ORIGINAL ARTICLE

Interobserver variability of modified Ferriman–Gallwey hirsutism score in a Turkish population

Murat Api · Beyhan Badoglu · Aysu Akca · Olus Api · Husnu Gorgen · Ahmet Cetin

Abstract

Objective The aim of the study was to evaluate the inter-observer variability of modified Ferriman–Gallwey (mFG) hirsutism scores on each body area in a Turkish population.
Design A cross-sectional study of simultaneous mFG scoring design was used. Observers did not make any interview with the subjects and were masked to the previous score results. Analyses included percentage of agreement, kappa coefficients, the Bland and Altman plot, confidence intervals, minimum and maximum kappa coefficients.
Setting The study was performed at a teaching and research hospital.
Patients Hundred and twenty-one Turkish women without any complaints of excessive body hair were studied.
Interventions Interventions included two special trained physicians, simultaneous and independent mFG scoring.
Main outcome measures The main outcome measures were mFG scores in each body area.
Results Agreement analysis demonstrated that the scores of the two physicians were quite concordant. The mean kappa value for nine body area was 0.744 and the highest kappa values from the upper back and the lowest kappa values from the upper lip revealed κ = 0.847, κ = 0.585, respectively. The highest (upper lip) and the lowest (arm) mean range scores for the two researchers among the 9 areas were 1.46–1.55 and 0.17–0.12, respectively. Only 68.6 and 67.8% of the mFG scores observed by each of the two observers were equal or below 8.
Conclusion The mFG scoring system was found to be clinically useful. The upper lip was observed to have the highest score of androgen sensitive area of the body as well as the highest interobserver variability. The cut-off value to establish the diagnosis of hirsutism should be population-specific.

Keywords Ferriman–Gallwey score · Hirsutism · Interobserver variability

M. Api (✉) · B. Badoglu · A. Akca · O. Api · H. Gorgen · A. Cetin
Department of Obstetrics and Gynecology,
Haseki Education and Research Hospital, Istanbul, Turkey
e-mail: muratapi@hotmail.com

O. Api
Department of Obstetrics and Gynecology,
Dr. Lutfi Kirdar Kartal Education and Research Hospital,
Istanbul, Turkey

Introductıon

Hirsutism is defined as the excessive growth of thick, dark hair in body parts where hair growth in women is normally absent or minimal. Such male pattern—terminal hair growth usually occurs in androgen—stimulated locations such as chin, face and chest.

What is considered as hirsutism may be considered normal in another setting according to ethnicity and cultural differences. For instance, women from Mediterrenean region have more facial and body hair than women from North Europe and Asia. Hirsutism—by itself, is a benign condition primarily of cosmetic concern. However, when hirsutism is accompanied by masculinizing signs or symptoms, it may be a manifestation of a serious underlying disorder [1].

In extensive search of the literature, it's observed that the Ferriman–Gallwey scoring system has been used to score the excess male pattern body hair since 1961. Additionally, the studies evaluating medical treatments for hirsutism, particularly use this instrument [2–7].

Facial and body terminal hair growth in a male-like pattern in women is the principal clinical sign of hyperandrogenism. Although its definition remains unclear, it is reported to affect 5–10% of women surveyed [8, 9].

The presence of hirsutism is extremely disturbing for women, with a significant negative impact on their psychosocial status [10, 11].

Visual methods of determining the degree of hirsutism usually follow those originally described in 1961 by Ferriman and Gallwey [12]. In their study, these investigators scored the density of terminal hairs at 11 different body sites (i.e., upper lip, chin, chest, upper back, lower back, upper abdomen, lower abdomen, arm, forearm, thigh, and lower leg). In each of these areas a score of 0 (absence of terminal hairs) through 4 (extensive terminal hair growth) was assigned. Hair growth over the forearm and lower leg was noted to be less sensitive or indifferent to androgens, and subsequent modifications of the Ferriman–Gallwey method have deleted scoring of these areas [13, 14]. Scoring of hair growth in the sideburn area, lower jaw and upper neck, and buttocks have been included in some other scoring systems [15].

The modified (i.e., only 9 body areas considered) Ferriman–Gallwey scoring system is the method in general use for visually scoring excess terminal body or facial hair growth for the clinical or investigational assessment of hirsutism.

Eventhough, several other objective instruments are defined (i.e., photography of body areas, microscopic assessment of hair diameter with extensive counting of shafts, computerized assessment of photographic evaluations, and others), they are impractical, complex, costly, or difficult to use [16].

The ease of use and low cost of the Ferriman–Gallwey system make it a potentially attractive tool. Despite its widely acceptance, the Ferriman–Gallwey system has a lot of limitations due to its subjectivity in its nature. The system can be affected by the operator who applied the score (nurse, technician, junior or senior physician or even patient herself), or which Ferriman–Gallwey system is used (the original score, modified score, reduced number of body area). Therefore there seems to be a need for a standardized, easily applicable, less costly, valid and reliable score. To our knowledge in the medical literature review since 1961, interobserver variability analysis has not been performed for the score.

The purposes of the present study were to define; (1) the degree of facial and body terminal hair, as assessed by the modified Ferriman–Gallwey (mFG) score, in a sample of women from the Turkish population without the complaint of hirsutism; (2) to assess the performance characteristics and interobserver agreement of scoring by the mFG and (3) the population-specific cut-off values of the instrument.

Materials and method

Hundred and twenty-one Turkish women without the complaints of hirsutism between 13 and 80 years of age participated in this trial. Each patient signed an informed consent in accordance with local hospital institutional review board approval of the protocol. All patients met the inclusion criteria (having no complaints of hirsutism or any endocrinologic disease that might cause high androgen levels or any other endocrine or chronic disorders such as diabetes, Cushing's syndrome, etc.). The mFG map scoring system has nine domains depicting portions of the body (upper lip, chin, chest, upper back, lower back, upper abdomen, lower abdomen, arm and thigh). There are five categories graded from 0 to 4 using an ordinal scale within each body surface domain. Total scores are obtained by adding the scores from all domains. The maximum score is 36. Our experience revealed that the interobserver agreement was inconsistent between the two researchers before being trained about the scoring system. Once the principal investigator (M.A) demonstrated that his intraobserver agreement was within 3 points (15%), two research residents were trained by him and were shown to agree with him within 15% before the study has begun. Then the principal investigator became blinded to the results of two residents' scoring until the end of the trial. Two observers (each observer were senior residents) independently and blindly scored each patient's hirsutism using the mFG map.

Assessment of interobserver variation in our study was designed in a way, to minimize ascertainment bias in order to determine accurate interobserver agreement. Bias can potentially be introduced by patients themselves if unintentionally they declare their laboratory findings to the researcher or by the investigators by learning the features about their patients. Since being aware of the laboratory results can be associated with ascertainment bias, the researchers were not permitted to see or ask about laboratory results of the cases if they had any. Consequently, this investigation was conducted with proper masking to avoid ascertainment bias.

Statistical analysis

On the basis of previously published studies assesing Ferriman–Gallwey scores in the hirsute population, we accepted that a 15% difference between the scores which corresponds to a difference of three points in the scoring system, would be a clinically significant variation between the investigators. Upon sample size calculation, it was found out that a total of 113 patients yielded a power of 80% at a type I error of of 0.05. SPSS version 13.0 for Windows (SPSS Inc, Chicago, IL, USA) was used for statistical analysis. Descriptive statistics were performed for each

variable. Agreement analysis was performed using the Kappa coefficient. The Bland and Altman plot was used to reveal a relationship between the differences and the averages, to look for any systematic bias and to identify possible outliers. We tested for normality of distribution by Shapiro–Wilke test of all variables. Modified Ferriman–Gallwey scores ranged between 11 and 34 for both the patient and the observer. The scores were normally distributed and were well represented across the range.

Parametric analysis was used to compare the normally distributed variables, and non-parametric analysis was used when significant deviation from normality was detected.

Results

Two observers successfully scored 121 women simultaneously by the modified Ferriman and Gallwey scoring system and agreement analyses demonstrated that the scores were quite concordant with each other. The highest score was given for the upperlip and the lowest for the arm (Fig. 1).

Demographic parameters of 121 cases are shown in the Table 1. Both observers completed the survey by mFG scoring with a 100% success. All women were white and Caucasian in the ethnic origin. None of them had any complaints of hirsutism either when asked or as the presenting symptom to the hospital.

The kappa values on the average for each body area were shown in Fig. 2. Agreement analysis demonstrated that the two observers scores were quite concordant. The mean kappa value for nine body areas was 0.744 and the highest

Table 1 Demographic parameters of the study group

	Mean ± SD
Age (year)	36.43 ± 13.76
Body mass index (kg/m^2)	26.55 ± 3.81
Gravida	2.88 ± 2.12
Parity	2.08 ± 1.4

kappa values from the upper back and the lowest kappa values from upper lip were to be 0.847, 0.585, respectively. The highest (upper lip) and the lowest (arm) mean scores for two researchers among the 9 areas were 1.46–1.55 and 0.17–0.12, respectively.

As it is shown in the histogram of 242 observation of measurements from 121 subjects, the mean mFG total score was 6.814 ± 5.46. The frequency distribution of 242 measurements obtained from 121 subjects by the two observers was shown in Fig. 3. According to the Gaussian distribution rule, 1.96 × SD contains the 95% of area under the curve of subjects. The 95th percentile cut-off value of our study group has been computed and it was found to be 10.71 (1.96 × 5.46533). In the Turkish population studied by each observer, only 68.6 and 67.8% of the population scores were equal or less than 8 for total scores.

The Bland and Altman graph revealed that there was a good relationship between the differences. Since most of the differences were within mean ± 1.96 SD, the difference between the total scores obtained from the observers were assessed to be clinically unimportant. Therefore scores of the two observers might be used interchangeably (Fig. 4).

Conclusion

In their original report, Ferriman and Gallwey noted that if only the nine hormonal (androgen sensitive) skin areas (i.e., excluding the lower leg and forearms) were considered, 9.9% of their 161 women had a score above 5, 4.3% had a score above 7, and 1.2% had a score greater than 10 [12].

From these data, a score of 8 or more has been considered to represent hirsutism. It should be kept in mind that these studies were performed predominantly in white populations. Although racial/ethnic differences in the number, distribution, or androgen sensitivity of hair follicles in normal individuals remain to be better defined, information regarding the prevalence of hirsutism in different racial groups is scanty.

There is no concensus in the medical literature for how many body regions are to be included in the scoring systems. While there is a study by Derksen et al. who evaluated 12 body regions, another study suggested only 2 body regions for the definition of hirsutism [17, 18].
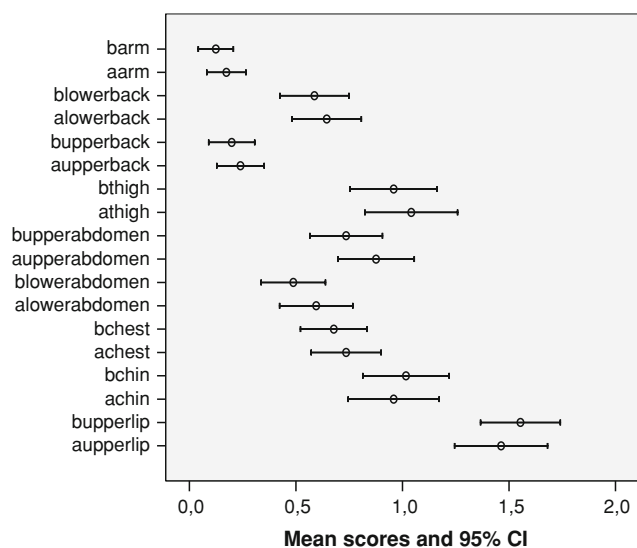

Fig. 1 Means and 95% confidence intervals of modified Ferriman–Gallwey scores for each body area of two observers. First letter of each body area (a and b) represents the different observers

Fig. 2 Interobserver agreement of researchers on nine body areas of the modified Ferriman–Gallwey hirsutism score. Values on *bars* represent the kappa values for each body area
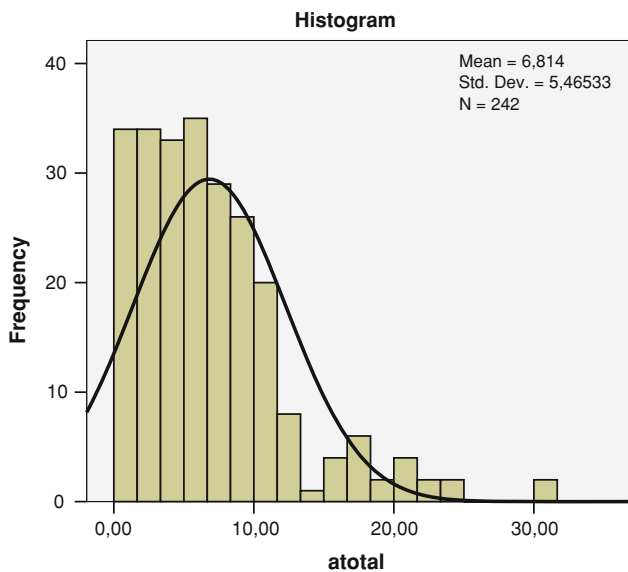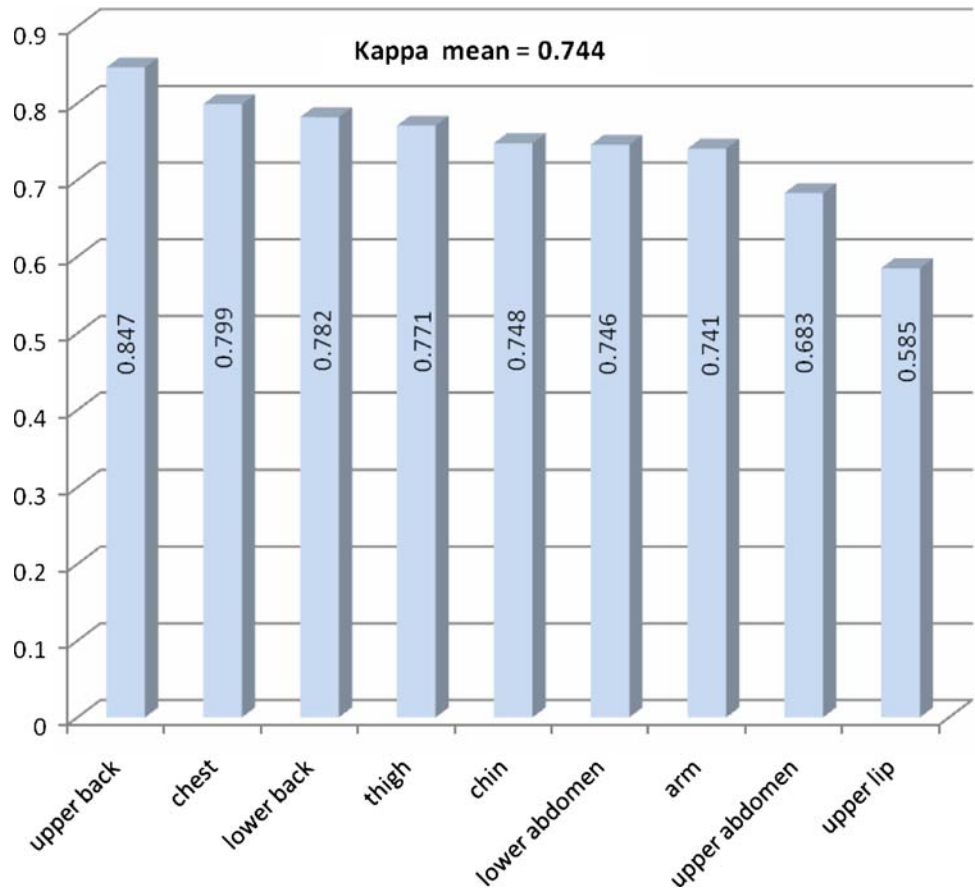
Kappa mean = 0.744

0.847 0.799 0.782 0.771 0.748 0.746 0.741 0.683 0.585

upper back chest lower back thigh chin lower abdomen arm upper abdomen upper lip

Histogram

Mean = 6,814
Std. Dev. = 5,46533
N = 242

Fig. 3 Histogram of 242 total scores

+1,96SD

Mean

-1,96SD

Fig. 4 The Bland and Altman graph displays a scatter diagram of the differences plotted against the averages of the two total Ferriman–Gallwey score measurements of two different operators. Horizontal line is drawn at the mean difference (−0.2066), and at the at the mean ± 1.96 times the standard deviation (SD = 0.618) of the differences (*dotted lines*)

In the majority of patients, hirsutism should be considered as a sign of other conditions [e.g., the polycystic ovary syndrome (PCOS), androgen-secreting tumors, nonclassic adrenal hyperplasia (NCAH), or syndromes of severe insulin resistance], rather than an isolated disorder.
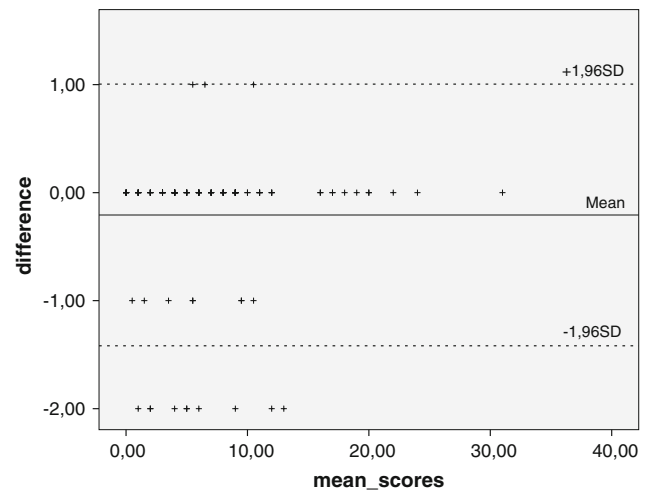
There appears to be different cut-off levels for Ferriman–Gallwey scores in different settings. Tellez and Frenkel have found that 95% of women had a score equal or less than 5 on 236 premenopausal women consulting in a birth

control clinic or consulting for acute non-endocrinological diseases. Their sample of women, coming from middle and low socioeconomic levels, appeared more hairless than European or North American Women. Thus, they depicted that hirsutism must be suspected with scores over 5 and suggested that their results cannot be extrapolated to all women, due to differences in ethnical backgrounds [19].

In a study by Hatch et al. [20] where they used the mFG scores, 7.6, 4.6, and 1.9% of their study population demonstrated scores of ≥ 6, 8, or 10, respectively. The overall cut-off values used to define hirsutism will decrease as the number of areas assessed (or the maximum score assigned to each area is reduced). For instance, Lorenzo studied 300 unselected female medical patients using a modification of the Ferriman–Gallwey score, in which only five areas of the body were scored (chin, upper lip, chest, abdomen, and thighs) [14]. Using this scoring method, they did not observe a hirsutism score over 5 in any of these women. While the exact numerical cut-off score used to define hirsutism will vary according to the quantifying system used, a value of 7 or greater is evident in only 5% of the general population when a scoring system assessing nine body areas is used [21].

The main objective of this investigation was to assess the performance characteristics and interobserver agreement of the mFG. If in this context, a physician's scoring agrees favorably with that by the other physician/researcher, then this would free up resources and facilitate group comparisons related to the treatment of hirsutism and the identification of PCOS since one of Rotterdam consensus criteria is clinical signs of hyperandrogenism. In contrast, if the level of agreement is found to be unacceptable, then the validity of studies that use only this instrument to score hirsutism should be further questioned.

Alternatively, various investigators have noted that, in comparison to white patients, hirsutism in Asian women is relatively uncommon even in the face of similar metabolic and endocrine abnormalities [22, 23].

In some earlier studies, the FG scoring has been described both as the instrument of the choice and as subjective and not useful. One of these studies reported that although the FG scoring showed the androgen excess, there was no interobserver agreement [24]. Nevertheless, in that study, none of the participants were trained by principal a investigator.

The Bland and Altman plot makes the point that any two methods that are designed to measure the same parameter (or property) will have a good correlation when a set of samples are chosen such that the property to be determined vary a lot between them. Therefore, we used this method for the assessment of observer variability in mFG. A high correlation for any two methods designed to measure the same property is thus in itself just a sign that one has cho-

sen a widespread sample. A high correlation does not automatically imply that there is good agreement between the two methods. The Bland and Altman is useful to reveal a relationship between the differences and the averages, to look for any systematic bias and to identify possible outliers. If there is a consistent bias, it can be adjusted for by subtracting the mean difference from the new method. If the differences within mean ± 1.96 SD are not clinically important, the two methods may be used interchangeably. In our study, most the observers' mean differences remained within mean ± 1.96 SD which implies acceptable interobserver variability for mFG scoring system.

According to the kappa values, in general, the scores of all nine areas were concordant between the observers. In this study, the upper lip showed the highest interobserver variability and it seems to have the highest androgen sensitivity between all the body areas studied.

We did not measure the serum androgen levels of the study population. As we declared in the methods section, observers were masked for the subjects' androgen levels to avoid ascertainment bias. It is true that our small sample does not represent the whole population, yet we highlighted the fact that patients without the complaint of hirsutism might have high mFG scores. These women can unnecessarily fulfill the 2003 Rotterdam concensus criteria for PCOS or become a candidate of a hirsutism treatment. The opposite can also happen unintentionally. Our suggestion and recommendation to an investigator in the field of endocrinology is they should be aware of the appropriate cut-off points according to the population characteristics. The definition of hirsutism may depend on self-perception of an individual women, relative comparison of an individual herself among the society, the degree of the body hair intensity as a pathology to be accepted or not accepted by the women and finally the priority of the hirsutism to become a health problem among the other health problems of women. So there is no standard definition of the hirsutism to be a complaint of an individual.

In a report from China, the suitable criterion of hirsutism for Chinese women in Shandong region was suggested to be ≥ 2 scores [25]. Because of the genetic variation of the different populations, the hair intensity and distribution shows a wide interracial spectrum. This variation forms the main objective of our research, and that was what we had tried to prove that one cut-off does not fit to everyone. In our opinion, at the population basis, the definition of hirsutism has to be worrisome for an individual and the general acceptance for most of the pathologies is that any value beyond the upper 95th percentile is said to be abnormal. In the same report, the hirsutism was significantly higher in PCOS patients (48.1%) than in controls (4.8%) by FG score ≥ 2. It is obvious that only 4.8% of the normal Chinese population have FG score more than 2 [25].

The perception of women to have a complaint about hirsutism may vary and seems to depends on not only the degree of body hair distribution and intensity but also the sensitivity of women to their body hair pattern which is accepted to be normal or abnormal. In a recent report of DeUgarte CM et al., a mFG score of at least 3 was observed in 22.1% of all subjects (i.e., the upper quartile); of these subjects, 69.3% complained of being hirsute, compared with 15.8% of women with an mFG score below this value, and similar to the proportion of women with an mFG score of at least 8 who considered themselves to be hirsute (70.0%). They concluded that white women and that an mFG of at least 3 signals the population of women whose hair growth falls out of the norm [26]. This research basically revealed that definition of hirsutism mostly depends on the complaints of women rather than the total mFG score of them. Because the hirsutism complaint was the exclusion criteria in our study population, subjects who considered themselves to be hirsute were not enroled.

Upper 95th percentile of our study population was found to be 10.71. Apparently, it is higher than the accepted and the mFG cut-off value of 8. Although the aim of this study was to reveal the agreement and performance characteristics of observers,we identified an interesting feature of our sample of Turkish women. Consequently it would not be wrong to speculate that a higher mFG cut-off value may be more appropriate to be used for the diagnosis of hirsutism in the Turkish population. We know that our results cannot be extrapolated to the whole Turkish women, due to regional differences. Additionally since our sample of women studied was not a random sample of unselected women from the community, it would not be appropriate to suggest that it would be representative of the general population. However, our study was conducted primarily with the aim of identifying the interobserver variability of mFG scoring and the cut-off value for the diagnosis of hirsutism in the Turkish population was a secondary outcome measure. Nevertheless, there seems to be a need for new trials in order to assess the cut-off value for the diagnosis of hirsutism in the Turkish population. One of the other limitations of our study was that, we had included a sample of women attending to our clinic with any other gynecological symptom other that hirsutism. It would have been more ideal if we had included a population of non-hirsute women without androgen excess. Furthermore its well-known that many hirsute women do not complain of it. Therefore, its possible that we might have included some hirsute women (FG score >8) without any complaint or with some androgen excess in our study. Although we have tried to minimize this bias by including women without any endocrinologic disease that might have caused high androgen levels, we believe that this bias might have had the potential to have led to higher scores in our study population, if there was any. However, the identification of a new cut-off value for the diagnosis of hirsutism was not our primary concern. The cut-off level for mFG scores are reported to be at a range of 2–8. We think that there is a need for further trials in order to determine the Turkish population norms. Our study points out the fact that Turkish women might have higher FG score cut-off to diagnose hirsutism. According to the results of our study, the mFG score has an acceptable interobserver variability and the cut-off value to establish the diagnosis of hirsutism should be population-specific.

Conflict of interest statement None.

References

1. Ferriman D, Gallwey JD (1961) Clinical assessment of body hair growth in women. J Clin Endocrinol 21:1440–1447
2. Kelestimur F, Everest H, Unluhizarci K, Bayram F, Sahin Y (2004) A comparison between spironolactone and spironolactone plus finasteride in the treatment of hirsutism. Eur J Endocrinol 150:351–354. doi:10.1530/eje.0.1500351
3. Sert M, Tetiker T, Kirim S (2003) Comparison of the efficiency of anti-androgenic regimens consisting of spironolactone, Diane 35, and cyproterone acetate in hirsutism. Acta Med Okayama 57:73–76
4. Muderris II, Bayram F, Guven M (2000) A prospective, randomized trial comparing flutamide (250 mg/d) and finasteride (5 mg/d) in the treatment of hirsutism. Fertil Steril 73:984–987. doi:10.1016/S0015-0282(00)00470-2
5. Harborne L, Fleming R, Lyall H, Sattar N, Norman J (2003) Metformin or antiandrogen in the treatment of hirsutism in polycystic ovary syndrome. J Clin Endocrinol Metab 88:4116–4123. doi:10.1210/jc.2003-030424
6. Sahin Y, Dilber S, Kelestimur F (2001) Comparison of Diane 35 and Diane 35 plus finasteride in the treatment of hirsutism. Fertil Steril 75:496–500. doi:10.1016/S0015-0282(00)01764-7
7. Bayram F, Muderris II, Guven M, Kelestimur F (2002) Comparison of high-dose finasteride (5 mg/day) versus low-dose finasteride (2.5 mg/day) in the treatment of hirsutism. Eur J Endocrinol 147:467–471. doi:10.1530/eje.0.1470467
8. McKnight E (1964) The prevalence of hirsutism in young women. Lancet 1:410–413. doi:10.1016/S0140-6736(64)92789-8
9. Hartz AJ, Barboriak PN, Wong A, Katayama KP, Rimm AA (1979) The association of obesity with infertility and related menstrual abnormalities in women. Int J Obesity Int J Obes 3:57–73
10. Barth JH, Catalan J, Cherry CA, Day A (1993) Psychological morbidity in women referred for treatment of hirsutism. J Psychosom Res 37:615–619. doi:10.1016/0022-3999(93)90056-L
11. Sonino N, Fava GA, Mani E, Belluardo P, Boscaro M (1993) Quality of life of hirsute women. Postgrad Med 69:186–189
12. Ferriman D, Gallwey JD (1961) Clinical assessment of body hair growth in women. J Clin Endocrinol Metab 21:1440–1447
13. Lorenzo EM (1970) Familial study of hirsutism. J Clin Endocrinol 31:556–564
14. Hatch R, Rosenfield RL, Kim MH, Tredway D (1981) Hirsutism: implications, etiology, and management. Am J Obstet Gynecol 140:815–830
15. Redmond GP (1995) Clinical evaluation of the woman with an androgenic disorder. In: Redmond GP (ed) Androgenic disorders. Raven Press, New York, pp 1–20

16. Hines G, Moran C, Huerta R, Folgman K, Azziz R (2001) Facial and abdominal hair growth in hirsutism: a computerized evaluation. J Am Acad Dermatol 45(6):846–850. doi:10.1067/mjd.2001.117386

17. Derksen J, Moolenaar AJ, Van Seters AP, Kock DF (1993) Semi-quantitative assessment of hirsutism in Dutch women. Br J Dermatol 128(3):259–263. doi:10.1111/j.1365-2133.1993.tb00168.x

18. Knochenhauer ES, Hines G, Conway-Myers BA, Azziz R (2000) Examination of the chin or lower abdomen only for the prediction of hirsutism. Fertil Steril 74(5):980–983. doi:10.1016/S0015-0282(00)01602-2

19. Tellez R, Frenkel J (1995) Clinical evaluation of body hair in healthy women. Rev Med Chil 123(11):1349–1354

20. Hatch R, Rosenfield RL, Kim MH, Tredway D (1981) Hirsutism: implications, etiology, and management. Am J Obstet Gynecol 140:815–830

21. Knochenhauer ES, Key TJ, Kahsar-Miller M, Waggoner W, Boots LR, Azziz R (1998) Prevalence of the polycystic ovarian syndrome in unselected black and white women of the Southeastern United States: a prospective study. J Clin Endocrinol Metab 83:3078–3082. doi:10.1210/jc.83.9.3078

22. Aono T, Miyazaki M, Miyake A, Kinugasa T, Kurachi K, Matsumoto K (1977) Responses of serum gonadotrophins to LH-releasing hormone and oestrogens in Japanese women with polycystic ovaries. Acta Endocrinol (Copenh) 85:840–849

23. Ewing JA, Rouse BA (1978) Hirsutism, race and testosterone levels: comparison of east Asians and Euroamericans. Hum Biol 50:209–215

24. Wild RA, Veseley S, Bebee L, Whitsett T, Owen W (2005) Ferriman Gallwey self scoring I: performance assessment in women with polycystic ovary syndrome. J Clin Endoc Methods 90(7):4112–4114. doi:10.1210/jc.2004-2243

25. Zhao JL, Chen ZJ, Shi YH, Geng L, Ma ZX, Li Y et al (2007) Investigation of body hair assessment of Chinese women in Shandong region and its preliminary application in polycystic ovary syndrome patients. Zhonghua Fu Chan Ke Za Zhi 42(9):590–594

26. DeUgarte CM, Woods KS, Bartolucci AA, Azziz R (2006) Degree of facial and body terminal hair growth in unselected black and white women: toward a populational definition of hirsutism. J Clin Endocrinol Metab 91(4):1345–1350. doi:10.1210/jc.2004-2301