Analysis of the Interobserver Variability in Penile Length Assessment

Mohamad Habous, FEBU, MD,* Gordon Muir, FRCS(uro),† Alaa Tealab, MD,‡ Ben Williamson, MBBS, BSc, MD,§ Mohamed Elkhoul, FECSM,* Waleed Elhadek, FECSM,* Saad Mahmoud, MD,* Osama Laban, FEBU, MD,§ Saleh Binsaleh, MD,** Osama Abdelwahab, MD,†† John P. Mulhall, MD, MSc, FECSM, FACS,‡‡ and David Veale, FRPSych, MD§§

*Urology and Andrology Department, Elaj Medical Centers, Jeddah, KSA; †Urology Department, King’s College, London, UK; ‡Urology Department, Zagazig University, Zagazig, Egypt; §University Hospitals Birmingham, Birmingham, UK; ¶King Khaled Hospital, Tabouk, KSA; **Division of Urology, Department of Surgery, Faculty of Medicine, King Saud University, Riyadh, KSA; ††Urology Department, Benha University, Benha, Egypt; ‡‡Sexual and Reproductive Medicine Program, Memorial Sloan Kettering Cancer, New York, NY, USA; §§ Institute of Psychiatry, Psychology, and Neuroscience, Kings College, London, UK

DOI: 10.1111/jsm.13005

ABSTRACT

Introduction. Many men seek penile augmentation treatments—a standard tool for their counseling is to inform them of what is “normal.” Although some studies suggest good correlation between stretched and erect penile length, those that have measured stretched and erect length simultaneously have shown significant variability.

Aim. To assess the accuracy of differing penile measurements with multiple observers.

Methods. We recruited 201 adult men (mean age 49.6 years) who achieved full erection using intracavernosal injection.

Main Outcome Measures. Penile measurements were taken by one of seven andrology specialists in a private, temperature-controlled (21°C, 72°F) environment. Stretched flaccid and erect length and circumference were measured. We analyzed the accuracy of each flaccid measurement using the erect measurements as a reference, for the overall patient population and for each observer.

Results. The mean underestimate of length from stretched flaccid to erect was 2.64 cm (21.4%) and girth 2.27 cm (19.5%). Interobserver variability ranged from a mean underestimate of 16–27% (length) and 15–27% (girth).


Key Words. Methodology; Penile Size Measurement; Interobserver Errors; Underestimate

INTRODUCTION

Urologists and psychotherapists often encounter patients who complain about the length of their penis, but these patients are usually well within the typical range for penis size [1]. Most men seeking advice on penile size, or requesting penile augmentation, have normal penile size [2]. Assessing penis length is critical when reassuring men about the “norm” and also when counselling such men. Men concerned about their penile size may fall into two groups [3–6]: those with small penis anxiety (SPA) and those with more severe symptoms of body dysmorphic disorder (BDD).
who may benefit from formal psychiatric assessment and treatment (including cognitive behavior therapy and selective serotonergic reuptake inhibitor medication) [7,8]. Both these groups, and in particular those men with BDD, have particular vulnerabilities possibly making them prone to exploitation with unethical and unproven treatments.

Cosmetic phalloplasty is regarded, at best, as experimental for men with SPA without any adequate outcome measures or evidence of safety [9]. There are no simple or risk-free strategies for penile augmentation, and great potential for harm [10]. In general, penile enhancement surgery can cause, at the very most, an apparent 1- to 2-cm increase in flaccid penile length and up to a 2.5-cm augmentation of penile girth. Unwanted outcomes and complications, namely penile deformity, paradoxical penile shortening, disagreeable scarring, granuloma formation, migration of injected material, and sexual dysfunction, were reported frequently in these studies [11,12]. Being able to help these men contextualize their concerns is vital in their management. Knowing what is “normal” and being able to help a man place himself on the scale of size are of key importance [13].

The first description of standard penile length for age was used by Schonfeld and Beebe [14]. The majority of studies examining penile anthropometric data has relied on measurements in the flaccid state. Veale et al. [13] recently conducted a systematic review and constructed nomograms for flaccid and erect penis length and circumference. In this meta-analysis, the stretched flaccid length appeared to be an excellent estimate of erect penile length (a ratio of 0.98). However, only three studies [15–17] actually measured the stretched and erect penile length simultaneously [15–17]. Of these, only the original study of Wessells et al. [17] showed a good correlation between erect length and stretched length. Both Promodu et al. [15] and Sengezer et al. [16] significantly underestimated erect length by 15–29%, respectively.

Some attempts have been made to standardize this aspect of measuring the flaccid penis, including development of an engineering model to approximate the optimal tensile force to be applied [18]. While it has been accepted that self-measurement may be inaccurate both due to variable technique and also potential exaggeration, no studies have addressed the issue of third party interobserver variability.

We were concerned by the potential inaccuracy both in measuring penis length between flaccid stretched and erect, and also the potential for observer bias in measuring penile size. Our hypothesis was that there may be significant error in predicting erect size by the stretched or flaccid measurements and that there may be significant measurement variation.

Methods

Study Population

Men presenting to a group of andrology clinics between January 2011 and June 2012 were evaluated by history and physical examination including penile length measurement. All men were having intracavernosal injection (ICI) to induce erection as part of their routine care. ICI is done for most (but not all) of our erectile dysfunction (ED) patients as part of optional investigations for ED. Organic causes for ED are common, and most of our patients request to ICI since most of them have failed oral medication prior to presentation. Consecutive patients who underwent ICI and met the inclusion criteria were assessed. Excluded from analysis were men with previous penile surgery (excluding circumcision), Peyronie’s disease, congenital curvature or hypospadias, and patients who did not achieve a fully rigid erection with diagnostic ICI. Comorbidities and patient demographics were recorded. Consent was collected from all participants, and the study was approved by institutional ethics committee.

Main Outcome Measures

Penile Length Measurement

Each patient was assessed by one of seven experienced andrology specialists in a controlled environment: private, air-conditioned consulting rooms at a constant temperature (21°C, 70°F). The temperature in the examination rooms (and all the medical center) is automatically controlled with central air condition. Data were collected and recorded in centimeters. Each patient had two parameters measured in the flaccid state and the same two parameters in the erect state: (i) pubic bone to tip of glans (bone-to-tip [BTT]); and (ii) circumference (girth) at the mid-penile shaft. Alprostadil 10 mcg was the injection agent utilized. For stretched flaccid length, the penis was extended to maximum capacity at a 90-degree angle to the body with the patient in the upright standing position. For erect length measurement, a fully rigid erection was required for inclusion.
Length was measured with a rigid plastic ruler, while girth was assessed using a disposable paper tape at the base of the penis.

**Statistical Analysis**

Data were collated and subsequently analyzed using SPSS software package (SPSS Statistics 20, IBM, Armonk, NY, USA). For each patient, the measurements in the erect state were considered the “true” values, and those in the stretched flaccid state were then compared with the corresponding “true” value, with absolute and percentage error recorded for each measurer.

**Results**

**Study Population**

Two hundred one patients with a mean age of 49.6 (range 20–75; SD 13.0) years were included (BMI: mean 28.77, range 16–51, SD 5.202). The comorbidity breakdown was: hypertension 12%, dyslipidemia 24%, diabetes 13%, obesity 24%, coronary artery disease 7%, and cigarette smoking 12%.

**Examiners**

The mean age and duration in clinical practice of the andrologists were 41 and 14 years, respectively (range 34–53; 8–26) years. Six of seven practitioners were formally fellowship trained, and all were fellows of the European Society of Sexual Medicine. The mean number of patients examined by each andrologist was 29 (range 12–59; SD 14).

**Penile Length Data**

The mean dimensions are listed in Table 1. The mean flaccid dimensions were BTT 11.7 ± 1.6 cm (8–17) and circumference 9.4 ± 1.5 cm (5–14). The mean erect dimensions were: BTT 14.3 ± 2.0 cm (8–19) and circumference 11.6 ± 1.5 cm (8–16). The mean interobserver discrepancy between stretched flaccid and erect BTT length was 2.9 ± 0.6 (range 2.0–3.5). The mean interobserver discrepancy between stretched and erect circumference was 2.5 ± 0.5 (range 1.7–3.1).

**Comorbidity**

The numbers with hypertension were 12% (24 patients), dyslipidemia 24% (48 patients), diabetes 13% (26 patients), obesity 24% (48 patients), coronary artery disease 7% (14 patients), and cigarette smoking 12% (24 patients).

**Discussion**

Our study suggests that not only is stretched length inaccurate in predicting erect length, but also experienced andrologists are likely to produce significantly different measurements. Put in context, a 20% underestimate of length can easily introduce an error of 3 cm in assessing patient’s true penile length—even much more than the most enthusiastic proponents of penile augmentation would ever claim. The interobserver variation was most marked when comparing skin-to-tip length. We hypothesize that this measurement generated the widest variation due to two principal factors, the first being the problem of how much axial traction the andrologist should apply. While it is possible to standardize the force, using a technique proposed by Chen et al. [18], this complicated methodology was neglected in our study in the interests of patient comfort. This concern for patient comfort is no doubt one of the main problems associated with measuring the stretched flaccid penis.

The second problem we associate with the wide interobserver variation in measuring the skin-to-tip length is that of the somewhat variable proximal end point for measuring at the supra-pubic skin. Different andrologists are likely to employ slightly different techniques, such as whether the patient should remain in full expiration for the duration of the measurement, or indeed the posture the patient should adopt. It can also be tricky to fully stretch the penis while simultaneously compressing the infrapublic fat pad with the measuring device.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Results of penile dimensions with age and BMI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age (years)</td>
</tr>
<tr>
<td>Mean</td>
<td>49.6</td>
</tr>
<tr>
<td>SD</td>
<td>13.0</td>
</tr>
<tr>
<td>Range</td>
<td>20–75</td>
</tr>
</tbody>
</table>

Nonetheless, a significant degree of interobserver variation remained when comparing BTT stretched length measurements, despite the fact that this parameter in theory negates perhaps the most marked source of error inherent in the skin-to-tip measurement noted above (i.e., that there is a definitive proximal end point in the bony pubis, as opposed to the variable end point of the supra-pubic skin). We have previously shown that bone-to-tip measurement seems more accurate that skin to tip in the erect penis, due to correction for BMI and thus the amount of infra-pubic fat (M. Habous et al., unpublished data).

Our data supports the findings of Promodu et al. [15] and Sengezer et al. [16], with a larger sample of both participants and andrologists. We have demonstrated the potential for misdiagnosis and misinformation of patients introduced by observer error.

Limitations

Our study could be improved by using the same group of patients and having each participant measured by two or more examiners in order to calculate the intraclass correlation coefficient. Many men are however reluctant to undergo repeated physical examination of their genitalia and dislike ICC induction of erection, apart from the ethics of repeated injection. A training program might improve interobserver reliability for the measurement of stretched length, although our preference is for the use of erect measurements.

We may be criticized for assuming that measurement of the erect penis is the reference value. Ideally, a number of specialists would have measured the same erect penises on a number of occasions to determine interobserver reliability when measuring the erect penis, but due to geographical and ethical issues, this was not possible, and probably never will be. It does however seem logical to assume that a fully erect penis, not requiring stretching or support, should be the easiest to measure. Men may be concerned with both erect and flaccid size, although a minority focus on flaccid size alone [4], as do nearly all physical intervention studies so far reported [4].

The difference in penile girth between the stretched flaccid state and the erect state is likely to be due to the morphological changes that occur with engorgement, rather than operator error. Our comparison of girth in this study was more for completeness and should be considered as a separate piece of information generated, rather than a measure to reflect on interobserver variation. Our data in this respect confirm that men can be assured that flaccid girth does not correlate with erect girth. Previous studies have differed when assessing the correlation between erect and stretched penis length. This study could suggest that interobserver variability may be a factor in such discrepancies.

All our men were given ICCs as part of routine andrology care. This is also our practice in the Middle East for men concerned by penile size when counseling them. In other cultures, video sex stimulation may well be a less invasive substitute, but there remain countries where the use of erotic images, even for medical reasons, is unacceptable. The study group was composed largely of men of Middle Eastern origin, but there is no reason to expect a different result with other ethnic groups even if the mean size measurements were larger or small.

Lastly, measurement of penile size is of course only one aspect of assessment of men with SPA. Equally as important is male genital image satisfaction [19,20], an understanding of the beliefs and attitudes about penile size [3], the frequency of avoidance, and safety-seeking behaviors to prevent the risk of shame or humiliation and the need to screen for BDD [21].

Conclusions

This large multicentre, multi-observer study of penis size demonstrates that flaccid penile measurements are both unreliable and observer dependent. For the first time, we have shown a further disadvantage of flaccid state measurement—marked and clinically significant interobserver variation. We therefore recommend that when counseling men about penile augmentation and reassuring men with regard to penis size, erect measurements should be made, recorded, and compared with population-based nomograms.

Acknowledgments

D.V. acknowledges salary support from the National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health at South London and Maudsley NHS Foundation Trust and the Institute of Psychiatry, King’s College London. This study was supported by a grant from the College of Medicine Research Center, Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia.

Corresponding Author: Mohamad Habous, FEBU, MD, Urology, Elaj Medical Centers, King Fahad Street,
Conflict of Interest: The author(s) report no conflicts of interest.

Statement of Authorship

Category 1
(a) Conception and Design
Mohamad Habous; Gordon Muir
(b) Acquisition of Data
Mohamad Habous; Alaa Tealab; Osama Laban; Osama Abdelwahab; Saad Mahmoud; Waleed Elhadek; Saleh Binsaleh; Mohamed Elkhouly
(c) Analysis and Interpretation of Data
Ben Williamson; Mohamad Habous; Gordon Muir; David Veale

Category 2
(a) Drafting the Article
Gordon Muir; Mohamad Habous; John P. Mulhall
(b) Revising It for Intellectual Content
Gordon Muir; David Veale; Mohamad Habous

Category 3
(a) Final Approval of the Completed Article
Mohamad Habous; John P. Mulhall

References