American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2017.

Vol. 186, No. 8
DOI: 10.1093/aje/kwx169
Advance Access publication:
May 24, 2017

Original Contribution

What Does a Single Semen Sample Tell You? Implications for Male Factor Infertility Research

Yu-Han Chiu, Regina Edifor, Bernard A. Rosner, Feiby L. Nassan, Audrey J. Gaskins, Lidia Mínguez-Alarcón, Paige L. Williams, Cigdem Tanrikut, Russ Hauser, and Jorge E. Chavarro* for the EARTH Study Team

* Correspondence to Dr. Jorge E. Chavarro, Department of Nutrition, Harvard T.H. Chan School of Public Health, 665 Huntington Avenue, Boston, MA 02115 (e-mail: jchavarr@hsph.harvard.edu).

Initially submitted August 11, 2016; accepted for publication December 16, 2016.

Semen parameters are variable within individuals, but it is unclear whether 1 semen sample could represent a man's long-term average values in epidemiologic studies. Between 2005 and 2014, a total of 329 men from a fertility clinic in Boston, Massachusetts, provided 768 semen samples as part of the Environment and Reproductive Health (EARTH) Study. Total sperm count, sperm concentration, morphology, motility, and ejaculate volume were assessed. We used linear mixed models to compare values from men's first semen samples with their long-term averages and to calculate intraclass correlation coefficients for each parameter. We calculated positive predictive values (PPVs) and negative predictive values (NPVs) by comparing agreement in classification according to World Health Organization reference limits. There were no differences in mean semen parameters between men's first samples and the remaining replicates. Intraclass correlation coefficients ranged from 0.61 for morphology to 0.75 for concentration, indicating consistently greater between-man variability than within-man variability. Nevertheless, using 1 sample alone resulted in high NPVs but low PPVs (range, 43%–91%). The average of 2 samples was needed to achieve high PPVs (range, 86%–100%) and NPVs (range, 91%–100%). We conclude that 1 semen sample may suffice for studies aimed at identifying average differences in semen quality between individuals. Studies aimed at classifying men based on World Health Organization reference limits may benefit from collection of 2 or more samples.

diagnosis; infertility; male factor infertility; reproducibility; semen parameters; semen samples; variability

Abbreviations: EARTH, Environment and Reproductive Health; ICC, intraclass correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; WHO, World Health Organization.

Infertility is estimated to affect approximately 10%–15% of couples attempting to conceive in their reproductive lifetime (1, 2), and a male factor is identified in 40%–60% of couples evaluated for infertility (3). The cornerstone of clinical evaluation of male infertility is the analysis of semen parameters, which also serves as a proxy for male fertility potential in epidemiologic studies designed to identify risk factors for male factor infertility (4–6).

It is well known that semen parameters are subject to inherent within-person variability over time (7–10). As a result, in clinical settings, the results of at least 2 semen samples are used for diagnosis of male factor infertility in order to minimize false-discovery and false-omission rates. Nonetheless, in clinical practice, a second diagnostic semen sample is usually requested only when the results of the first analysis fall below the World Health Organization (WHO) reference limits (11, 12). The extent to which this practice may result in underdiagnosis of male factor infertility is unclear.

In research applications, on the other hand, investigators face the challenge of making the most efficient allocation of a fixed budget in order to maximize the amount of information obtained. Specifically, investigators often face the decision of whether to obtain 2 or more semen samples per man,

mirroring clinical practice, or obtain a single sample per man in order to maximize sample size. The optimal conclusion for research purposes continues to be a matter of debate. Previous work among healthy individuals has found no differences, on average, between semen parameters obtained from duplicate samples (13). Nevertheless, since within-man variability in semen quality may be higher among subfertile men than among healthy men or donors (9), it is unclear whether study findings for fertile men are applicable to men recruited at fertility centers. In the present study, we sought to address the relative benefits of obtaining more than 1 semen sample per man in research settings, in a cohort of men in subfertile couples seeking evaluation and treatment at an academic fertility center.

METHODS

Study population

Study participants were male partners in subfertile couples enrolled in the Environment and Reproductive Health (EARTH) Study, an ongoing prospective cohort study of couples seeking infertility evaluation and treatment at the Massachusetts General Hospital Fertility Center (14). Men were eligible if they were between ages 18 and 55 years, had no history of vasectomy, and were partners in couples using their own gametes for intrauterine insemination or assisted reproduction (4). A total of 329 men provided 768 semen samples (range, 1–9 samples per subject) between January 2005 and August 2014. At enrollment, participants completed a general health questionnaire on demographic factors, lifestyle factors, and reproductive history. Clinical information was abstracted from electronic medical records. The study was approved by the human subjects committees of the Harvard T.H. Chan School of Public Health and Massachusetts General Hospital. Signed informed consent was provided by all participants before joining the study.

Semen analysis

All semen samples were obtained on-site by masturbation into a sterile plastic cup. Men were instructed to abstain from ejaculation for at least 48 hours, but no more than 5 days, before semen sample collection and to report the specific time of last ejaculation. Of 768 semen samples, 78 samples (from 75 men) did not have information provided on abstinence time. Analysis was initiated after completion of liquefaction at 37°C and within 30 minutes after ejaculation. Ejaculate volume was measured with a graduated serological pipette. To measure both sperm concentration and motility, 5 μL of semen from each sample was placed into a prewarmed (37°C) Makler counting chamber (Sefi Medical Instruments Ltd., Haifa, Israel). A minimum of 200 spermatozoa from at least 4 different fields were analyzed from each specimen. Sperm concentration and motility were assessed by means of computer-aided semen analysis (IVOS II, version 14; Hamilton Thorne, Beverly, Massachusetts). Percentage of motile sperm was classified in terms of progressive motility (i.e., percentage of progressively motile spermatozoa) and total motility (i.e., percentage of progressively plus nonprogressively motile spermatozoa) (10). Sperm morphology was assessed manually under high-resolution oil immersion microscope optics according to the strict criteria of Kruger et al. (15). For quality control purposes, the laboratory conducted weekly monitoring of sperm morphology smears and reevaluated any slides when reports deviated from acceptable ranges of variation. In addition, the laboratory performed a quarterly competency evaluation of all technicians and conducted proficiency testing every 6 months using an outside evaluator.

Total sperm count was calculated as (sperm concentration) × (ejaculate volume). Total motile sperm count was defined as (concentration) × (ejaculate volume) × (% total motility) and total normal count as (concentration) × (ejaculate volume) × (% morphologically normal). In addition, we dichotomized semen parameters according to the WHO's lower reference limits: 39 million per ejaculate for total sperm count, 15 million per mL for sperm concentration, 40% for total sperm motility, 32% for progressive motility, 4% morphologically normal sperm for morphology, and 1.5 mL for ejaculate volume (10). We defined good-semen-quality samples as those in which all semen parameters were above the WHO lower reference limits.

Statistical analyses

All statistical analyses were conducted using SAS 9.4 (SAS Institute, Inc., Cary, North Carolina). We classified participants into 3 groups according to the total number of semen samples provided (1, 2, or ≥3 samples), calculated descriptive statistics, and tested for differences between groups using Kruskal-Wallis and Fisher's exact tests for continuous and categorical variables, respectively. Approximately 10% of the samples ($n = 78$ samples) had missing data on abstinence time. To increase power and reduce bias, we used multiple imputation (PROC MI in SAS) to generate 10 imputed data sets. We used linear mixed models to evaluate whether the mean value of men's first samples differed systematically from the average of the remaining samples among 197 men who provided at least 2 semen samples ($n = 636$ samples). Specifically, mean semen quality in the population for the first sample and that for the remaining replicates was estimated in each imputed data set. The mean difference between men's first sample results and the average of the remaining samples was estimated by introducing an indicator for first sample into the regression model. The final estimates and associated 95% confidence intervals were computed using PROC MIANALYZE in SAS. Model-data agreement was assessed by means of conditional Studentized residuals for normality and homoscedasticity and by Cook's distance for influential outliers. No influential observation was identified using a rule of thumb with a Cook's distance value over 1.0 (16).

To improve normality and homoscedasticity, we log-transformed data on sperm concentration, sperm count, total motile sperm count, and total normal count and square-root-transformed data on ejaculate volume, progressive motility, and morphology. We also calculated intraclass correlation coefficients (ICCs) based on the estimates of within- and between-man variance obtained from the mixed-effects regression models. The models adjusted for age (<35 years, ≥35 years) and abstinence time (<2 days, 2–2.9 days, 3–3.9 days, or ≥4 days)

corresponding to each sample. Analyses of total motility and progressive motility were additionally adjusted for time between semen collection and the start of semen analysis. We also created spaghetti plots for men who had provided 7 or more samples ($n = 12$) for each of the semen quality parameters to visualize within-man and between-man variation over time.

We then assessed how well each man's first sample classified him according to the WHO 2010 values in a subset of men who provided at least 3 semen samples ($n = 104$ men and 450 samples) (10). To assess the discordance between the first and second samples, we calculated the conditional probability of the second-sample results given the first-sample results, focusing on the probability of discordant findings. We estimated sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false-discovery rate, and false-omission rate by comparing the agreement of the classification based on the first sample only and the classification based on each man's long-term average, defined as the average of the first to *N*th samples for each man. We also compared the agreement between the classification based on the average of the first and second samples and the classification based on the average of the first to *N*th samples for each man. These metrics were also calculated when the average of all of the semen samples was considered as the gold standard. Two-sided *P* values less than 0.05 were considered significant.

RESULTS

Characteristics of the EARTH study participants are summarized in Table 1. The median number of samples per man was 2 (range, 1–9), and the median amount of time between the collection of each man's first and last samples was 156 days (range, 6–2,239). There were no appreciable differences in demographic and reproductive characteristics by the number of samples given (1, 2, or ≥3) (Table 1).

As has been reported by others (9, 17–20), visual evaluation of the results of semen analyses over time suggested considerable within-man variability (Figure 1). Nevertheless, there were no significant differences between the mean value of the men's first samples and the mean of their remaining samples (Table 2). Furthermore, when variability was quantified, between-man variability was consistently larger than within-man variability, as reflected in the ICCs. The ICCs across all semen parameters ranged from 0.61 for morphology to 0.75 for sperm concentration. The results were consistent when the analysis was restricted to men who had at least 1 abnormal semen parameter based on their long-term average. All results were nearly identical when we included men who provided only 1 semen sample in the analysis in order to improve the estimation of between-man variability (see Web Table 1, available at https://academic.oup.com/aje).

We then calculated how much statistical power would be gained when taking 2 samples from each man versus 1 sample from each man at any given sample size. When the number of men in each group was fixed, additional samples improved power (Web Figure 1). Obtaining replicates was noticeably better for power for a parameter with a lower ICC compared with one with a higher ICC, especially when the sample size was low. Nonetheless, when the total number of semen samples was fixed, power decreased with increasing number of replicates per man (Web Figure 2).

When the results of semen analysis were dichotomized according to the WHO lower reference limits, 27% of men had discordant results between their first 2 samples. Specifically, among men with at least 1 parameter below the reference limits in their first sample, 8.5% had all semen parameters above the WHO reference limits in their second sample. On the other hand, among men with all semen parameters above the WHO reference limits in their first sample, 51% had at least 1 semen parameter below the WHO reference limits in their second sample.

We also evaluated the accuracy of the first sample to classify a man according to the WHO lower reference limits by comparing it with the classification obtained from each man's long-term average. The population prevalence of at least 1 semen parameter below WHO reference values based on each man's first sample was slightly lower than that based on each man's long-term average values (56.7% vs. 67.3%; Table 3). When we assessed the accuracy of classification at the individual level, classification based on the first semen sample alone had high NPVs but low PPVs for all semen quality parameters, with the exception of total and progressive motility (Table 3). The average of 2 semen samples was necessary to accurately classify men on all semen parameters with high sensitivity (range, 91%–100%), specificity (range, 89%–100%), PPV (range, 86%–100%), and NPV (range, 91%–100%) (Table 4).

Lastly, we estimated the frequency of false-discovery and false-omission classification of male factor infertility using the results from the first semen sample compared with those from the first 2 samples. As Table 3 shows, 6.8% of men with at least 1 abnormal semen parameter would have been classified as normal if only 1 semen sample had been considered. The false discovery rate was reduced to 2.9% when the first 2 samples were considered (Table 4). On the other hand, 33.3% of men classified as having a normal semen analysis on the basis of their first sample had at least 1 semen parameter below WHO reference limits when they were classified on the basis of long-term averages (Table 3). The results of a second sample improved classification: 8.6% of men were misclassified as normal when the results of their first 2 semen samples were considered relative to their average values from all of the samples.

DISCUSSION

We evaluated the long-term variability in semen parameters among men visiting a fertility center in the EARTH Study. As previously reported (9, 17–20), graphical evaluation of variability over time suggested substantial within-man variability. Nevertheless, we found that between-man variability for each of the parameters investigated was consistently larger than within-man variability. In addition, the mean value of men's first samples was not significantly different from the mean of multiple semen samples. On the other hand, when semen parameters were dichotomized, a second sample was necessary to estimate the prevalence of below-WHO-reference values at population levels and to achieve adequate classification of all parameters for an individual man. In fact, our findings suggest that not collecting a second semen sample when the results of the first one are

Table 1. Characteristics of Participants According to Total Number of Semen Samples Provided, Environment and Reproductive Health Study, Boston, Massachusetts, 2005–2014

Characteristic	No. of Semen Samples Provided								*P* Value[a]
	Total (n = 329 Subjects, n = 768 Samples)		1 (n = 132 Subjects, n = 132 Samples)		2 (n = 93 Subjects, n = 186 Samples)		≥3[b] (n = 104 Subjects, n = 450 Samples)		
	Median (IQR)	%	Median (IQR)	%	Median (IQR)	%	Median (IQR)	%	
Time metrics									
Time between first and last samples, days	182 (87–331)				88 (53–147)		292 (186–473)		<0.0001
Time between subsequent samples, days	91 (57–150)				88 (53–147)		98 (57–165)		0.31
Abstinence time before first sample, hours	60 (48–79)		60 (48–74)		58 (48–85)		61 (49–83)		0.79
Demographic characteristics									
Age, years	36 (33–39)		36 (33–39)		35 (33–39)		36 (33–39)		0.44
Race/ethnicity									0.83
White, not Hispanic		88		89		86		88	
Black		3		4		2		2	
Asian		6		5		6		8	
Hispanic or Latino		4		3		5		3	
Current smoker		6		5		11		5	0.16
Body mass index[c]	27 (24–30)		27 (24–30)		27 (25–29)		27 (24–30)		0.91
Quality of first semen sample									
Ejaculate volume, mL	3 (2–4)		3 (2–4)		3 (2–4)		3 (2–4)		0.50
Sperm concentration, millions/mL	63 (30–106)		65 (29–113)		64 (32–96)		61 (28–116)		0.91
Total sperm count, millions	156 (81–266)		168 (93–263)		163 (71–266)		121 (77–276)		0.63
Total motility, %	48 (28–63)		51 (30–65)		44 (24–59)		48 (28–67)		0.22
Progressive motility, %	26 (15–37)		27 (16–38)		22 (13–34)		29 (16–39)		0.19
Normal morphology, %	6 (4–8)		6 (4–10)		5 (3–8)		6 (4–8)		0.18
Self-reported reproductive history									
Diagnosis of male factor infertility[d]		35		34		40		33	0.55
Undescended testes		5		8		3		2	0.10
Varicocele		9		6		12		9	0.25
Reproductive surgery[e]		20		22		14		22	0.26

Abbreviation: IQR, interquartile range.
[a] Kruskal-Wallis test for continuous variables; χ^2 test for categorical variables.
[b] The maximum number of samples was 9.
[c] Weight (kg)/height (m)^2.
[d] Based on Society for Assisted Reproductive Technology diagnoses.
[e] Report of any of the following: orchidopexy, varicocelectomy, hydrocelectomy, hernia repair, urethral repair, hypospadias repair, sympathectomy, bladder neck surgery, or other reproductive surgery.

normal could potentially lead to underdiagnosis of male factor infertility. These findings support the collection of 2 or more semen samples, regardless of the results of the first one, when the goal is to classify individual men according to relevant cutoffs, as is the case with clinical evaluation of men in subfertile couples and certain research settings. Nevertheless, they also suggest that in research settings where semen quality parameters are treated as continuous variables—that is, when the aim of the study is to identify average differences between groups of men—obtaining more than 1 sample per man does not offer any appreciable advantage over obtaining a single sample.

We found no significant differences between the mean value of men's first semen samples and the mean value of the remaining replicates. These findings are in agreement with those of previous studies (13, 21). In a multicenter study of 615 fertile men, Stokes-Riner et al. (13) showed that semen parameters were not significantly different between the first and second samples after adjusting for abstinence time, time from ejaculation to the start of the analysis, and study center. Similarly, Bae et al. (21) reported no significant differences in semen parameters between the first and second semen samples among 227 male partners with a singleton baby. On the other hand, 2 studies have found significant differences in certain semen parameters between 2 samples (19, 20). Francavilla et al. (19) found that only volume and combined parameters including volume (total sperm count and total

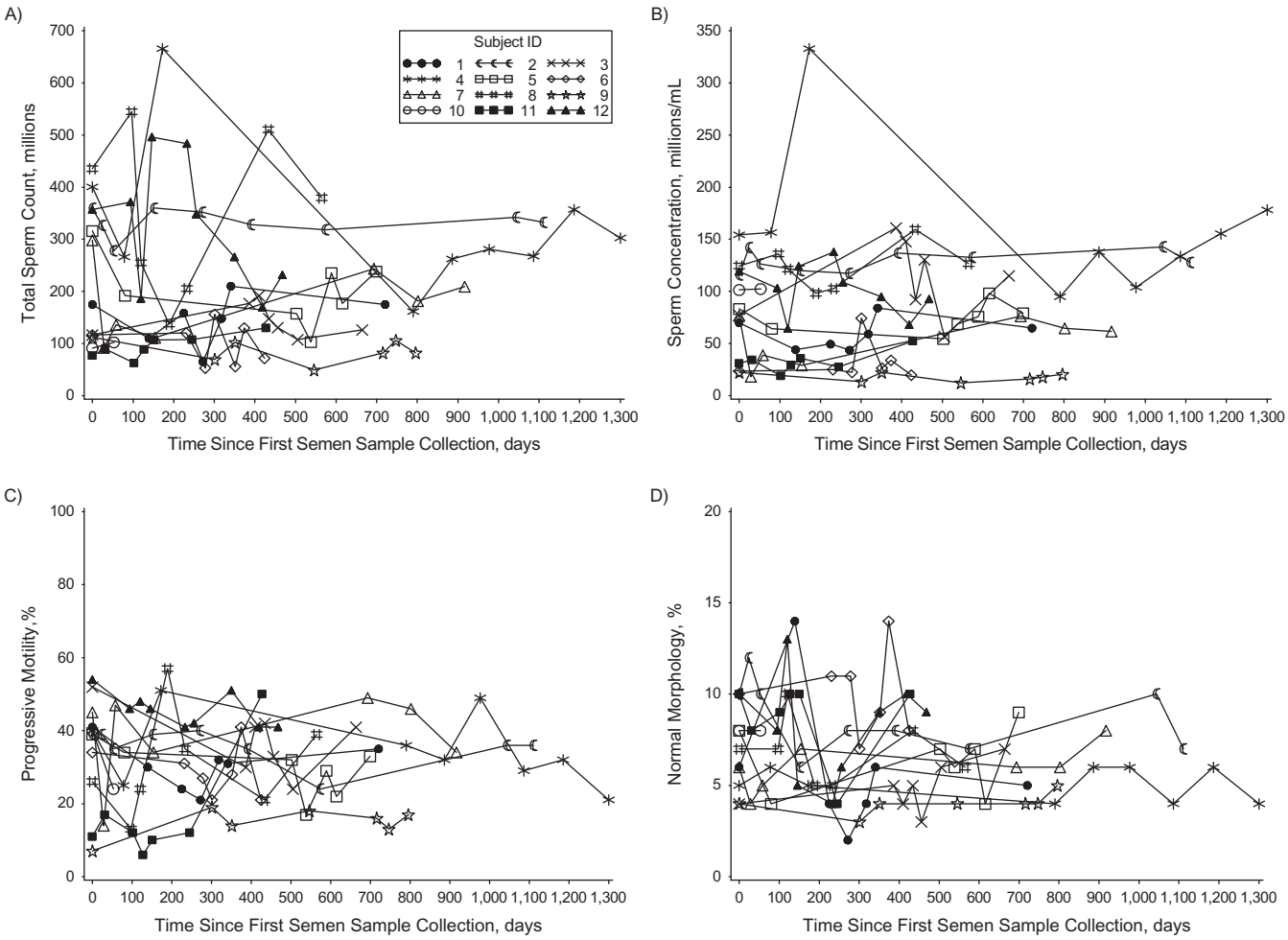

Figure 1. Variation in semen parameters over time among 12 men who provided 7 or more semen samples in the Environment and Reproductive Health Study, Boston, Massachusetts, 2005–2014. The figure shows variation in total sperm count (A), sperm concentration (B), progressive motility (C), and morphology (D) over time among men who provided 7–9 samples (total number of samples = 92). ID, identification.

motile sperm count) were significantly lower in a repeat sample obtained after 1 day of abstinence. These results are not directly comparable to our findings, given that samples obtained on consecutive days are expected to be systematically different, especially for semen volume and sperm concentration, which are the parameters most influenced by an insufficient abstinence interval (22–24). Secondly, in a study of 5,240 subfertile men from 2 medical centers in the Netherlands, Leushuis et al. (20) found a small but statistically significant difference between 2 semen analyses for ejaculate volume (−0.04 mL, 95% confidence interval: −0.07, 0) and progressive motility (−2.10%, 95% confidence interval: −2.62, −1.57) but not for other semen parameters, including concentration, morphology, and total motile sperm count. Nonetheless, in contrast with the present report and those of Bae et al. (21) and Stokes-Riner et al. (13), the results of the Leushuis et al. study were not adjusted for duration of abstinence, time from ejaculation to the start of analysis, and study center, which may explain the small differences between the 2 semen analyses (20). In addition, the magnitude of the differences identified, while statistically significant, may not be clinically relevant.

In agreement with previous studies by other investigators evaluating the relative contributions of between- and within-man variation in semen parameters (9, 19, 20), we too found that between-man variability was consistently larger than within-man variability and that results of semen analyses were highly reproducible. In our study, sperm concentration had the highest reproducibility (ICC = 0.75), whereas progressive motility and sperm morphology had the lowest (ICC = 0.63 and ICC = 0.61, respectively). Similar to our findings, Leushuis et al. reported the highest ICCs for sperm concentration (ICC = 0.89) and lower ICCs for motility (ICC = 0.58) and morphology (ICC = 0.60) among subfertile men (20). Another study consisting of men whose partners were undergoing intrauterine insemination also showed a higher ICC for sperm concentration (ICC = 0.92) and a lower ICC for rapid motility (ICC = 0.78) (19). Higher ICCs for sperm count and sperm concentration and a lower ICC for motility were also reported in the studies conducted among healthy

Table 2. Comparison of Mean Values[a] for Semen Parameters Between the First Semen Sample and the Long-Term Average of Other Samples Among 197 Men Who Provided 2 or More Samples ($n = 636$ Samples), Environment and Reproductive Health Study, Boston, Massachusetts, 2005–2014

Semen Parameter	First Sample		Remaining Samples		Difference[b] Between First Sample and Remaining Samples		ICC	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Ejaculate volume[c], mL	2.7	2.5, 2.9	2.6	2.4, 2.8	0.02	−0.02, 0.06	0.71	0.65, 0.76
Sperm concentration[d] $\times 10^6$/mL, %	54.1	48.0, 61.1	52.4	46.3, 59.3	3	−5, 13	0.75	0.70, 0.80
Total sperm count[d] $\times 10^6$, %	128	113, 144	135	119, 153	6	−5, 17	0.64	0.57, 0.70
Total motility[e],%	44.4	40.6, 48.2	43.3	39.6, 47.0	1.12	−1.31, 3.55	0.66	0.59, 0.72
Progressive motility[c,e], %	22.3	19.6, 25.1	21.6	19.2, 24.1	0.08	−0.11, 0.26	0.63	0.56, 0.70
Normal morphology[c], %	5.5	5.0, 6.1	5.7	5.3, 6.2	−0.04	−0.11, 0.04	0.61	0.54, 0.68
Total motile sperm count[d] $\times 10^6$, %	51.4	42.0, 62.7	46.5	38.0, 56.9	10	−5, 29	0.71	0.65, 0.76
Total normal sperm count[d] $\times 10^6$, %	6.6	5.5, 8.0	6.5	5.4, 7.8	2	−11, 17	0.71	0.65, 0.76

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient.

[a] Adjusted for age (<35 years, ≥35 years) and abstinence time (<2 days, 2–2.9 days, 3–3.9 days, or ≥4 days).

[b] Relative differences (%) are presented for total sperm count, sperm concentration, total motile sperm count, and total normal sperm count; absolute differences (unit for corresponding semen parameters) are presented for other semen parameters. For ejaculate volume, progressive motility, and morphology, data are presented as the difference between square roots of the mean value. Estimates were obtained from linear mixed models comparing the mean value of men's first samples with the mean of the other samples while accounting for within-man variation.

[c] Data on ejaculate volume, progressive motility, and morphology were square-root-transformed, and back-transformed mean values are presented on the original scale.

[d] Data on these semen parameters were log-transformed, and back-transformed mean values are presented on the original scale.

[e] Additionally adjusted for time between semen collection and semen analysis (≤30 minutes, >30 minutes).

men (17, 25). Collectively, the findings from studies evaluating average differences between first and subsequent samples and the findings from studies evaluating reproducibility over time suggest that a single semen sample can adequately represent a man's average semen quality in studies aimed at identifying average differences in semen quality between groups of men.

In research settings, although using a single sample provides an unbiased estimate of the population mean, it provides a larger variance estimate than what would be obtained

Table 3. Accuracy in Classification (%) of Dichotomized Semen Parameters (Based on World Health Organization 2010 Reference Values) From the First Semen Sample Among 104 Men Who Provided 3 or More Samples ($n = 450$ Samples), Environment and Reproductive Health Study, Boston, Massachusetts, 2005–2014

Semen Parameter (Dichotomized Based on WHO Reference Values)	First Sample Below WHO Reference Limits	Average of All Samples Below WHO Reference Limits	Sensitivity[a]	Specificity[a]	PPV[a]	NPV[a]	FDR[a]	FOR[a]
Ejaculate volume <1.5 mL	14.4	12.5	61.5	92.3	53.3	94.4	46.7	5.6
Concentration <15 million/mL	7.7	6.7	71.4	96.9	62.5	97.9	37.5	2.1
Total sperm count <39 million	6.7	5.8	50.0	95.9	42.9	96.9	57.1	3.1
Total motility <40% motile	39.4	45.2	78.7	93.0	90.2	84.1	9.8	15.9
Progressive motility <32% motile	52.9	65.4	73.5	86.1	90.9	63.3	9.1	36.7
Morphology <4% normal	22.1	20.2	76.2	91.6	69.6	93.8	30.4	6.2
Good-semen-quality samples[b]	43.3	32.7	88.2	78.6	66.7	93.2	33.3	6.8
At least 1 parameter fell below WHO reference limits	56.7	67.3	78.6	88.2	93.2	66.7	6.8	33.3

Abbreviations: FDR, false-discovery rate; FOR, false-omission rate; NPV, negative predictive value; PPV, positive predictive value; WHO, World Health Organization.

[a] The gold standard was based on the average of values from the first to *N*th semen samples.

[b] Good-semen-quality samples were defined as those having all evaluated semen parameters above the WHO reference values.

Table 4. Accuracy in Classification (%) of Dichotomized Semen Parameters (Based on World Health Organization 2010 Reference Values) From the First 2 Semen Samples Among 104 Men Who Provided 3 or More Samples ($n = 450$ Samples), Environment and Reproductive Health Study, Boston, Massachusetts, 2005–2014

Semen Parameter (Dichotomized Based on WHO Reference Values)	Average of First 2 Samples Below WHO Reference Limits	Average of All Samples Below WHO Reference Limits	Sensitivity[a]	Specificity[a]	PPV[a]	NPV[a]	FDR[a]	FOR[a]
Ejaculate volume <1.5 mL	11.5	12.5	92.3	100	100	98.9	0	1.1
Concentration <15 million/mL	6.7	6.7	100	100	100	100	0	0
Total sperm count <39 million	5.8	5.8	100	100	100	100	0	0
Total motility <40% motile	46.2	45.2	95.7	94.7	93.8	96.4	6.2	3.6
Progressive motility <32% motile	66.4	65.4	95.6	88.9	94.2	91.4	5.8	8.6
Morphology <4% normal	21.2	20.2	90.5	96.4	86.4	97.6	13.6	2.4
Good-semen-quality samples[b]	33.7	32.7	94.1	95.7	91.4	97.1	8.6	2.9
At least 1 parameter fell below WHO reference limits	66.4	67.3	95.7	94.1	97.1	91.4	2.9	8.6

Abbreviations: FDR, false-discovery rate; FOR, false-omission rate; NPV, negative predictive value; PPV, positive predictive value; WHO, World Health Organization.

[a] The gold standard was based on the average of values from the first to *N*th semen samples.

[b] Good-semen-quality samples were defined as those having all evaluated semen parameters above the WHO reference values.

from taking multiple samples from each subject. Therefore, taking additional semen samples from each man will still be helpful to improve study precision. It is interesting that when the ICC is low, obtaining replicates is noticeably better for statistical power, especially when the sample size is low. On the other hand, if the ICC is high, adding replicate samples has minimal benefits for power. However, when an investigator must further consider the optimal allocation of resources on a fixed budget (leading to a fixed number of samples), taking a single sample from as many men as possible is the most efficient strategy.

The WHO recommends that characterizing a man's baseline semen quality in relation to the WHO 2010 lower reference limits requires considering the results of at least 2 semen samples (10, 26). Consistent with this recommendation, we found that using a single sample resulted in low PPVs (range, 43% for total sperm count to 91% for progressive motility) and NPVs (range, 63% for progressive motility to 98% for sperm concentration) for some parameters. In order to achieve acceptable classification levels, the average of 2 samples was necessary (PPV range, 86%–100%; NPV range, 91%–100%). The recommendation to assess at least 2 semen samples does not exactly reflect the management of infertility work-up in clinical practice. For example, the clinical guidelines from the American Society for Reproductive Medicine recommend that a repeat confirmatory test be performed if the initial evaluation is abnormal, implying that clinicians may stop at 1 semen analysis if the results of the first sample are normal (12). Our data reveal a pitfall of this practice, potentially resulting in underdiagnosis of male factor infertility. In addition, our study showed that 27% of men had discordant results between their first 2 samples and that the probability of having discordant findings was higher among men whose first sample was normal than among men whose first sample had abnormal values. Taken together, our findings suggest that for studies aimed at classifying men according to WHO reference limits, collecting 2 or more samples would be recommended.

Strengths of this study include its prospective design, the large number of repeated semen analysis results available per man, and the use of a single laboratory, reducing the possibility of interlaboratory variation (27). A potential limitation of the study is that the majority of study participants were non-Hispanic whites, and thus the findings may not be generalizable to minority men. Similarly, because the study is restricted to men visiting a fertility center and their average semen quality was lower than that of men in the general population (13, 21, 28), we cannot be certain that results will be generalizable to men with untested fertility or fertile men. However, the consistency of our findings with previous work carried out among fertile men (13, 21) decreases concern about this issue. Lastly, while the analysis of discrimination ability based on the WHO lower reference limits provided insights into the strengths and pitfalls of diagnostic strategies for male factor infertility, it is well known that classifying men according to these cutoffs does not provide accurate prediction of fertility in natural settings or among couples undergoing infertility treatment (29, 30).

In conclusion, our results show that the usefulness of a single semen sample depends on the intended goal. Using the results of a single semen sample per man may be the most efficient allocation of resources for studies aimed at identifying average differences in semen quality between groups of men. Nevertheless, for clinical diagnosis or studies where the goal is to classify an individual according to the WHO reference limits, at least 2 samples per man should be collected, regardless of the results of the first sample.

ACKNOWLEDGMENTS

Author affiliations: Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Yu-Han Chiu, Audrey J. Gaskins, Jorge E. Chavarro);

Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Regina Edifor); Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Bernard A. Rosner, Paige L. Williams); Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Feiby L. Nassan, Lidia Mínguez-Alarcón, Russ Hauser); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Yu-Han Chiu, Paige L. Williams, Russ Hauser, Jorge E. Chavarro); Department of Urology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts (Cigdem Tanrikut); Department of Obstetrics and Gynecology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts (Russ Hauser); and Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Bernard A. Rosner, Audrey J. Gaskins, Jorge E. Chavarro).

This work was supported by the National Institute of Environmental Health Sciences (grants R01-ES009718 and ES000002), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant L50-HD085359), and Harvard Catalyst | The Harvard Clinical and Translational Science Center (National Center for Research Resources and National Center for Advancing Translational Sciences, National Institutes of Health (award UL1 TR001102)), with additional financial support from Harvard University and its affiliated academic health-care centers.

We gratefully acknowledge all members of the EARTH study team, specifically research nurses Jennifer B. Ford and Myra G. Keller, senior research staff Ramace Dadd and Patricia Morey, and the physicians and staff at the Massachusetts General Hospital Fertility Center.

Conflict of interest: none declared.

REFERENCES

1. Louis JF, Thoma ME, Sørensen DN, et al. The prevalence of couple infertility in the United States from a male perspective: evidence from a nationally representative sample. *Andrology*. 2013;1(5):741–748.
2. Thoma ME, McLain AC, Louis JF, et al. Prevalence of infertility in the United States as estimated by the current duration approach and a traditional constructed approach. *Fertil Steril*. 2013;99(5):1324.e1–1331.e1.
3. Thonneau P, Marchand S, Tallec A, et al. Incidence and main causes of infertility in a resident population (1,850,000) of three French regions (1988–1989). *Hum Reprod*. 1991;6(6): 811–816.
4. Hauser R, Meeker JD, Duty S, et al. Altered semen quality in relation to urinary concentrations of phthalate monoester and oxidative metabolites. *Epidemiology*. 2006;17(6): 682–691.
5. Chavarro JE, Furtado J, Toth TL, et al. *Trans* fatty acid levels in sperm are associated with sperm concentration among men from an infertility clinic. *Fertil Steril*. 2011;95(5): 1794–1797.
6. Bergman Å, Heindel JJ, Jobling S, et al., eds. *State of the Science of Endocrine Disrupting Chemicals—2012. An Assessment of the State of the Science of Endocrine Disruptors Prepared by a Group of Experts for the United Nations Environment Programme and World Health Organization*. Geneva, Switzerland: World Health Organization; 2013.
7. Poland ML, Moghissi KS, Giblin PT, et al. Variation of semen measures within normal men. *Fertil Steril*. 1985;44(3): 396–400.
8. Mallidis C, Howard EJ, Baker HW. Variation of semen quality in normal men. *Int J Androl*. 1991;14(2):99–107.
9. Keel BA. Within- and between-subject variation in semen parameters in infertile men and normal semen donors. *Fertil Steril*. 2006;85(1):128–134.
10. World Health Organization. *WHO Laboratory Manual for the Examination and Processing of Human Semen*. 5th ed. Geneva, Switzerland: World Health Organization; 2010.
11. National Institute for Health and Care Excellence. *Fertility Problems: Assessment and Treatment. Investigation of Fertility Problems and Management Strategies*. (NICE clinical guideline 156). London, United Kingdom: National Institute for Health and Care Excellence; 2013. https://www.nice.org.uk/guidance/cg156/chapter/Recommendations#investigation-of-fertility-problems-and-management-strategies. Updated August 2016. Accessed September 28, 2016.
12. Practice Committee of the American Society for Reproductive Medicine. Diagnostic evaluation of the infertile male: a committee opinion. *Fertil Steril*. 2015; 103(3):e18–e25.
13. Stokes-Riner A, Thurston SW, Brazil C, et al. One semen sample or 2? Insights from a study of fertile men. *J Androl*. 2007;28(5):638–643.
14. Chavarro JE, Ehrlich S, Colaci DS, et al. Body mass index and short-term weight change in relation to treatment outcomes in women undergoing assisted reproduction. *Fertil Steril*. 2012; 98(1):109–116.
15. Kruger TF, Acosta AA, Simmons KF, et al. Predictive value of abnormal sperm morphology in in vitro fertilization. *Fertil Steril*. 1988;49(1):112–117.
16. Lomax RG, Hahs-Vaughn DL. *Statistical Concepts: A Second Course*. 4th ed. New York, NY: Routledge; 2012.
17. Alvarez C, Castilla JA, Martínez L, et al. Biological variation of seminal parameters in healthy subjects. *Hum Reprod*. 2003; 18(10):2082–2088.
18. Nallella KP, Sharma RK, Said TM, et al. Inter-sample variability in post-thaw human spermatozoa. *Cryobiology*. 2004;49(2):195–199.
19. Francavilla F, Barbonetti A, Necozione S, et al. Within-subject variation of seminal parameters in men with infertile marriages. *Int J Androl*. 2007;30(3):174–181.
20. 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(7): 2631–2635.
21. Bae J, Kim S, Chen Z, et al. Human semen quality and the secondary sex ratio. *Asian J Androl*. 2017;19(3):374–381.
22. Sauer MV, Zeffer KB, Buster JE, et al. Effect of abstinence on sperm motility in normal men. *Am J Obstet Gynecol*. 1988; 158(3 Pt 1):604–607.
23. Blackwell JM, Zaneveld LJ. Effect of abstinence on sperm acrosin, hypoosmotic swelling, and other semen variables. *Fertil Steril*. 1992;58(4):798–802.
24. De Jonge C, LaFromboise M, Bosmans E, et al. Influence of the abstinence period on human sperm quality. *Fertil Steril*. 2004;82(1):57–65.

25. Poland ML, Moghissi KS, Giblin PT, et al. Stability of basic semen measures and abnormal morphology within individuals. *J Androl*. 1986;7(4):211–214.
26. Sharlip ID, Jarow JP, Belker AM, et al. Best practice policies for male infertility. *Fertil Steril*. 2002;77(5): 873–882.
27. Jorgensen N, Auger J, Giwercman A, et al. Semen analysis performed by different laboratory teams: an intervariation study. *Int J Androl*. 1997;20(4):201–208.
28. Cooper TG, Noonan E, von Eckardstein S, et al. World Health Organization reference values for human semen characteristics. *Hum Reprod Update*. 2010;16(3):231–245.
29. Jedrzejczak P, Taszarek-Hauke G, Hauke J, et al. Prediction of spontaneous conception based on semen parameters. *Int J Androl*. 2008;31(5):499–507.
30. Buck Louis GM, Sundaram R, Schisterman EF, et al. Semen quality and time to pregnancy: the Longitudinal Investigation of Fertility and the Environment Study. *Fertil Steril*. 2014;101(2):453–462.