $p=0.373$ and $\rho=0.153$, $p=0.178$, respectively) suggesting an independent relationship of these parameters. However, density ($\rho=0.781$, $p<0.001$), count ($\rho=0.845$, $p<0.001$), and TMC ($\rho=0.887$, $p<0.001$), all demonstrated increasing individual patient deviation with increasing mean. Therefore, CV_w would be recommended for comparing the variability of these parameters. The CV_w was high for each parameter. Motility and volume had the least variation (each 36%), while TMC had the greatest (82%). Table2 summarizes the reproducibility of each parameter.

Semen volume, density, count, and TMC were not normally distributed. Thus, to calculate ICC via the ANOVA model, these data were transformed into parametric data in the following manner: square root (volume), density^(1/4), count^(1/5), and LN(TMC+0.1). No transformation was necessary for motility.

ICC for a single measure demonstrated moderate agreement for motility ($ICC(1,1) = 0.552$ [0.378–0.688]) and moderate reliability for volume. TMC was substantially reliable with $ICC(1,1) = 0.778$ [0.674–0.852]. Count and density were the most reliable measures reflecting outstanding agreement ($ICC(1,1) = 0.876$ [0.813–0.919] and 0.871 [0.806–0.915], respectively). Table 2 includes data for the estimates and 95% confidence intervals of the ICC for each parameter, for both a single measure and an average of two measures. When using the average of two SAs, the reliability improved for each parameter. Motility -- the least reliable parameter -- and volume improved to moderate-to-substantial reliability when SA were averaged. All other parameters demonstrated at least substantial agreement based on an average of two SAs. Figure 2 depicts the reliability for each parameter based on both a single measure and on an average of two SA.

Discussion

In general, in this population of transitioning youths at risk for subfertility, it appears that the individual semen parameters demonstrate a high degree of variation which may limit reproducibility. However, the critical parameters of density, count, and TMC can be interpreted to be reliably consistent measures for an individual patient relative to the population, thus allowing for classification of a patient's risk for infertility based on a single SA.

The validity of our results is confirmed in a comparison with literature from adult populations of subfertile men. The results from three previous studies documenting CV_w in the infertile adult population found the ranges of values for each parameter to be: volume (27.1% to 29.6%), density (29% to 56.8%), count (54.2% to 63.4%), motility (34% to 34.4%), and TMC (30% to 73.2%).¹¹⁻¹³ Calculated values of CV_w from our study were slightly above the upper ranges for each of these parameters: volume (36%), density (63%), count (69%), motility (36%), and TMC (82%). This suggests that the degree of within-patient variation in the transitioning adolescent at risk for subfertility is similar to, but perhaps slightly greater than, the variation seen in an infertile adult population.

These previous studies all found that the between-patient variation was greater than the within-subject variation, which is also consistent with our findings.¹¹⁻¹³ Leushuis et al. analyzed 5,240 subfertile men who submitted two semen analyses.¹³ From this group of patients, they found the ICCs to be 0.70 for volume, 0.89 for density, 0.58 for motility, and 0.73 for TMC; only volume was outside of the confidence interval of our results (albeit by only 0.01). The remaining estimates from that study were well within range of the estimates of our study population.

Our results also reveal that the semen parameters trend towards increased reliability with a second sample. The overlapping 95% confidence intervals shown in Figure 2 preclude us from saying that these results are significantly different compared to those obtained from a single SA. Previous studies have shown that the ICC values are higher for each parameter as the number of samples increases, thereby improving the ability to characterize the patient.^{11, 12} However, while this trend for improvement is clear, it may take as many as five samples before a statistically significant improvement in the ICC can be shown.¹¹ Guidelines generally recommend two samples in the evaluation of infertile men. By following these guidelines we are accepting inherent variability that may limit our ability to fully characterize the patient. Indeed, more samples would do a better job. If we are willing to accept these shortcomings for two samples and we cannot show a significant difference between two samples and one sample, would we be willing to accept the potential for mischaracterization based off of just one sample?

This reliability of a single SA, particularly TMC, should be sufficient to guide clinical decisions and to provide counseling for the patient. However, a second SA allows for an average of the two values, which provides even better classification of the patient. In this population of at-risk youths, one could consider obtaining just one sample in a patient with a normal result, who would appear to be at lower risk for infertility compared to his peers. From Figure 1D, we can see that the odds that a patient would have a severely diminished TMC on repeat SA, given a normal TMC on the 1st SA, was 4:17. Increased concern in those with an abnormal result could be further addressed with a second sample to confirm the findings. In the setting of a low TMC on the first SA, 7 patients out of 35 (20%) saw an improvement in TMC on their 2nd SA. Count and density each performed better than TMC. Indeed, Keel found that sperm donors could more appropriately be described by a single SA compared to male infertility patients in a study of 479 subjects with a total of 2,043 SA.¹² We are not suggesting this selective use of a single SA in the infertile adult population; but it may be a reasonable approach to use, for instance, in adolescents with a varicocele as we attempt to characterize the individual's future risk and to decide whether to surgically intervene at a certain point in time. If there is any question regarding the collection of the first sample, however, a second one should be requested.

There are some limitations of these data. This study was conducted in patients with a history of either cryptorchidism or varicocele. The retrospective nature of the study approach increases the potential for both selection and observation bias. As a result, it must be clearly stated that these data are not necessarily generalizable to the peri-adolescent population as a whole. Furthermore, statistics of agreement, such as ICC, are affected by the prevalence of the disease in the population. In this study, by selecting a group of patients at-risk for future fertility issues, the ICC could therefore be biased towards a lower result than in comparison to the population at large. This is because a more homogenous population will have lower between-patient variability. Similarly, potential heterogeneity of sexual maturation in our cohort, beyond what we were able to control for with Tanner staging, could affect the results. This within-patient heterogeneity would tend to bias the measure of reliability that we calculated to a value lower than the true value. Thus, this limitation actually strengthens the conclusion that semen analysis is a reliable measure with just one sample being tested. In addition, observation bias in this study could result from a non-standard time of abstinence between the SAs. Studies assessing the extent to which abstinence duration affects the within-patient variability are somewhat conflicting in the literature.^{12, 13} Our results may also be limited by the fact that different laboratories were used for evaluation of the SA; lack of standardization between laboratories has the potential to affect the reliability of the results.²¹ However, in our study, we feel that the use of different labs reflects a real-world scenario for our cohort due to the diversity and geography of our catchment area. Finally, it should also be recognized that pregnancy is the true gold standard test for fertility and anything short of that – including SA – is a surrogate marker with inherent limitations.

Conclusions

A large degree of within-patient variability of individual SA parameters exists in this population of youths at risk for infertility, as has been shown in adult men. However, overall there is substantial agreement and reliability between consecutive samples in this population. Thus, although multiple samples may improve SA reliability, it is possible to appropriately classify these young men based on the result of a single SA as they transition from pediatric to adult care.

Acknowledgments

The authors would like to thank Meg Richter, Angela Kalmus, and Aileen Schast, PhD for their tireless commitment to this research project; it could not have been completed without them.

References

1. Lee PA, O'Leary LA, Songer NJ, et al. Paternity after unilateral cryptorchidism: a controlled study. *Pediatrics*. 1996; 98:676. [PubMed: 8885945]
2. Greenberg SH, Lipshultz LI, Wein AJ. Experience with 425 subfertile male patients. *J Urol*. 1978; 119:507. [PubMed: 25971]
3. Gorelick JI, Goldstein M. Loss of fertility in men with varicocele. *Fertil Steril*. 1993; 59:613. [PubMed: 8458466]
4. Report on optimal evaluation of the infertile male. *Fertil Steril*. 2006; 86:S202. [PubMed: 17055823]
5. Alvarez C, Castilla JA, Martinez L, et al. Biological variation of seminal parameters in healthy subjects. *Hum Reprod*. 2003; 18:2082. [PubMed: 14507825]
6. Cooper TG, Jockenhovel F, Nieschlag E. Variations in semen parameters from fathers. *Hum Reprod*. 1991; 6:859. [PubMed: 1757526]
7. Poland ML, Moghissi KS, Giblin PT, et al. Stability of basic semen measures and abnormal morphology within individuals. *J Androl*. 1986; 7:211. [PubMed: 3745007]
8. Poland ML, Moghissi KS, Giblin PT, et al. Variation of semen measures within normal men. *Fertil Steril*. 1985; 44:396. [PubMed: 4029428]
9. Wijchman JG, de Wolf BT, Graafe R, et al. Variation in semen parameters derived from computer-aided semen analysis, within donors and between donors. *J Androl*. 2001; 22:773. [PubMed: 11545289]
10. Schwartz D, Laplanche A, Jouannet P, et al. Within-subject variability of human semen in regard to sperm count, volume, total number of spermatozoa and length of abstinence. *J Reprod Fertil*. 1979; 57:391. [PubMed: 513028]
11. Francavilla F, Barbonetti A, Necozione S, et al. Within-subject variation of seminal parameters in men with infertile marriages. *Int J Androl*. 2007; 30:174. [PubMed: 17239086]
12. Keel BA. Within- and between-subject variation in semen parameters in infertile men and normal semen donors. *Fertil Steril*. 2006; 85:128. [PubMed: 16412742]
13. Leushuis E, van der Steeg JW, Steures P, et al. Reproducibility and reliability of repeated semen analyses in male partners of subfertile couples. *Fertil Steril*. 2010; 94:2631. [PubMed: 20434148]
14. Mallidis C, Howard EJ, Baker HW. Variation of semen quality in normal men. *Int J Androl*. 1991; 14:99. [PubMed: 1869320]
15. World Health Organization Laboratory Manual for the Examination and Processing of Human Semen. 5th. 2010.
16. Cayan S, Erdemir F, Ozbey I, et al. Can varicocelectomy significantly change the way couples use assisted reproductive technologies? *J Urol*. 2002; 167:1749. [PubMed: 11912402]
17. Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. *Clin Diagn Lab Immunol*. 2002; 9:1235. [PubMed: 12414755]
18. Bland JM, Altman DG. Measurement error proportional to the mean. *BMJ*. 1996; 313:106. [PubMed: 8688716]
19. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979; 86:420. [PubMed: 18839484]
20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977; 33:159. [PubMed: 843571]
21. Keel BA. How reliable are results from the semen analysis? *Fertil Steril*. 2004; 82:41. [PubMed: 15236987]

Abbreviations

SA	semen analysis/es
CV_w	within-subject coefficient of variation
ICC	intraclass correlation coefficient

Figure1.

Scatterplots for individual semen parameters comparing the first to the second SA. Reference lines are drawn at the lower normal limit according to WHO standards, dividing the plot into 4 quadrants. The upper left and lower right quadrants represent discordance between the two SA; the lower left and upper right quadrants represent concordance (agreement) between the first and second SA result. A) Semen density, B) semen volume, C) sperm motility, D) total motile count.

Figure 2. Individual semen parameter reliability for a single measure vs. an average of two measures. A scale for the degree of reliability is shown on the axis on the right. While adding a second measure of each parameter improves the reliability, the confidence intervals overlap.

Table1

Results of Consecutive Semen Analyses

Semen Parameter	SA #1	SA #2
Volume (mL)		
mean +/- SD	2.8+/-1.6	3.1 +/-1.6
median (IQR [*])	2.8(1.5-4.0)	3.0(2.0-4.1)
range	0.1-7.6	0.5-7.2
Density (N/mL)		
mean +/- SD	23.6 +/- 34.5	23.6 +/- 31.3
median (IQR)	6.5(1.7-29.0)	8.0(3.0-30.7)
range	0 -138.0	0 -122.7
Count (N, million)		
mean +/- SD	74.5 +/-153.5	84.0 +/-143.2
median (IQR)	16.0(4.0-56.0)	27.0(6.0-81.9)
range	0 - 854.6	0-871.2
Motility (%)		
mean +/- SD	39.3 +/- 23.0	45.1 +/- 22.3
median (IQR)	44.0(20.0-56.0)	48.0(34.0-58.0)
range	0-100	0 - 87.0
TMC (N, million)		
mean +/- SD	39.8 +/- 92.8	52.1+/-103.1
median (IQR)	6.5(0.8-243)	10.2(2.1-48.4)
range	0 - 521.3	0 - 670.8

^{*} IQR = interquartile Range

Table 2

Reproducibility and Reliability of Semen Analyses

Semen Parameter	Concordance	CVw	ICC(1, 1) [~]	ICC (1, 2) [†]
Volume <i>95% CI</i>	85% (67/79) [*]	36%	0.555 (0.382-0.690)	0.714 (0.553-0.817)
Density <i>95% CI</i>	91% (72/79) [*]	63%	0.871 (0.806-0.915)	0.931 (0.892-0.956)
Count <i>95% CI</i>	92% (73/79) [*]	69%	0.876 (0.813-0.919)	0.934 (0.897-0.958)
Motility <i>95% CI</i>	77% (61/79) [*]	36%	0.552 (0.378-0.688)	0.711 (0.549-0.815)
TMC <i>95% CI</i>	86% (68/79) [^]	82%	0.778 (0.674-0.852)	0.875 (0.805-0.920)

^{*} Normal is Defined by WHO 2010 standards: volumes 1.5mL, density 15N/mL, count 39 N, motility 40%

[^] Normal is defined by TMC 20 N/ejaculate

[~] ICC for a single measure (i.e., a single SA) using a one-way random effects model

[†] ICC for the average of two measures using a one-way random effects model