Title

Poor reproducibility of percentage of normally shaped sperm using the World Health Organization Fifth Edition strict grading criteria

Journal

F&S Reports, 3(2)

ISSN

2666-3341

Authors

Baker, Karen C
Steiner, Anne Z
Hansen, Karl R
et al.

Publication Date

2022-06-01

DOI

10.1016/j.xfre.2022.03.003

Peer reviewed

Poor reproducibility of percentage of normally shaped sperm using the World Health Organization Fifth Edition strict grading criteria

Karen C. Baker, M.D.,[a] Anne Z. Steiner, M.D., M.P.H.,[b] Karl R. Hansen, M.D., Ph.D.,[c] Kurt T. Barnhart, M.D.,[d] Marcelle I. Cedars, M.D.,[e] Richard S. Legro, M.D.,[f] Michael P. Diamond, M.D.,[g] Stephen A. Krawetz, Ph.D.,[h] Rebecca Usadi, M.D.,[i] Valerie L. Baker, M.D.,[j] R. Matthew Coward, M.D.,[k] Fangbai Sun, M.P.H.,[l] Robert Wild, M.D., M.P.H., Ph.D.,[m] Puneet Masson, M.D.,[n] James F. Smith, M.D., M.S.,[o] Nanette Santoro, M.D.,[p] and Heping Zhang, Ph.D.,[q] for the Reproductive Medicine Network

[a] Division of Urology, Duke University, Durham, North Carolina; b Department of Obstetrics and Gynecology, Duke University, Durham, North Carolina; [c] Department of Obstetrics and Gynecology, Health Sciences Center, University of Oklahoma, Oklahoma City, Oklahoma; [d] Department of Obstetrics and Gynecology, University of Pennsylvania, Philadelphia, Pennsylvania; [e] Department of Obstetrics and Gynecology, University of California–San Francisco, San Francisco, California; [f] Department of Obstetrics and Gynecology, Pennsylvania State University, Hershey, Pennsylvania; [g] Department of Obstetrics and Gynecology, Medical College of Georgia, Augusta University, Augusta, Georgia; [h] Department of Obstetrics and Gynecology & Center for Molecular Medicine and Genetics, Wayne State University, Detroit, Michigan; [i] Department of Reproductive Endocrinology and Infertility, Atrium Health, Charlotte, North Carolina; [j] Department of Gynecology and Obstetrics, Johns Hopkins University, Lutherville, Maryland; [k] Department of Urology, University of North Carolina, Chapel Hill, North Carolina; [l] Department of Biostatistics, Yale School of Public Health, New Haven Connecticut; [m] Department of Obstetrics and Gynecology, Biostatistics and Epidemiology University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma; [n] Department of Urology, University of Pennsylvania, Philadelphia, Pennsylvania; [o] Department of Urology, University of California–San Francisco, San Francisco, California; [p] Department of Obstetrics and Gynecology, University of Colorado, Aurora, Colorado; and [q] Collaborative Center for Statistics in Science, Yale School of Public Health, New Haven, Connecticut

Objective: To determine the reproducibility of the World Health Organization Fifth Edition (WHO5) strict grading methodology by comparing the percentage of morphologically normal sperm (PNS) recorded by the core laboratory with results obtained at the fertility centers participating in a multisite clinical trial.
Design: Secondary cohort analysis of data from the Males, Antioxidants, and Infertility trial.
Setting: Fertility centers.
Patient(s): Semen values of 171 men participating in a multicenter, double-blind, randomized, placebo-controlled trial evaluating the effect of antioxidants on male fertility.

Received January 3, 2022; revised February 22, 2022; accepted March 13, 2022.

K.C.B. reports honoraria for the American Urologic Association Oral Board Review Course. A.Z.S. reports grant from the NICHD/NIH for the submitted work; consulting fees from Seikagaku Corporation and Prima-Temp; and Editor-in-Chief, *F&S Reviews*. K.R.H. reports grant NIH U10HD077680 from the NIH/NICHD for the submitted work, grants R03HD101893 and R01HD100305 from the NIH; and consulting fees from AblaCare outside the submitted work. K.T.B. has nothing to disclose. M.I.C. has nothing to disclose. R.S.L. reports grants from NIH R01 HD091350-04, NIH/NCCIH R01AT009484-02, Guerbet USA, Data Coordinator Center for the RMN U10HD055925-10REV, Hass Avocado Board; Penn State Clinical and Translational Science Institute, NIH/NICHS R01 HD083323-04, NIH R01HD100630-01, National Center for Advancing Translational Sciences UL1 TR002014, NIH U10 HD038992-15 (Ext), NIH/NICHD U10 HD038992-10 (Ext), NIH/NHLBI R01 HL119245-05, and Patty Brisben Foundation for Women's Sexual Health; consulting fees from Insudd (2020), Bayer (2019), Fractyl (2019), AbbVie (2019), and Ferring (2018); honorarium/payment for lectures from the National Research Center for Assisted Reproductive Technology and Reproductive Genetics, Shandong University, Jinan, China; Member, Steering Committee, Eastern Siberia PCOS Epidemiology & Phenotype study, Scientific Center of Family Health and Human Reproduction, Irkutsk, Russian Federation; and Member, International Advisory Panel, Developing, disseminating and implementing a core outcome set for infertility, Funded by the Royal Society of New Zealand Catalyst Fund and the University of Auckland, New Zealand. M.P.D. reports grant U10HD039005 from the NIH/NICHD for the submitted work. S.A.K. reports funding from the NICHD and Endowed Chair to his institution for the submitted work, royalties from Springer, consulting fees from Taylor and Frances, and advisory board for KINBRE outside the submitted work. R.U. has nothing to disclose. V.L.B. has nothing to disclose. M.C. reports grant U10HD077844 from the NIH/NICHD for the submitted work. F.S. has nothing to disclose. R.W. has nothing to disclose. P.M. has nothing to disclose. J.F.S. has nothing to disclose. N.S. has nothing to disclose. H.Z. reports grants from the NIH and NSF for the submitted work.

This work was supported by National Institutes of Health (NIH)/Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Grants R25HD075737 (to N.S.), U10HD077844 (to A.Z.S.), U10HD077680 (to K.R.H., V.L.B.), 1U10 HD077841 (to M.I.C.), U10HD027049 (to C.C.), U10HD038992 (to R.S.L.), U10HD039005 (to M.P.D., R.U., S.K.), and U10HD055925 and UL1 TR001863 (to H.Z.).

Reprint requests: Karen Baker, M.D., Division of Urology, Duke University Medical Center DUMC 3146, Durham, North Carolina 27710 (E-mail: karen.baker@duke.edu).

Fertil Steril Rep® Vol. 3, No. 2, June 2022 2666-3341

https://doi.org/10.1016/j.xfre.2022.03.003

Intervention(s): Not applicable.
Main Outcome Measure(s): Strict morphology expressed as PNS as determined at each fertility center and the core central laboratory for the same semen sample.
Result(s): No correlation was found in the PNS values for the same semen sample between the core laboratory and fertility center laboratories either as a group or by individual site. Interobserver agreement was similarly low ($\kappa = 0.05$ and 0.15) between the core and fertility laboratories as a group for strict morphology, categorized by the WHO5 lower reference limits of 4% and 0, respectively. Moderate agreement was found between the core and 2 individual fertility laboratories for the cutoff value of 0 ($\kappa = 0.42$ and 0.57). The remainder of the comparisons demonstrated poor to fair agreement.
Conclusion(s): Strict morphology grading using the WHO5 methodology demonstrated overall poor reproducibility among a cohort of experienced fertility laboratories. This lack of correlation and agreement in the PNS values calls into question the reproducibility, and thereby the potential applicability, of sperm strict morphology testing. (Fertil Steril Rep® 2022;3:110–5. ©2022 by American Society for Reproductive Medicine.)
Key Words: Semen analysis, teratozoospermia, spermatozoa, quality control, male factor infertility

The semen analysis is the standard test for quantifying male reproductive fitness because of its accessibility, lack of invasiveness, and low cost. In addition to semen volume, sperm concentration, and assessment of sperm motility, the grading of sperm morphology, typically expressed as the percentage of morphologically normal sperm (PNS), is commonly reported as part of a standard semen analysis.

In 2010, the World Health Organization (WHO) endorsed the use of "strict" grading criteria and adopted a threshold value of 4% as the lower reference limit for PNS—citing "evidence supporting the relationship between the percentage of normal forms […] and fertilization rates in vivo" (1). The methodology is detailed in the *WHO Laboratory Manual for the Examination and Processing of Human Semen*, Fifth Edition (WHO5) (1). In brief, semen samples are mixed, fixed on duplicate slides, and stained. A total of 200 individual, randomly selected sperm per slide are graded as normal or abnormal based on the presence, size, and/or appearance of the head, midpiece, principle piece (tail), and excessive retained cytoplasm. The replicated slide is graded, and the PNS is calculated.

The broad overlap of semen parameters between fertile and infertile couples is a recognized limitation of semen analysis. Despite efforts to standardize methodology for grading sperm morphology, publications examining the impact morphology on spontaneous pregnancy, intrauterine insemination, and in vitro fertilization reach varying conclusions (2–10). Likewise, outcomes are inconsistent for studies correlating lifestyle and environmental exposures with sperm morphology (11–15). These divergent results are distressing to couples seeking fertility care and potentially confusing to practitioners. Additionally, the lay press, government, and scientific community increasingly recognize that inconsistent outcomes have the potential to erode confidence in both the scientific method and investigators (16–19).

One potential cause for variable outcomes reported for sperm morphology is poor reproducibility of the strict grading method. For the purposes of this manuscript, reproducibility is defined as the ability to duplicate the results of a prior study/test using the same material and procedures used by the original investigator (20). This definition aligns with the recommendations of the National Academies of Sciences, Engineering, and Medicine, which defines reproducibility as "obtaining consistent results using the same input data, computational steps, methods, and code; and conditions of analysis" (21). "Reproducibility is the minimum necessary condition for a finding to be believable and informative"(20), and as such, it is appropriate to ask if strict grading methodology delivers consistent results across a panel of experienced fertility laboratories.

The primary objective of this study was to investigate the reproducibility of the WHO5 strict sperm morphology grading by examining the degree of agreement between the PNS reported by the core laboratory and the values reported for the same semen sample by the site laboratories during the Reproductive Medicine Network (RMN)'s Males, Antioxidants, and Infertility (MOXI) trial. The secondary objective was to investigate patient factors associated with teratospermia.

MATERIALS AND METHODS

We performed a secondary analysis of data from the RMN MOXI trial (22). The full details of the MOXI trial are available at ClinicalTrials.gov (NCT02421887). In brief, the MOXI trial was a multicenter, randomized, double-blind, placebo-controlled trial analyzing the effect of antioxidants on semen parameters among couples with mild male factor infertility. Participants completed an extensive questionnaire that included tobacco, alcohol, and drug use; presence and laterality of varicocele; occupation; and exposures to pesticides, toxic chemicals, radiation, and heat. Participants provided semen samples to their fertility site at time of the randomization (visit 1) and after 90 days of treatment (visit 3). Participants received standardized instructions for precollection abstinence and collection methods. Standard semen analyses, including sperm morphology, were performed at each clinical site's andrology laboratory using the WHO5 methodology. Semen smears were then shipped to the RMN core laboratory for centralized grading of sperm morphology using the WHO5 "strict" criteria. All study sites are College of American Pathologists certified and perform internal quality control for sperm morphology. Both local and central laboratories were blinded to the treatment assignment of all participants. Approval for the study was obtained from the University of

Pennsylvania, which served as the single institutional review board for each site, with additional local site review.

Our study cohort is comprised of MOXI participants with strict morphology graded by the core laboratory at visit 1. We analyzed pairs of strict morphology values by comparing the PNS as graded by the core (PNScore) to the PNS as graded by the RMN clinical site laboratories (PNSsite) for the semen samples submitted by our cohort for MOXI visits 1 and 3. In additional to the PNS values, the pairs were also analyzed by the cutoff values of $<4\%$ or $\geq 4\%$ and 0 or >0, which correspond to the threshold values commonly used in clinical practice (4).

Continuous data were expressed as mean $\pm$ standard deviation and analyzed using the one-way ANOVA if the data were normally distributed. Otherwise, they were expressed as median with interquartile range and analyzed using the Kruskal-Wallis test. Categorical variables were presented as number and frequency. The relationship between the PNScore and PNSsite was investigated using the Pearson's correlation and Spearman's correlation for normally and not normally distributed data, respectively. Strong correlation was defined as $r \geq +0.7$. Agreement was calculated by simple κ, and meaningful agreement was defined a $\kappa > 0.6$. Univariable analyses were performed to evaluate the influence of patient characteristics, lifestyle and occupation exposures, and/or geographic location on the PNScore. Ethnicity, category of baseline semen abnormality, and occupational category were excluded from the overall analysis due to insufficient numbers. Race, pesticide exposure, radiation exposure, and study site were excluded from the analysis of the PNS of 0 due to insufficient numbers. Subsequently, a multivariable logistic regression model was created including the variables found to be significantly associated with the PNScore in the univariable analyses. Multivariable logistic regression analysis was conducted in a stepwise fashion, with a P value of $<.2$ to enter and P value of $<.05$ to stay. Due to the relatively large number of sites, the site variable was treated as a random effect in the multivariable analysis and was shown to be insignificant. Tables are presented with odds ratios and the corresponding 95% confidence intervals for the predictors for the logistic regression analysis. A P value of $<.05$ was considered statistically significant (23). All statistical tests were two-sided. Analyses were performed with SAS, version 9.4 (SAS Institute, Cary NC).

RESULTS

Baseline Characteristics of the Cohort

A total of 171 male subjects participated in the MOXI trial, of whom 126 had PNScore recorded in the dataset for visit 1 at the termination of the trial. These 126 subjects constitute the cohort for this study. Detailed descriptive statistics of the cohort are available in Supplemental Table 1 (available online). The mean age and median body mass index of subjects were 33.6 years and 28.6, respectively. The cohort was predominantly white (75.4%), college or higher educated (69.1%), and from households with an annual income of >75K (59.5%). The majority had primary infertility (65.9%), had no previous fertility treatment (74.6%), and did not currently smoke (88.9%). Nineteen percent of subjects reported an occupation potentially associated with gonadotoxic exposure (e.g., mining/extraction, healthcare, and farming). Less than 10% of subjects reported varicocele, recreational drug use, or known exposure to pesticides, toxic chemicals, radiation, or heat. Forty-five percent of subjects had >1 semen parameter below the WHO5 lower reference limit at enrollment in the MOXI trial, whereas 39.7% of subjects had isolated teratospermia (PNS $\leq$ 4).

The median PNScore at visit 1 was 5% (interquartile range, 3%–9%) (Table 1). Thirty-five percent of participants had a PNScore of <4%, and 9.5% had a PNScore of 0. The median PNScore did not differ significantly between the sites. The prevalence of teratospermia, defined by the threshold value of either <4 or 0, did not differ between the sites.

There were no statistically significant differences in patient characteristics, semen parameters at enrollment, or treatment group assignments across the RMN sites (Supplemental Tables 1 and 2).

Reproducibility of Strict Morphology Grading

At the time the MOXI trial was terminated, the dataset contained 110 pairs of PNS values consisting of 48 pairs of PNScore and PNSsite from visit 1 and 62 pairs of PNScore and PNSsite from visit 3. These pairs represent 77.8% of participants, 43.6% of semen samples, and 6 of the 9 fertility sites. We found no correlation between the paired PNScore and PNSsite scores overall (Table 2). Likewise, analysis by site demonstrated poor and nonsignificant correlation between strict morphology values reported by the core and those reported individual fertility sites (Table 2).

Teratospermia was then analyzed by the threshold values of PNS of <4% and 0 (Table 3). Interobserver agreement between the core laboratory and the sites as a group was poor for PNS of <4% and 0% ($\kappa = 0.05$ and 0.15, respectively). When analyzed by individual fertility center, there was moderate agreement between the core and 2 fertility centers when teratospermia was defined as a PNS of 0 ($\kappa = 0.42$ and 0.57). Interobserver agreement was poor to fair for the remainder of the sites and comparisons.

Analysis of Variables Associated with PNS

Univariable and multivariable analyses found no association found between the PNScore of <4% and any variable (data not shown). Younger age and self-reported exposure to toxic chemicals were associated with a PNScore of 0 during univariable analysis; however, only age remained significant during multivariable analysis (Supplemental Table 3, available online). Notably, there was no association between fertility center site and teratospermia.

DISCUSSION

This study demonstrates poor reproducibility of strict sperm morphology values between a central core laboratory and a cohort of experienced, licensed andrology laboratories grading the same semen samples. Not only did we find no correlation between strict morphology values between the core

TABLE 1

Percentage of morphologically normal sperm at visit 1.

Variables	All sites combined N = 126	Site 1 N = 27	Site 2 N = 7	Site 3 N = 15	Site 4 N = 9	Site 5 N = 4	Site 6 N = 12	Site 7 N = 35	Site 8 N = 10	Site 9 N = 7
PNS, median (range)	5.0 (3.0, 9.0)	5.0 (3.0, 8.0)	3.0 (2.0, 7.0)	8.0 (2.0, 11.0)	5.0 (4.0, 5.0)	5.0 (2.5, 6.5)	4.5 (3.5, 9.5)	5.0 (2.0, 9.0)	6.5 (5.0, 10.0)	8.0 (0.0, 19.0)
PNS < 4%, n (%)	43 (34.1)	11 (40.7)	4 (57.1)	6 (40.0)	1 (11.1)	1 (25.0)	3 (25.0)	13 (37.1)	2 (20.0)	2 (28.6)
PNS ≥ 4%, n (%)	83 (65.9)	16 (59.3)	3 (42.9)	9 (60.0)	8 (88.9)	3 (75.0)	9 (75.0)	22 (62.9)	8 (80.0)	5 (71.4)
PNS > 0, n (%)	114 (90.5)	26 (96.3)	6 (85.7)	14 (93.3)	9 (100.0)	3 (75.0)	12 (100.0)	29 (82.9)	10 (100.0)	5 (71.4)
PNS = 0, n (%)	12 (9.5)	1 (3.7)	1 (14.3)	1 (6.7)	0 (0.0)	1 (25.0)	0 (0.0)	6 (17.1)	0 (0.0)	2 (28.6)

Note: The percentage of morphologically normal sperm by strict grading criteria as assessed by core laboratory at Males, Antioxidants, and Infertility trial visit 1. The difference among sites did not reach statistical significance. PNS = percentage of morphologically normal sperm.

Baker. Reproducibility of strict morphology. Fertil Steril Rep 2022.

TABLE 2

Correlation of the percentage of morphologically normal sperm between the core and site laboratories for all visits.

		Core laboratory	
		# Specimens	Correlation
Site laboratory	All sites	110	r = 0.124[a], P = .20
	Site 1	31	r = 0.071[b], P = .71
	Site 2	11	r = 0.435[b], P = .18
	Site 3	29	r = 0.040[b], P = .84
	Site 7	8	r = 0.025[a], P = .95
	Site 8	19	r = -0.200[b], P = .41
	Site 9	12	r = 0.432[a], P = .16

Note: The results of strict morphology graded at sites 4, 5, and 6 were not incorporated into the dataset before the termination of the Males, Antioxidants, and Infertility trial.

[a] Spearman's correlation.

[b] Pearson's correlation.

Baker. Reproducibility of strict morphology. Fertil Steril Rep 2022.

laboratory and the fertility center laboratories either as a group or by individual site, but we also found no agreement between the core and site laboratories for teratospermia defined as a PNS of <4% and only poor agreement overall for teratospermia defined as a PNS of 0. Of note, teratospermia defined as a PNScore of <4% was not associated with any patient characteristics in our dataset. Teratospermia defined as a PNScore of 0 was associated with younger age and self-reported toxic chemical exposure during univariable analysis; however, only age remained a significant association during multivariable analysis. In light of the poor reproducibility of morphology grading, the clinical applicability of these associations should be interpreted with caution.

Our results show that assessing sperm morphology appears to remain a highly subjective exercise despite adoption of the strict grading criteria by the WHO in 1999 and addition of the WHO5 lower threshold value for normal morphology of ≥4% in 2010. Our finding is consistent with several studies that have documented high interobserver variability when assessing sperm morphology with the strict grading criteria.

In 2016, Punjabi et al. (24) published the outcomes of over 100 Belgian laboratories that participated in a thrice yearly voluntary external quality control (EQC) program for semen analysis over a 15-year period spanning the publication of WHO5. Two centrally prepared air-dried semen smears were sent to each laboratory per EQC event to assess the performance of strict morphology grading. The overall coefficient of variation (CV) for PNS was 79.4% for the duration of the study. The investigators noted that performance improved after adoption of the WHO5 methods; however, the variation in morphology grading remained "unacceptably high" with an extrapolated PNS CV of ≥40% in the years after 2010. It is notable that 20% of the results were discordant in the final year of the study when analyzed by PNS above or below the WHO5 threshold value.

Similarly, Filimberti et al. (25) analyzed the outcomes of 56 Tuscan laboratories that participated in a dedicated training program in the WHO5 methods and found that

the PNS CV was 88.6% despite targeted training. The investigators concluded that despite improvements in performance with training "the course was not sufficient to limit variability in the results of morphology, as the overall average CV of the laboratories remained very high." These studies highlight the marked variation in grading sperm morphology under "real-world" conditions despite adoption of the WHO5 methodology and participating in an EQC program.

In 2014, Wang et al. (26) reported good agreement for PNS but moderate or worse agreement for categorization of sperm defects among experienced graders. In brief, high-resolution pictures of 5,296 sperm from healthy donors were sent to 3 experienced graders at 3 different high volume centers. Each grader individually scored all sperm in accordance with the WHO5 criteria. The investigators found that the mean PNS was 20.87% and the CV was 4.8%. Agreement among the graders for the overall PNS was good ($\kappa = 0.47$–0.52). There was marked variation in scoring among graders based on by defect category (e.g., head defect) and specific defect type (e.g., tapered head) with CV varying between 6.8% and 15.6% for defect category and 11.2% and 133% for specific defect. This large degree of variation underscores the subjectivity inherent in grading sperm morphology even when experienced graders are assessing the same individual spermatozoa.

Our dataset demonstrated no association between teratospermia and geographic location (as defined by a fertility center). This finding echoes the results of Swan et al. (15) who found no difference in the PNS among 4 US cities despite a difference in concentration and motility. Likewise, Auger et al. (14) compared the morphology values for 1,001 men from 4 European cities and found no difference in the PNS. The investigators did find an association between poorer sperm morphology and occupational and lifestyle exposures (e.g., metal welding, alcohol consumption, and chemical spraying). In contrast, our analysis found no association between lifestyle exposure and a PNS of <4%. Younger age was associated with the absence of any morphologically normal sperm in our dataset during both univariable and multivariable analyses. As previously mentioned, exposure to toxic chemicals was associated with higher odds of a PNScore of 0 during univariable analysis, but this association did not remain significant during multivariable analysis. These associations should be interpreted with caution, however, given that the "true value" for PNS cannot be established due to the poor reproducibility of strict morphology grading between the core and site laboratories.

Limitations

This study is a secondary analysis—the primary trial was not designed to evaluate the reproducibility of morphological grading. Because the original MOXI trial was terminated early in accordance with a prespecified internal pilot study, our dataset was limited by the absence of data pairs because some PNSsite values were not entered into the dataset before data lock. The identity of the interpreting technician was not recorded at the core or site laboratories; therefore, the analysis of the performance of individual technicians is not possible. All sites were certified andrology laboratories with internal quality assurance/quality improvement programs, and there is ample literature assessing individual performance of sperm morphological grading using both the WHO5 and other grading criteria for experienced and inexperienced graders (14, 24–27). Therefore, we felt it was both timely and valuable to focus the reproducibility of the WHO5 method, rather than further scrutinize individual laboratory personnel. The broader applicability of our results is based on our presumption that the performance of strict morphology grading at RMN facilities is not inferior to other fertility centers and laboratories. Despite these limitations and presumptions, this study mirrors "real-world" conditions as training and quality control practices differ among laboratories, and the resulting variations in morphological grading may influence both generalizability of reproductive science and its applicability to patient care.

CONCLUSION

In conclusion, our analysis demonstrates that almost 10 years after the adoption the WHO5, inconsistencies remain in the scoring of sperm morphology, even among a cohort of highly experienced fertility laboratories grading the same semen smears. Our study found no correlation in PNS as a continuous variable and little to no agreement in PNS for clinically meaningful categories as defined by the WHO threshold value and by the complete absence of morphologically normal

TABLE 3

Level of agreement for the percentage of morphologically normal sperm between the core and site laboratories for all visits.

			Core laboratory	
		# Specimen	<4%	0
Site laboratory	All sites	110	$\kappa = 0.053$, $P = .16$	$\kappa = 0.146$, $P = .04$
	Site 1	31	$\kappa = 0.073$, $P = .28$	$\kappa = 0.073$, $P = .28$
	Site 2	11	$\kappa = 0.000$, $P = 1.00$	$\kappa = 0.421$, $P = .09$
	Site 3	29	$\kappa = 0.004$, $P = .96$	$\kappa = 0.074$, $P = .29$
	Site 7	8	$\kappa = 0.040$, $P = .69$	$\kappa = 0.000$, $P = 1.00$
	Site 8	19	$\kappa = 0.000$, $P = 1.00$	$\kappa = 0.000$, $P = 1.00$
	Site 9	12	$\kappa = 0.125$, $P = .37$	$\kappa = 0.571$, $P = .03$

Note: The results of strict morphology graded at sites 4, 5, and 6 were not incorporated into the dataset before the termination of the Males, Antioxidants, and Infertility trial. κ = simple kappa coefficient.

Baker. Reproducibility of strict morphology. Fertil Steril Rep 2022.

sperm. The poor reproducibility of strict sperm morphology grading calls into question the applicability of morphology values between laboratories and, by extension, the generalizability of strict morphology in assessing male reproductive potential and predicting treatment outcomes. Combined with recent publications demonstrating fecundity in the absence of morphologically normal sperm, the clinical relevance of strict sperm morphology seems increasingly uncertain.

REFERENCES

1. WHO Department of Reproductive Health and Research. World Health Organization. Clinical and Laboratory Standards Institute, Centers for Disease Control and Prevention (U.S.). Laboratory quality management system: handbook. 5th ed. Geneva: World Health Organization; 2011. p. 56–102.
2. Guzick DS, Overstreet JW, Factor-Litvak P, Brazil CK, Nakajima ST, Coutifaris C, et al. Sperm morphology, motility, and concentration in fertile and infertile men. N Engl J Med 2001;345:1388–93.
3. Slama R, Eustache F, Ducot B, Jensen TK, Jorgensen N, Horte A, et al. Time to pregnancy and semen parameters: a cross-sectional study among fertile couples from four European cities. Hum Reprod 2002;17:503–15.
4. Kovac JR, Smith RP, Cajipe M, Lamb DJ, Lipshultz LI. Men with a complete absence of normal sperm morphology exhibit high rates of success without assisted reproduction. Asian J Androl 2017;19:39–42.
5. Erdem M, Erdem A, Mutlu MF, Ozisik S, Yildiz S, Guler I, et al. The impact of sperm morphology on the outcome of intrauterine insemination cycles with gonadotropins in unexplained and male subfertility. Eur J Obstet Gynecol Reprod Biol 2016;197:120–4.
6. Lemmens L, Kos S, Beijer C, Brinkman JW, van der Horst FA, van den Hoven L, et al. Predictive value of sperm morphology and progressively motile sperm count for pregnancy outcomes in intrauterine insemination. Fertil Steril 2016;105:1462–8.
7. Kruger TF, Acosta AA, Simmons KF, Swanson RJ, Matta JF, Oehninger S. Predictive value of abnormal sperm morphology in in vitro fertilization. Fertil Steril 1988;49:112–7.
8. French DB, Sabanegh ES Jr, Goldfarb J, Desai N. Does severe teratozoospermia affect blastocyst formation, live birth rate, and other clinical outcome parameters in ICSI cycles? Fertil Steril 2010;93:1097–103.
9. Grow DR, Oehninger S, Seltman HJ, Toner JP, Swanson RJ, Kruger TF, et al. Sperm morphology as diagnosed by strict criteria: probing the impact of teratozoospermia on fertilization rate and pregnancy outcome in a large in vitro fertilization population. Fertil Steril 1994;62:559–67.
10. Hotaling JM, Smith JF, Rosen M, Muller CH, Walsh TJ. The relationship between isolated teratozoospermia and clinical pregnancy after in vitro fertilization with or without intracytoplasmic sperm injection: a systematic review and meta-analysis. Fertil Steril 2011;95:1141–5.
11. Chen Q, Yang H, Zhou N, Sun L, Bao H, Tan L, et al. Phthalate exposure, even below US EPA reference doses, was associated with semen quality and reproductive hormones: prospective MARHCS study in general population. Environ Int 2017;104:58–68.
12. Bloom MS, Whitcomb BW, Chen Z, Ye A, Kannan K, Buck Louis GM. Associations between urinary phthalate concentrations and semen quality parameters in a general population. Hum Reprod 2015;30:2645–57.
13. Swan SH. Semen quality in fertile US men in relation to geographical area and pesticide exposure. Int J Androl 2006;29:62–8, discussion 105–8.
14. Auger J, Eustache F, Andersen AG, Irvine DS, Jorgensen N, Skakkebaek NE, et al. Sperm morphological defects related to environment, lifestyle and medical history of 1001 male partners of pregnant women from four European cities. Hum Reprod 2001;16:2710–7.
15. Swan SH, Brazil C, Drobnis EZ, Liu F, Kruse RL, Hatch M, et al. Geographic differences in semen quality of fertile U.S. males. Environ Health Perspect 2003;111:414–20.
16. Federal Register. Strategy for American innovation. Available at: https://www.federalregister.gov/documents/2014/07/29/2014-17761/strategy-for-american-innovation. Accessed July 21, 2021.
17. Carroll AE. Science needs a solution for the temptation of positive results. Available at: https://www.nytimes.com/2017/05/29/upshot/science-needs-a-solution-for-the-temptation-of-positive-results.html. Accessed July 21, 2021.
18. The Economist. Are research papers less accurate and truthful than in the past?. Available at: https://www.economist.com/science-and-technology/2018/03/17/are-research-papers-less-accurate-and-truthful-than-in-the-past. Accessed July 21, 2021.
19. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med 2016;8:341ps12.
20. Bollen K, Cacioppo JT, Kaplan RM, Krosnick JA, Olds JL. Social, behavioral, and economic sciences perspectives on robust and reliable science. Available at: https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf. Accessed July 21, 2021.
21. National Academies of Sciences. Engineering, and Medicine. Reproducibility and replicability in science. Available at: https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science. Accessed July 21, 2021.
22. Steiner AZ, Hansen KR, Barnhart KT, Cedars MI, Legro RS, Diamond MP, et al. The effect of antioxidants on male factor infertility: the Males, Antioxidants, and Infertility (MOXI) randomized clinical trial. Fertil Steril 2020;113:552–60.e3.
23. Andrade C. The P value and statistical significance: misunderstandings, explanations, challenges, and alternatives. Indian J Psychol Med 2019;41:210–5.
24. Punjabi U, Wyns C, Mahmoud A, Vernelen K, China B, Verheyen G. Fifteen years of Belgian experience with external quality assessment of semen analysis. Andrology 2016;4:1084–93.
25. Filimberti E, Degl'Innocenti S, Borsotti M, Quercioli M, Piomboni P, Natali I, et al. High variability in results of semen analysis in andrology laboratories in Tuscany (Italy): the experience of an external quality control (EQC) programme. Andrology 2013;1:401–7.
26. Wang Y, Yang J, Jia Y, Xiong C, Meng T, Guan H, et al. Variability in the morphologic assessment of human sperm: use of the strict criteria recommended by the World Health Organization in 2010. Fertil Steril 2014;101:945–9.
27. Eustache F, Auger J. Inter-individual variability in the morphological assessment of human sperm: effect of the level of experience and the use of standard methods. Hum Reprod 2003;18:1018–22.