The Agreement Between the Two- and Three-Step Method for Identifying Subtle Menstrual Disturbances

Noordhof D.A.[1], Taylor M.Y.[2], De Martin Topranin V.[1,3], Engseth T.P.[2], Sandbakk Ø.[1,2], Osborne J.O.[2]

[1] Department of Neuromedicine and Movement Science, Centre for Elite Sports Research, Norwegian University of Science and Technology, Trondheim, Norway.

[2] School of Sport Sciences, UiT The Arctic University of Norway, Tromsø, Norway.

[3] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway.

As accepted for publication in International Journal of Sports Physiology and Performance,

Corresponding author:

Dionne A. Noordhof

Centre for Elite Sports Research

Department of Neuromedicine and Movement Science

Norwegian University of Science and Technology

dionne.a.noordhof@ntnu.no

Preferred running head: The Two- vs. the Three-Step Method

Abstract word count: 249

Text only word count: 1798

Number of figures and tables: 3

Abstract

Recent methodological recommendations suggest the use of the 'three-step method', consisting of calendar-based counting, urinary ovulation testing, and serum blood sampling for the identification of subtle menstrual disturbances (SMDs). However, the use of the three-step method is not always feasible, so a less demanding combination of calendar-based counting and urinary ovulation testing i.e., the two-step method, may be a viable alternative. **Purpose**: To investigate the agreement between the two- and three-step methods for the detection of SMDs. **Methods**: Menstrual cycles (MCs, 98) of 59 athletes were assessed using the two- and three-step methods. Regular length MCs (i.e., ≥21 and ≤35 days) were classified as either having no SMD (luteal phase length ≥10 days, mid-luteal progesterone concentration ≥16 $nmol \cdot L^{-1}$, and being ovulatory), or having an SMD (e.g., short luteal phase (<10 days), inadequate luteal phase (mid-luteal progesterone concentration <16 $nmol \cdot L^{-1}$) or being anovulatory). Method agreement was assessed using McNemar's test and Cohen's kappa (κ). **Results**: Substantial agreement was observed between methods (κ=.72; 95%CI ([.53, .91]), but the two-step method did not detect all MCs with a SMD, resulting in evidence of systematic bias ($\chi 2=5.14$; $p=.023$). The two-step method detected 61.1% of MCs that had an SMD ([51.4, 70.8]), as verified using the three-step method, and correctly identified 100% of MCs without an SMD. **Conclusions**: MCs classified as being disturbed using the two-step method could be considered valid evidence of SMDs. However, MCs classified without SMDs do not definitively confirm their absence, due to the proven underdetection via the two-step method.

Introduction

In naturally menstruating women, the sex hormones, estrogen and progesterone, fluctuate cyclically throughout a 21–35-day menstrual cycle (MC). A eumenorrheic MC comprises four distinct hormonal environments: the early follicular phase; the late follicular phase; the ovulatory phase; and, the mid-luteal phase.[1,2] However, about 52% of exercising women has been shown to suffer from subtle menstrual disturbances (SMDs).[3] SMDs are characterized by a regular MC length, but have underlying hormonal irregularities,[4] resulting in a shortened luteal phase (i.e., <10 days between the day after a positive ovulation test until the first day of menses of the subsequent MC)[3,5] and/or an inadequate luteal phase (i.e., mid-luteal progesterone concentration <16 $nmol \cdot L^{-1}$)[1,2] or anovulation.[1] The verification of these different MC phases, as well as the detection of menstrual disturbances are accomplished using three additive ‘steps’, which become progressively more invasive and expensive: 1) calendar-based counting to assess MC length; 2) urinary ovulation testing to confirm ovulation and determine the length of the luteal phase; and, 3) serum hormone verification to confirm the mid luteal rise in progesterone.[6] The recommended method, known as the ‘three step method’ uses all three measurements, and can therefore identify both severe menstrual disturbances, such as amenorrhea (via the first step), and SMDs (via the two- and three-step methods, see Figure 1).

The identification of an inadequate luteal phase requires a serum progesterone measurement, rendering it detectable solely using the three-step method (Figure 1). In contrast, the other two SMDs are detectable with calendar-based counting and urinary ovulation testing (i.e., the two-step method; Figure 1).[1,3,4] Concerningly, all SMDs are undetectable without urinary ovulation testing and/or serum hormone testing, and so naturally-menstruating athletes with regular MCs lasting 21-35 days may remain unaware of the presence of an SMD without thorough MC monitoring.[7]

Early detection of SMDs is important for athletes and their support team, since this could prevent athletes from progressing along the continuum from SMDs towards more severe menstrual disturbances, such as oligomenorrhea and amenorrhea.[7] The clinical consequences of amenorrhea, as a result of prolonged or severe low energy availability, are well documented and include among others, infertility and low bone mineral density.[8] For sport scientists it is also important to detect SMDs, since those data should be excluded *a posteriori;* as depressed hormonal concentrations can result in a potentially confounding effect when investigating aspects related to MC phase and, for example, training.[2] Although the use of the three-step method results in the detection of all SMDs and strengthens research quality, this method is also invasive, expensive, and time demanding. Therefore, a less-invasive, cheaper, and faster method, consisting of calendar-based counting and urinary ovulation testing (i.e., the two-step method) would be preferable for both sports practice and sport scientists. Thus, the aim of the current study was to investigate the agreement between the two- and three-step methods for identifying SMDs.

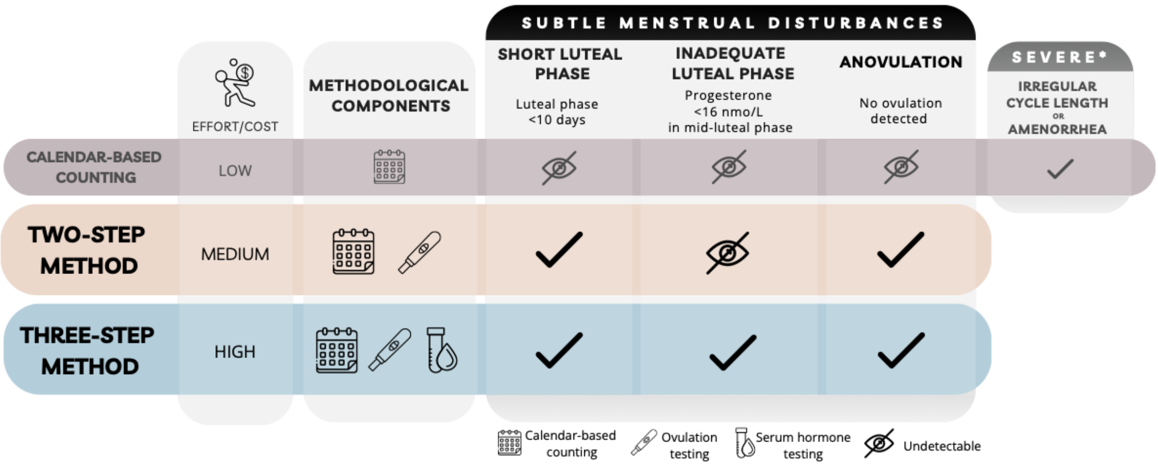

Figure 1. The detection of subtle menstrual disturbances using the two- and three-step methods.

Methods

Participants

Endurance-trained females (n=63) were recruited as part of the Female Endurance Athlete (FENDURA) project.[9] The inclusion criteria were: 1) having self-reported MC lengths of 21-35 days for the prior 6 months; 2) no hormonal contraceptive use in the prior 3 months; 3) aged 17-45 years; and, 4) active in an endurance sport (Tier 1 n=8, Tier 2 n=33, Tier 3 n=17, Tier 4 n=1)[10]. Exclusion criteria were: 1) having an injury or illness that prevented them from training; 2) having a known clinically diagnosed menstrual disorder, 3) being pregnant or within 12 months following parturition[1]. The Regional Committee for Medical and Health Research Ethics waived the requirement for ethical approval for this study (Project-ID: 230505). The project was performed according to institutional ethical requirements and was pre-approved by the Norwegian Centre for Research Data (NSD, Project-ID: 955558).

Methodology

Female athletes recorded their first day of menses in an online training diary (the Norwegian Olympic Sport Centre (Olympiatoppen) training diary, or BESTR training diary (Oslo, Norway)), used a Clearblue Digital Ovulation test (SPD Swiss Precision Diagnostics GmbH, Geneva, Switzerland) from day 8 of their MC until a positive urinary ovulation test or the start of menses in the subsequent MC, and provided a single venous blood sample in the mid-luteal phase (7-9 days after the day of a positive ovulation test).[1,2] Blood was sampled from the antecubital vein after an overnight fast and collected in a serum separator tube (Vacutainer SST 8.5 mL, BD, Franklin Lakes, NJ, United States). The sample clotted for 30 min, before being centrifuged (4200 revolutions$\cdot$min^{-1} for 10 min). Subsequently, the serum was frozen at -80 °C until analysis. Analyses were performed at the University Hospital of Northern Norway (accredited according to ISO/IEC 15189), Tromsø, Norway. Serum progesterone concentrations were determined using liquid chromatography-tandem mass spectrometry.

MCs with irregular cycle length (<21 or >35 days) were identified and excluded. Regular length MCs were subsequently classified as having or not having an SMD using the two- and three-step method. A MC classified as having no SMD was defined as a cycle lasting ≥ 21 and ≤ 35 days, with a positive urinary ovulation test and mid-luteal progesterone concentration ≥16 nmol$\cdot$L^{-1} (see Figure 1).

Statistical analysis

All analyses were undertaken using R[11] in the RStudio[12] environment. A systematic difference between methods was assessed using McNemar's test with continuity correction.[13] The agreement between the three- vs two-step method was evaluated using Cohen's kappa (κ).[13] The strength of agreement was categorized as: *slight* ($0.00 \leq \kappa \leq 0.20$), *fair* ($0.21 \leq \kappa \leq 0.40$), *moderate* ($0.41 \leq \kappa \leq 0.60$), *substantial* ($0.61 \leq \kappa \leq 0.80$), and *almost perfect* ($0.81 \leq \kappa$).[14] The validity of the two-step method, when compared to the three-step method, was evaluated via test sensitivity, specificity, positive predictive value, and negative predictive values (see Figure 2). Estimates are reported with 95% confidence intervals.

Results

Seven MCs were excluded from the analysis because of irregular menstrual cycle length (polymenorrhea: n=1; oligomenorrhea: n=6). In total, 98 unique MCs from 59 endurance athletes (age 28.9 [7.9] y) were assessed. There was evidence of systematic bias between the two methods ($\chi^2=5.14$; $p=.023$), with the two-step method under detecting luteal-phase inadequate MCs in the absence of other SMDs. There was also *substantial* agreement between methods, with a Cohen's κ of .72 ([.53 to .91]).

The two-step method, when compared to the three-step method, correctly identified 61.1% ([51.4, 70.8]) of MCs with an SMD (i.e., test sensitivity; see Figure 2 and Table 1), while it correctly identified 100% of MCs without an SMD (i.e., test specificity).

The probability that a MC was correctly classified with an SMD using the two-step method was 100% (i.e., positive predictive value). On the other hand, a MC with no SMD via the two-step method, had a 92.0% ([86.4, 97.3]) probability of being correctly classified (i.e., having no SMD; negative predictive value), while having an 8.0% ([2.7, 13.4]) chance of a false negative, i.e., having an SMD.

		Three-step method		
		SMD	No SMD	
Two-step method	SMD	**True positive**	**False positive**	← Row for *positive predictive value*
	No SMD	**False negative**	**True negative**	← Row for *negative predictive value*
		↑ Column for *sensitivity*	↑ Column for *specificity*	

Figure 2. Schematic representation of the calculations for sensitivity, specificity, positive-, and negative predictive value (adapted from Trevethan[17]).

SMD, subtle menstrual disturbance.

Table 1. Contingency table of frequencies for subtle menstrual disturbances (SMD) using the three- (gold standard) and the two-step method.

		Three-step method		
		SMD	No SMD	Total
Two-step method	SMD	11	0	11
	No SMD	7	80	87
	Total	18	80	**98**

Discussion

The aim of the current study was to investigate the agreement between the two- and three-step methods for the detection of SMDs. Our key findings showed evidence of systematic bias, due to the limitation of the two-step method to detect an inadequate luteal phase, although there was also *substantial* agreement between methods. The two-step method accurately classified all eumenorrheic cycles (MCs without a SMD), and correctly detected SMDs in 61.1% of all cycles with a true SMD, as verified with the three-step method. However, a MC classified as eumenorrheic by the two-step method had an 8.0% chance of an undetected inadequate luteal phase.

To the best of our knowledge, this is the first study that assessed the agreement between the two- and three-step methods for identifying SMDs. While substantial agreement was observed between methods, the two-step method did not detect 7 out of 18 cycles with a SMD, resulting in a systematic difference between methods. When classifying individual MCs, the 100% positive predictive value of the two-step method assures that the detection of a SMD using this method is always correct (i.e., no false positives). However, the two-step method resulted in an 8% underdetection of SMDs, as it failed to identify inadequate luteal phases. Consequently, the exclusive use of the two-step method in research may lead to inclusion of cycles with an inadequate luteal phase in data analyses, which could mask potential changes in, for instance, exercise performance between MC phases.[2] Besides, failure to accurately identify such SMDs may have health and performance implications for athletes, particularly if undetected menstrual disturbances persist over an extended period or progress into severe menstrual disturbances.[4,15]

Recent methodological guidelines have highlighted the importance of implementing the three-step method in sports science research for both MC phase determination and the detection of menstrual disturbances.[1,2] However, a shortened luteal phase (i.e., detectable via the two-step method) was not mentioned in these guidelines as an indicator of an SMD. While a short luteal phase (i.e., clinical luteal phase deficiency) and inadequate luteal phase (i.e., biochemical luteal phase deficiency) may operate through separate mechanisms,[5] some overlap between the two has been found as well.[3,5] Therefore, the detection of short luteal phases via the two-step method may also coincidentally capture MCs with inadequate luteal phases. It should be noted that the exact performance of the two-step method will be influenced by the sample included, the method used to detect ovulation, and the definitions of the different SMDs. Future research should therefore evaluate the effect of different ovulation-detection methods and SMD definitions on the performance of the two-step method, as well as the likelihood of a coincident presence of both short- and inadequate luteal phases in a more homogeneous athletic sample.

Practical applications

The three-step method is the preferred method when investigating the influence of menstrual cycle phase on, for instance, training. However, due to the high demands related to blood sampling, the two-step method, as presented here, is a viable alternative. In sports practice, assessing MC status of one MC might not be reflective of MC status throughout the entire season, due to large changes in physical and psychological stress influencing the functioning of the hypothalamus-pituitary-ovarian axis.[16] Thus, the two-step method could be used as an option for monitoring over an extended period. When using the two-step method, MCs classified with an SMD should be considered valid evidence of an SMD. However, the two-step method may incorrectly classify MCs as ‘eumenorrheic' (i.e., without an SMD) despite the presence of an undetected luteal phase deficiency. Therefore, it is essential to convey this potential limitation of the two-step method when communicating MC results with coaches, support staff, and athletes and when presenting such results in scientific publications.

Conclusions

There is substantial agreement between the two- and three-step methods for identifying SMDs. However, there was also systematic bias between methods, with the two-step method correctly classifying cycles that are considered eumenorrheic but only identifying SMDs in 61.1% of all cycles with a true SMD diagnosis, as verified using the three-step method.

Acknowledgements

The authors thank the participants for their participation in one of the FENDURA studies, the bioengineers and laboratory technicians (Anna Lena Müller, Dita Valsdottir, Hege Nymo Østgaard, Heleen Docter, Liv Landmark Nordengen, Madeleine-Emilie Killingstad) for their assistance during the data collection and the University Hospital of Northern Norway (Guri Grimnes, Ole Martin Fuskevåg, Silja Breivik) for their support with processing and analyzing blood samples.

This study was funded by the Tromsø Research Foundation (Project-ID: 19_FENDURA_BW) and UiT The Arctic University of Norway.

References

1. Elliott-Sale KJ, Minahan CL, de Jonge XAKJ, et al. Methodological considerations for studies in sport and exercise science with women as participants: a working guide for standards of practice for research on women. *Sports Med*. Published online March 16, 2021. doi:10.1007/s40279-021-01435-8

2. Janse de Jonge X, Thompson B, Han A. Methodological recommendations for menstrual cycle research in sports and exercise. *Med Sci Sports Exerc*. Published online June 25, 2019. doi:10.1249/MSS.0000000000002073

3. De Souza MJ, Toombs RJ, Scheid JL, O'Donnell E, West SL, Williams NI. High prevalence of subtle and severe menstrual disturbances in exercising women: confirmation using daily hormone measures. *Hum Reprod*. 2010;25(2):491-503. doi:10.1093/humrep/dep411

4. Prior JC. Adaptive, reversible, hypothalamic reproductive suppression: More than functional hypothalamic amenorrhea. *Front Endocrinol*. 2022;13. Accessed January 18, 2024. https://www.frontiersin.org/articles/10.3389/fendo.2022.893889

5. Schliep KC, Mumford SL, Hammoud AO, et al. Luteal phase deficiency in regularly menstruating women: prevalence and overlap in identification based on clinical and biochemical diagnostic criteria. *J Clin Endocrinol Metab*. 2014;99(6):E1007-1014. doi:10.1210/jc.2013-3534

6. Schaumberg MA, Jenkins DG, Janse de Jonge XAK, Emmerton LM, Skinner TL. Three-step method for menstrual and oral contraceptive cycle verification. *J Sci Med Sport*. 2017;20(11):965-969. doi:10.1016/j.jsams.2016.08.013

7. Taim BC, Ó Catháin C, Renard M, Elliot-Sale KJ, Madigan S, Ní Chéilleachair N. The prevalence of menstrual cycle disorders and menstrual cycle-related symptoms in female athletes: a systematic literature review. *Sports Med*. Published online June 30, 2023. doi:10.1007/s40279-023-01871-8

8. Mountjoy M, Ackerman KE, Bailey DM, et al. 2023 International Olympic Committee's (IOC) consensus statement on relative energy deficiency in sport (REDs). *Br J Sports Med*. 2023;57(17):1073-1097. doi:10.1136/bjsports-2023-106994

9. The FENDURA research project | UiT. Accessed December 5, 2023. https://uit.no/research/fendura?p_document_id=685241&Baseurl=%2Fresearch%2F

10. McKay AKA, Stellingwerff T, Smith ES, et al. Defining training and performance caliber: a participant classification framework. *Int J Sports Physiol Perform*. 2021;17(2):317-331. doi:10.1123/ijspp.2021-0451

11. R Core Team. R: A language and environment for statistical computing. Published online 2021. https://www.r-project.org/

12. RStudio: Integrated development for R. Published online 2023. http://www.rstudio.com/

13. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology. *Theriogenology*. 2010;73(9):1167-1179. doi:10.1016/j.theriogenology.2010.01.003

14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. doi:10.2307/2529310

15. De Souza MJ. Menstrual disturbances in athletes: a focus on luteal phase defects. *Med Sci Sports Exerc*. 2003;35(9):1553-1563. doi:10.1249/01.MSS.0000084530.31478.DF

16. Hackney AC, ed. *Sex Hormones, Exercise and Women: Scientific and Clinical Aspects*. Springer International Publishing; 2023. doi:10.1007/978-3-031-21881-1

17. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health*. 2017;5:307. doi:10.3389/fpubh.2017.00307