Published in final edited form as:

Fertil Steril. 2011 November ; 96(5): 1266–1270.e1. doi:10.1016/j.fertnstert.2011.08.022.

REANALYZING THE MODIFIED FERRIMAN-GALLWEY SCORE: IS THERE A SIMPLER METHOD FOR ASSESSING THE EXTENT OF HIRSUTISM?

Heather Cook, M.D.^a, Kathleen Brennan, M.D.^{a,b}, and Ricardo Azziz, M.D., M.P.H., M.B.A.^c

^aDepartment of Obstetrics and Gynecology, the David Geffen School of Medicine at UCLA, Los Angeles, California

^bDepartment of Obstetrics and Gynecology, Cedars-Sinai Medical center, Los Angeles California

^cGeorgia Health Sciences University, Augusta, Georgia

Abstract

Objective—To determine whether assessing the extent of terminal hair growth in a subset of the traditional 9 areas included in the modified Ferriman-Gallwey (mFG) score can serve as a simpler predictor of total body hirsutism when compared to the full scoring system, and to determine if this new model can accurately distinguish hirsute from non-hirsute women.

Design—Cross-sectional analysis

Setting—Two tertiary care academic referral centers.

Patients—1951 patients presenting for symptoms of androgen excess.

Interventions—History and physical examination, including mFG score.

Main Outcome Measures—Total body hirsutism.

Results—A regression model using all nine body areas indicated that the combination of upper abdomen, lower abdomen and chin was the best predictor of the total full mFG score. Using this subset of three body areas is accurate in distinguishing true hirsute from non-hirsute women when defining true hirsutism as mFG>7.

Conclusion—Scoring terminal hair growth only on the chin and abdomen can serve as a simple, yet reliable predictor of total body hirsutism when compared to full body scoring using the traditional mFG system.

Keywords

hirsutism; Ferriman-Gallwey; modified Ferriman-Gallwey; androgen excess; polycystic ovary syndrome; PCOS; hair growth

© 2011 American Society for Reproductive Medicine. Published by Elsevier Inc. All rights reserved.

Corresponding Author/Reprint Requests: Ricardo Azziz, MD, MPH, MBA, Georgia Health Sciences University, 1120 15th Street, AA-311, Augusta Georgia 30912. Phone: (706) 721-2301, Fax: (706) 721-2303. razziz@mcg.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

INTRODUCTION

Hirsutism is defined as the development of male pattern terminal hair growth in women and affects approximately 5-8% of the population (1-5). It is often associated with androgen excess disorders including non-classic adrenal hyperplasia, androgen secreting tumors, and polycystic ovary syndrome (PCOS) (1,6). For purposes of clinical evaluation and diagnosis, it is important to have a standardized system to evaluate hirsutism (4). The most widely used scoring system was first developed by Ferriman and Gallwey in 1961, and involves subjective tabulation of terminal hair growth in various areas of the body (7). The original Ferriman-Gallwey system involved the scoring of 11 body areas including the lip, chin, chest, upper abdomen, lower abdomen, upper arm, forearm, thigh, lower leg, upper back, and lower back. This was later modified (i.e. the modified Ferriman-Gallwey or mF-G method) to include only nine body areas, excluding the forearm and lower leg as these areas were found not to correlate with androgen excess (2,4,7,8).

In the mFG scoring system, each individual body area is visually scored on a scale of zero to four, where zero indicates no terminal hair growth and four indicates full male pattern terminal hair growth (4). However, this system does have its limitations. First, it involves a detailed, full body exam assessing and scoring all nine body areas (2), which may be considered invasive by many patients and is cumbersome when performing epidemiologic studies. Additionally, studies have shown wide inter-observer variations when using the mFG score (8-9). For these reasons, it would be prudent to identify a simpler and more widely applicable screening system for hirsutism.

Previous studies, the majority of which are observational, have suggested that specific body areas may be more related to total body hirsutism than others, including the lip, chin and lower abdomen (2-3, 9-10). Given these data, we hypothesized that scoring terminal hair growth on a subset of the traditional nine body areas used in the mFG scoring system could serve as a simpler, less invasive, yet reliable predictor of total body hirsutism.

MATERIALS AND METHODS

Subjects

One thousand one hundred sixteen patients presenting to the University of Alabama Birmingham (UAB) for evaluation of symptoms potentially related to androgen excess between 1987 and 2002 served as the *discovery cohort*. The *validation cohort* consisted of 835 patients presenting to Cedars-Sinai Medical Center for similar complaints between 2003 and 2009.

Patients who were premenarchal or postmenopausal, had undergone prior hysterectomy, bilateral oophorectomy, had been previously diagnosed with PCOS or another disorder of androgen excess, or were receiving hormonal treatment for at least three months prior to their evaluation were excluded from the study.

Study protocol

A complete physical examination was performed on each patient, including a hirsutism exam using the mFG scoring system (4). Total mFG scores, as well as individual body area scores, were prospectively recorded. Each body area was visually scored on a scale of zero to four; a score of zero indicated no terminal hair growth, while a score of four indicated full male pattern terminal hair growth. Additionally, height, weight, age, and race/ethnicity were noted, and body mass index (BMI) was calculated for each patient. All exams were performed by one examiner (RA), and all data were recorded prospectively. The study was approved by the Institutional Review Boards for Human Protection of UAB and CSMC.

Statistical Analysis

A multiple linear regression model was used to evaluate all 9 body areas and their combinations to find which subset could serve as the best predictor of the *full mFG score*. The summation of individual body area scores from this subset was termed the *simplified (sFG) FG score*. The regression model was applied to the discovery cohort in order to generate an equation that would estimate the *full mFG score* based only on this subset of body areas. The score calculated by this equation was termed the *estimated mFG score*. A predetermined definition of hirsutism was not selected, as opinions differ regarding the appropriate cut-off point to distinguish hirsute from non-hirsute women using the *full mFG score* (11-14). Therefore, analysis was performed for *full mFG scores* of 6, 7, and 8. A non-parametric receiver operating characteristic analysis was then carried out on the discovery cohort to determine the threshold *estimated mFG score* that would most accurately distinguish hirsute from non-hirsute patients, which was determined by maximizing both the sensitivity and specificity.

In addition to evaluating the *estimated mFG score* that would separate hirsute from non-hirsute women, a separate receiver operating characteristic analysis was performed to evaluate the threshold *sFG scores* that would distinguish hirsute from non-hirsute women. Sensitivities, specificities, positive predictive values and negative predictive values were then calculated using data from the validation cohort for both the *estimated mFG score* and the *sFG score* for different population prevalences.

RESULTS

The discovery cohort did not differ significantly from the validation cohort with regards to age (28 v 31.9 years) and BMI (31.7 v 29.5 kg/m²), although ethnic/racial differences were found. Compared to the discovery cohort, the validation cohort had fewer white patients (67.2 v 82.6%, $p<0.001$) and more patients in the 'other' ethnicity category (18.8% v 2.4%, $P<0.001$) (most of whom were Hispanic-American); the proportion of black patients was similar in both groups (15% v 14%).

Data from our discovery cohort indicated that the best single body area predictor of total body hirsutism was the lower abdomen. Using only the lower abdomen as a predictor, 55% of the variability in the data set was explained ($p<0.001$). Table 1 demonstrates the variance accounted for by each individual body area. When all combinations of 3 body areas were examined, two models were shown to be the best predictors of full mFG score and statistically equivalent. These models were: a) *upper abdomen + lower abdomen + chin* ($R^2=78.7\%$, $p<0.001$), and b) *upper abdomen + lower abdomen + thigh* ($R^2=79.6\%$, $p<0.001$). Increasing the subset to four body areas, including lower abdomen, upper abdomen, chin, and thigh, did not significantly increase the predictive ability ($R^2=79.6\%$, $p<0.001$). Because our objective was to find a simpler and less invasive method for assessing the extent of terminal body and facial hair growth, the three body area model without the thigh was chosen for the *estimated mFG score*, as examination of the chin is less invasive than examination of the thigh. Subsequently, using our regression model, the following *estimated mFG score* equation was created in order to predict the *full mFG score* within ± 2.5 points:

$$\text{Estimated mFG score} = (\text{lower abdomen} + \text{upper abdomen} + \text{chin}) \times 1.8$$

The predictive value of the *estimated mFG score* was then tested on our validation cohort of 835 patients and was found to still have a high correlation with the *full mFG score* ($R^2=80\%$ and $p<0.001$). Additionally, we examined the interaction between our estimated score and race. Not surprisingly, the means of both *full mFG scores* and *estimated mFG scores* did vary across race. However, the relationship between the *full mFG scores* and the

corresponding *estimated mFG scores* was the same in all three races. Therefore, the equation for the *estimated mFG score* did not need to be altered according to race.

We then determined the best *estimated mFG score* threshold to distinguish true hirsute from true non-hirsute women. As there is no standard *full mFG score* differentiating hirsute from non-hirsute women (4), three different mFG cut-off points (≥ 6 , ≥ 7 , and ≥ 8) that have been used to define hirsutism were explored (11). For each cut-off value, a non-parametric receiver operating characteristic analysis was carried out on the discovery cohort to determine the threshold *estimated mFG score* that would most accurately distinguish hirsute from non-hirsute patients, which was determined by the receiver operating characteristic curve that maximized both the sensitivity and specificity (Figure 1). The sensitivity, specificity, and accuracy of our *estimated mFG score* equation were calculated using the validation data (Table 2). The corresponding *estimated mFG score* cutoffs are 5.4, 5.4 and 7.2 respectively for the *full mFG score* thresholds of 6, 7 and 8. For the *full mFG score* cutoff of 7 and the *estimated mFG score* cutoff of 5.4, the validation data accuracy for distinguishing hirsute and non-hirsute women is maximized at 86.1%, with a sensitivity of 74.9% and a specificity of 97.2% ($p < 0.001$).

To further simplify the hirsutism score for clinical use, we wanted to eliminate the multiplier of 1.8 in the *estimated mFG score* equation. Instead, we examined the value of using a threshold based simply on adding the absolute scores from three body areas to distinguish hirsute from non-hirsute women. Again, this is defined as the *simplified FG (sFG) score*. Not surprisingly, the best *sFG score* cutoffs in the discovery set corresponding to the *full mFG score* thresholds of 6, 7 and 8 are the *estimated mFG score* cut-offs divided by our multiplier of 1.8, or cutoffs of 3, 3 and 4 respectively.

For example, using the *full mFG score* threshold of ≥ 7 and a *sFG score* cutoff of ≥ 3 , the sensitivity, specificity and accuracy are the same as an *estimated mFG score* of 5.4 (Table 3). Using the *sFG score* threshold of ≥ 3 to distinguish hirsute from non-hirsute women, the positive predictive value (PPV) and negative predictive values (NPV) were calculated for different population prevalences. The prevalence of total body hirsutism in our validation cohort was 31.5% based on a *full mFG score* ≥ 7 . Using this population prevalence, the PPV of our model was found to be 92.5% and NPV 89.4% (Table 3).

DISCUSSION

Screening women for hirsutism in a clinical setting has traditionally involved a full body examination, which many patients and practitioners consider invasive (2). Previous studies have made attempts to correlate terminal hair growth at a specific body area with total body hirsutism, but these correlations have usually been secondary observations and not the primary intention of the study (2,3,9,10). Knochenhauer et al. (15) chose to look only at the chin and lower abdomen as predictors for total body hirsutism, which was based on the investigators' clinical experiences. Based on their data, both areas proved to have high sensitivity, but low PPV, as predictors for total body hirsutism. Our study examined all subset combinations of the nine body areas included in the mFG scoring system, and found in the discovery cohort that the best single body area correlating to the *full mFG score* was the lower abdomen. As Knochenhauer and colleagues pointed out, evaluating only this single body area is likely not useful in a clinical setting as the predictive value is low, with only 45% of the variability in our data explained by this body area score alone. Lorenzo (16) took a slightly different approach by attempting to use only the face (lip+chin) as an estimate of hair growth on other individual body areas. A significant correlation was seen when comparing scores from the face to individual scores of the abdomen, chest, thighs, and forearms, but these areas were not correlated to total body hirsutism scores.

Based on the results of our study, there appears to be a good alternative to the traditional mFG scoring system when screening for hirsutism, involving the evaluation of a subset of three body areas in place of the full hirsutism examination. For predictive purposes, the best subsets were found to be the combination of upper abdomen, lower abdomen and chin, or upper abdomen, lower abdomen, and thigh. We chose to primarily study the former subset because the two subsets were shown to be statistically equivalent as predictors, and evaluation of the chin was felt to be less invasive and cumbersome than examination of the thigh. A *sFG score* of ≥ 3 , considering these three body areas only, will distinguish hirsute from non-hirsute women with an accuracy of 87.5%.

It is important to consider the impact of the prevalence of hirsutism in certain populations when assessing the predictive ability of the screening method. In a population with a high prevalence of hirsutism, such as those encountered by physicians that treat patients with androgen excess, the predictive value is higher (PPV of 92.5% in our validation cohort) compared to populations with a lower prevalence of hirsutism, such as in general population screening studies, where the PPV would be expected to fall somewhere between 58-75% based on the 5-8% reported prevalence of hirsutism in the United States (4). Thus, when assessing populations of unselected women, a confirmatory full body assessment for hirsutism may be necessary in women who have a positive initial screen. Finally, we should note that this study was performed in mainly white, black and Hispanic-American women and may not be applicable to other ethnic populations, such as Asians.

In conclusion, we have demonstrated in two large cohorts that examination of only the chin and abdomen is a simple and reliable screening method for detecting hirsutism. Furthermore, reducing the number of body areas assessed may increase the willingness of study participants or patients to be evaluated. By simplifying the method of evaluation and by reducing the number of body areas evaluated, the potential for error and for inter-observer variability may be lowered. This tool would be most useful in the execution of large epidemiologic studies of androgen excess features in women. However, we are uncertain of its usefulness in a clinical setting where more complete scoring of body hair growth may be possible. Additionally, this new tool has not been tested in monitoring patients' responses to therapeutic interventions for the treatment of hirsutism. It is also important to determine if our findings remain consistent when tested in different ethnic populations. As such, further research is needed to test our simplified model of hirsutism scoring and to assess its value in the clinical setting, as well as to assess patient and clinician comfort with this method. Because hirsutism was assessed by one examiner in our study, the model also needs to be tested for inter-observer variability to determine the generalizability in various clinical settings.

Acknowledgments

Grant Support: This study was supported in part by NIH grants R01-HD29364 (to RA) and an endowment from the Helping Hands of Los Angeles, Inc.

REFERENCES

1. Azziz R, Sanchez LA, Knochenhauer ES, Moran C, Lazenby J, Stephens KC, Taylor K, Boots LR. Androgen excess in women: experience with over 1000 consecutive patients. *J Clin Endocrinol Metab.* 2004; 89:453–62. [PubMed: 14764747]
2. Knochenhauer E, Hines G, Conway-Myers B, Azziz R. Examination of the chin or lower abdomen only for the prediction of hirsutism. *Fertil Steril.* 2000; 74:980–3.
3. Lunde O, Grottump P. Body hair growth in women: Normal or hirsute. *Am J Anthropol.* 1984; 64:307–13.

4. Yildiz BO, Bolour S, Woods K, Moore A, Azziz R. Visually scoring hirsutism. *Hum Reprod Update*. 2010; 16:51–64. [PubMed: 19567450]
5. DeUgarte CM, Woods KS, Bartolucci AA, Azziz R. Degree of facial and body terminal hair growth in unselected black and white women: toward a populational definition of hirsutism. *J Clin Endocrinol Metab*. 2006; 91:1345–50. [PubMed: 16449347]
6. Landay M, Huang A, Azziz R. Degree of hyperinsulinemia, independent of androgen levels, is an important determinant of the severity of hirsutism in PCOS. *Fertil Steril*. 2009; 92:643–7. [PubMed: 18722607]
7. Ferriman D, Gallwey JD. Clinical assessment of body hair in women. *J Clin Endocrinol Meta*. 1961; 21:1440–7.
8. Wild RA, Vesely S, Beebe L, Whitsett T, Owen W. Ferriman-Gallwey self scoring: performance assessment in women with polycystic ovary syndrome. *J Clin Endocrinol Metab*. 2005; 90:4112–14. [PubMed: 15827102]
9. Api M, Badogly B, Alecra A, Api O, Gorgen H, Cetin A. Interobserver variability of modified Ferriman-Gallwey hirsutism score in a Turkish population. *Arch Gynecol Obstet*. 2009; 279:473–9. [PubMed: 18677501]
10. Hines G, Moran C, Huerta R, Folgman K, Azziz R. Facial and abdominal hair growth in hirsutism: a computerized evaluation. *J Acad Dermatol*. 2001; 45:846–50.
11. Hatch R, Rosenfield RL, Kim MH, Tredway D. Hirsutism: implications, etiology, and management. *Am J Obstet Gynecol*. 1981; 140:815–30. [PubMed: 7258262]
12. Ferriman D, Purdie AW. The aetiology of oligomenorrhoea and/or hirsuties: a study of 467 patients. *Postgrad Med J*. 1983; 59:17–20. [PubMed: 6866869]
13. Ferriman D, Purdie AW. Association of oligomenorrhoea, hirsuties, and infertility. *Br Med J*. 1965; 2:69–72. [PubMed: 14305368]
14. Diamanti-Kandarakis E, Kouli CR, Bergiele AT, Filandra FA, Tsianateli TC, Spina GG, et al. A survey of the polycystic ovary syndrome in the Greek island of Lesbos: hormonal and metabolic profile. *J Clin Endocrinol Metab*. 1999; 84:4006–11. [PubMed: 10566641]
15. Knochenhauer ES, Key TJ, Kahsar-Miller M, Waggoner W, Boots LR, Azziz R. Prevalence of the, polycystic ovary syndrome in unselected Black and White women of the Southeastern United States: A prospective study. *J Clin Endocrinol Metab*. 1998; 83:3078–82. [PubMed: 9745406]
16. Lorenzo EM. Familial study of hirsutism. *J Clin Endocrinol Metab*. 1970; 31:556–64. [PubMed: 4248490]

CAPSULE

Scoring terminal hair growth on only the chin and abdomen can serve as a simple, yet reliable predictor of total body hirsutism.

Figure 1. Non-parametric ROC curves displaying the optimal *estimated mFG* score thresholds that most accurately distinguishes hirsute from non-hirsute patients. This was determined by the ROC curve that maximized both the sensitivity and the specificity for a *full mFG score* threshold of ≥ 7 .

Figure 2. Simplified Ferriman-Gallwey hair growth scoring system. A sum score of greater than or equal to 3, using the three displayed body areas, determine hirsutism.

Table 1

Percentage of variability explained (R^2) and 95% Prediction Interval (PI) for each individual body area component in decreasing order

	Discovery (N=1116)		Validation (N=835)	
Variable	R-square	PI	R-square	PI
lower abdomen	45.0%	+/- 6.8	54.5%	+/- 6.9
upper abdomen	43.5%	+/- 6.9	56.1%	+/- 6.7
chin	39.9%	+/- 7.1	50.1%	+/- 7.2
thighs	36.4%	+/- 7.3	53.3%	+/- 7.0
chest	35.0%	+/- 7.4	50.3%	+/- 7.2
upper lip	31.4%	+/- 7.6	27.4%	+/- 8.7
lower back	28.7%	+/- 7.7	49.7%	+/- 7.2
upper back	26.6%	+/- 7.8	34.5%	+/- 8.2
arms	21.0%	+/- 8.1	25.5%	+/- 8.8
Proxy Score	78.3%	+/- 4.3	80.3%	+/- 4.5

Table 2
Threshold determination of *estimated mFG score* and *simplified (sFG) score* for differing definitions of hirsutism.

Hirsutism Definitions	Threshold Estimated Score	Threshold Simplified Score	Sensitivity	Specificity	Accuracy
mFG ^a ≥ 6	5.4	3	64.9%	98.5%	81.7%
mFG ^a ≥ 7	5.4	3	74.9%	97.2%	86.1%
mFG ^a ≥ 8	7.2	4	65.9%	97.4%	81.7%

^a mFG is modified Ferriman-Gallwey score

Table 3

Positive predictive (PPV) and negative predictive (NPV) values for differing prevalence rates of hirsutism based on a *simplified (sFG) score* ≥ 3

Expected Population Prevalence (%)	PPV for simplified score ≥ 3	NPV for simplified score ≥ 3
5	58.5%	98.7%
10	74.8%	97.2%
20	87%	93.9%
31.5 ^a	92.5%	89.4%
70	98.4%	62.4%

^aPrevalence of hirsutism using an mFG score ≥ 7 in the validation cohort