Introducing the HOPE (Hypospadias Objective Penile Evaluation)-score: A validation study of an objective scoring system for evaluating cosmetic appearance in hypospadias patients

Fred van der Toorn a,*, Tom P.V.M. de Jong b, Robert P.E. de Gier c, Piet R.H. Callewaert d, Eric H.J.R. van der Horst e, Martijn G. Steffens f, Piet Hoebeke g, Rien J.M. Nijman h, Nicol C. Bush i, Katja P. Wolffensbuttel a, Marleen M.C. van den Heijkant b, Jan-Willem van Capelle f, Mark Wildhagen j, Reinier Timman k, Jan J.V. van Busschbach k

a Division Pediatric Urology, Sophia Children’s Hospital, Erasmus Medical Center, P.O. Box 2060, 3000 CB Rotterdam, The Netherlands
b Pediatric Renal Center, University Children’s Hospital UMC Utrecht and AMC Amsterdam, The Netherlands
c Division Pediatric Urology, Radboud University Medical Center Nijmegen, The Netherlands
d Division Pediatric Urology, Maastricht University Medical Center, The Netherlands
e Division Pediatric Urology, VU University Medical Center Amsterdam, The Netherlands
f Division Pediatric Urology, Isala Hospital Zwolle, The Netherlands
 g Division Pediatric Urology, University Hospital Gent, Belgium
h Division Pediatric Urology, University Medical Center Groningen, The Netherlands
i Division Pediatric Urology, Children’s Medical Center and University of Texas Southwestern Medical Center, Dallas, USA
j Unit of Trials and Research Coordination, Erasmus Medical Center Rotterdam, The Netherlands

* Corresponding author. Tel.: +31 10 70 36559; fax: +31 10 70 36802.
E-mail addresses: f.vandertoorn.1@erasmusmc.nl (F. van der Toorn).
Abstract  Objective: To determine the reliability and internal validity of the Hypospadias Objective Penile Evaluation (HOPE)-score, a newly developed scoring system assessing the cosmetic outcome in hypospadias.

Patients and methods: The HOPE scoring system incorporates all surgically-correctable items: position of meatus, shape of meatus, shape of glans, shape of penile skin and penile axis. Objectivity was established with standardized photographs, anonymously coded patients, independent assessment by a panel, standards for a “normal” penile appearance, reference pictures and assessment of the degree of abnormality. A panel of 13 pediatric urologists completed 2 questionnaires, each consisting of 45 series of photographs, at an interval of at least 1 week.

The inter-observer reliability, intra-observer reliability and internal validity were analyzed.

Results: The correlation coefficients for the HOPE-score were as follows: intra-observer reliability 0.817, inter-observer reliability 0.790, “non-parametric” internal validity 0.849 and “parametric” internal validity 0.842. These values reflect good reproducibility, sufficient agreement among observers and a valid measurement of differences and similarities in cosmetic appearance.

Conclusions: The HOPE-score is the first scoring system that fulfills the criteria of a valid measurement tool: objectivity, reliability and validity. These favorable properties support its use as an objective outcome measure of the cosmetic result after hypospadias surgery.

Introduction

Hypospadias is one of the most common congenital malformations of the genito-urinary tract with an incidence rate of around 3.8 per 1,000 male births [1,2]. The goals of modern hypospadias repair are creating a functional neo-urethra, straightening curvature and achieving a cosmetically normal appearance of the penis with minimal complications [3,4].

In more than 80% of patients the meatus is in a distal- or mid-shaft position for which improvement in cosmetic appearance is the most important indication for operative treatment [4]. Still, boys who underwent such correction tend to have a more negative genital perception and may be more inhibited in sexual contacts than boys without hypospadias, especially when cosmetic outcome is poor [5–10]. Also quality of life of children with hypospadias might be impaired by negative penile self-perception, sense of shame or teasing remarks from peers [11]. In addition, poor cosmetic outcome can also result in worse school performance [12]. Since modern hypospadias repair demonstrates an acceptably low complication rate, current hypospadias surgery should focus on (improving) cosmetic outcome [4,13–15].

To offer our patients the best care, in 2009 the “Dutch Hypospadias Study Group” initiated a web-based prospective multicenter study to evaluate the outcomes of hypospadias surgery (complications, cosmetic and functional outcomes). The group’s goals were quality assurance based on prospective data collection, periodic evaluation, and practice change if necessary, in keeping with standard aspects of professional practice [16,17]. To date, however, there is no validated and objective measurement tool for cosmetic appearance incorporating all relevant and surgically correctable aspects to be used in this study [6–8,18–23]. To fill the gap, we designed the Hypospadias Objective Penile Evaluation (HOPE)-score and evaluated its reliability and validity.

Patients and methods

The HOPE scoring system evaluates penile appearance on the basis of six surgically correctable items: the position of meatus, shape of meatus, shape of glans, shape of penile skin and penile axis including penile torsion and (if erection is observed) penile curvature. The total HOPE score ranges from a minimum score of 1 to a maximum of 10 (see Table 1).

The HOPE-score uses the following 6 elements of objectivity (see Table 2). First, standardized pictures of the penile appearance: the penis is photographed in a standard way from 5 different views, respectively dorsal, right lateral, left lateral, ventral overview and ventral detail of glans/meatus (see Fig. 1). Second, the use of anonymously coded patients. Third, independent assessment by a panel of pediatric urologists. Fourth, recognition of standards for a “normal” penile appearance by the panel (see Table 3). The standard for a “normal” meatus is a slit-like shape without scars, no irregularity/ asymmetry and no swelling, no lump(s) and located on the ventral-distal aspect of the
The standard for "normal" shape of the glans is acorn shaped with a solid bridge of normal looking glans tissue proximal to the meatus, with absence of scars, irregularity/asymmetry swelling and lump(s) [25]. The standard of a "normal" shape of penile skin is an evenly covered penis with absence of scars, irregularity/asymmetry, swelling and lump(s). The standard of the penile axis is straight without curvature or torsion [3,4].

Fifth, assessment of the degree of abnormality instead of the subjective degree of satisfaction by the rater. Sixth,
the use of reference pictures for the different degrees of abnormality in the items of the HOPE-score (see Fig. 2).

In our prospective multicenter “Dutch Hypospadias Study” standardized photographs of the penile appearance are made on all patients pre-operatively (under general anesthesia) and post-operatively at six months as well as at ages 5, 10, and 15 years. For this validation study we selected randomly and on the basis of photographically good quality the series of photographs of 70 (60 six months post-operatively and 10 pre-operatively) of the first 352 patients.

These pictures were divided over two questionnaires A and B as follows: series of 10 (randomly selected) patients were included in both questionnaires; of the remaining 60 series, 30 were included in questionnaire A, the other 30 in questionnaire B. A panel of 13 pediatric urologists first completed questionnaire A and at least 1 week later questionnaire B. Besides the HOPE-score, the panel also rated the “general cosmetic appearance” of the penis on a 10-point scale ranging from 10 (=perfect) to 1 (=one of worst appearances imaginable). In addition, each questionnaire included pictures in 5 sets of 3 (randomly

### Table 3  Standards of “normal” cosmetic appearance HOPE-score.

<table>
<thead>
<tr>
<th>Position meatus:</th>
<th>Location on the ventral-distal aspect of the glans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shape Meatus:</td>
<td>- Slit-like shape</td>
</tr>
<tr>
<td></td>
<td>- No scar(s)</td>
</tr>
<tr>
<td></td>
<td>- No irregularity/asymmetry</td>
</tr>
<tr>
<td></td>
<td>- No swelling/lump(s)</td>
</tr>
<tr>
<td>Shape Glans:</td>
<td>- Acorn shaped</td>
</tr>
<tr>
<td></td>
<td>- Solid bridge of normal looking glans tissue underneath the meatus</td>
</tr>
<tr>
<td></td>
<td>- No scar(s)</td>
</tr>
<tr>
<td></td>
<td>- No irregularity/asymmetry</td>
</tr>
<tr>
<td></td>
<td>- No swelling/lump(s)</td>
</tr>
<tr>
<td>Shape Skin:</td>
<td>- Evenly covered penis</td>
</tr>
<tr>
<td></td>
<td>- No scar(s)</td>
</tr>
<tr>
<td></td>
<td>- No irregularity/asymmetry</td>
</tr>
<tr>
<td></td>
<td>- No swelling/lump(s)</td>
</tr>
<tr>
<td>Penile axis:</td>
<td>- Straight without curvature or torsion</td>
</tr>
</tbody>
</table>
selected) patients to determine a rank order in "general cosmetic appearance" without using the HOPE-score. The panel saw these "rank order" pictures in one questionnaire again for a separate HOPE-score assessment in the other questionnaire. So, both questionnaires A and B incorporated 45 series of photographs.

To analyze the intra-observer reliability of the HOPE-score, the panel members assessed series of photographs of the 10 cases twice in a blinded design. Judging a total of 45 series of photographs in each questionnaire at an interval of at least 1 week the panel was assumed not to realize they assessed 10 cases for a second time. For intra-observer reliability we computed the mean Pearson’s correlation coefficient between the first and second assessments of these 10 cases by the 13 panel members. We pre-defined a correlation coefficient of 0.7–0.8 a sufficient correlation and >0.8 a strong correlation [26].

Inter-observer reliability was analyzed by the intraclass correlation coefficient (ICC) between the panel members’ HOPE-scores. We used the two-way random procedure and consistency agreement. We considered the individual ratings as the unit of analysis and hence report the single measure reliability. Again, a correlation coefficient of 0.7–0.8 was considered a sufficient correlation and >0.8 a strong correlation [26].

To analyze the “non-parametric” internal validity the panel in each questionnaire rank ordered cosmetic outcome in 5 sets of 3 (randomly selected) patients without using the HOPE-score. Then they evaluated the cosmetic result in each of these patients using the HOPE-score in the other questionnaire. These HOPE scores served to construct a ranking of the same triplets of patients. This “non-parametric” internal validity was assessed by averaging the Spearman correlation coefficients between the rank order by direct judgment (without the HOPE-score) and the rank order constructed with the HOPE-score in each set. A correlation coefficient of 0.7–0.8 was considered to reflect sufficient internal validity and a correlation of >0.8 to reflect a good internal validity [26].

To analyze the “parametric” internal validity the panel determined in 70 patients both the HOPE-score and (on a 10-point scale) the “general cosmetic appearance”. The “parametric” internal validity was assessed by averaging the Spearman correlation coefficients between the rank order following from the HOPE-score in these 70 patients and the rank order following from the assessment of the “general cosmetic appearance”. A Spearman correlation of 0.7–0.8 was considered to reflect sufficient internal validity and a correlation of >0.8 to reflect a good internal validity [26].

Results

The intra-observer reliability in the panel showed an average Pearson correlation coefficient of 0.817, reflecting strong intra-rater reproducibility. The inter-observer reliability in the panel demonstrated an intraclass correlation coefficient of 0.790, indicating sufficient agreement among observers. The “non-parametric” and “parametric” internal validity scores by the panel demonstrated average Spearman rank correlation coefficients of 0.849 and 0.842 respectively, indicating a valid measurement of differences and similarities in cosmetic appearance (see Table 4).

Discussion

The ideal scoring system for cosmetic result after hypospadias surgery is an objective, reproducible and validated tool measuring all relevant and surgically correctable aspects of the hypospadias.

Important elements of objectivity are: standardized photographs, anonymous patients, independent scoring by

---

**Table 4** Results reliability and validity.

<table>
<thead>
<tr>
<th></th>
<th>Correlation coefficient</th>
<th>95% CI</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-observer reliability</td>
<td>0.817 (Pearson)</td>
<td>0.759–0.876</td>
<td>0.095</td>
</tr>
<tr>
<td>Inter-observer reliability</td>
<td>0.790 (intraclass)</td>
<td>0.733–0.843</td>
<td></td>
</tr>
<tr>
<td>Internal validity—“non-parametric”</td>
<td>0.849 (Spearman rank)</td>
<td>0.822–0.876</td>
<td>0.075</td>
</tr>
<tr>
<td>Internal validity—“parametric”</td>
<td>0.842 (Spearman rank)</td>
<td>0.827–0.856</td>
<td>0.066</td>
</tr>
</tbody>
</table>
2. Shape of Meatus

Figure 2  (Continued).
3. Shape of Glans:

Figure 2  (Continued).
a panel, recognition of standards for a “normal” penile appearance and assessment of the degree of abnormality using reference pictures. Reproducibility is determined by intra- and inter-observer reliability. Parametric and non-parametric correlations are used to determine the internal validity of a measurement tool.

In 1995 Mureau et al. evaluated (parental or patient) satisfaction with (different items of) genital appearance with the use of the validated "Genital Penile Perception Score" (GPPS) [6–8]. The GPPS includes relevant items such as position of the meatus and glanular shape but also surgically non-correctable items such as penile size, penile thickness and glanular size, as well as a non-relevant item scrotum/testes [6–8,15]. Furthermore, these authors did not use the important elements of objectivity mentioned above.

Another scoring system assessing the operative results is the "Hypospadias Objective Scoring Evaluation" (HOSE), designed by Holland et al., in 2001 [18]. This HOSE system was not exclusively developed to assess cosmetic outcome as it also incorporates an assessment of the urinary stream and the presence of fistula. This scoring system does not assess relevant cosmetic items like shape of the glans and shape of the penile skin. Also, with regard to the psychometric properties of the HOSE-score, only the interobserver reliability was tested.

Baskin in 2001 used photography to analyze the cosmetic outcome of hypospadias surgery but did not assess the degree of abnormality [19]. Furthermore, the reliability and validity of the scoring method were not tested.

The latter also holds true for a scoring system described in 2005 by Ververidis et al. using standardized photographs assessed on cosmetic outcome by an independent panel [20].

In 2006 the corresponding author of this current study developed and presented the Hypospadias Penile Perception Score (HPPS), which incorporated all relevant surgically correctable items, a set of standardized photographs and a panel assessing anonymous patients [21]. Internal validity and reliability of the HPPS were found to be good. However, the HPPS still lacked reference pictures as an important element of objectivity. Furthermore, like the scoring systems used by Mureau or by Ververidis, the HPPS evaluates the degree of satisfaction with cosmetic outcome – which is an inherent subjective assessment.

In 2008 Weber et al. copied the HPPS with permission and renamed it the Pediatric Penile Perception Score (PPPS) [22]. However, the PPPS has the same drawbacks as noted above concerning the HPPS: scoring the inherent subjective degree of satisfaction and the lack of reference pictures.

In 2008 Snodgrass et al. devised a questionnaire to determine the outcome of hypospadias repair [23]. However, this method also incorporates an assessment of urinary stream and lacks important elements of objectivity like anonymous patients, standardized photographs and evaluation by a panel.

In developing and validating our HOPE-score we set three basic goals. First, it should incorporate all relevant
4b. Shape of Skin without foreskin:

Figure 2 (Continued).

and surgically correctable items of the penile appearance. Second, the items should refer to well established standards for a “normal” cosmetic appearance. Third, it should introduce objectivity through reference pictures helping to score the degrees of abnormality of the HOPE-score items, standardized photographs, anonymous patients and independent assessment by a panel of health professionals [19,20].

In the present study the HOPE-score demonstrated a good reliability and validity, which makes it a valuable outcome measure to evaluate the general cosmetic outcome after hypospadias surgery. The use of this scoring system before and after surgery can help to evaluate to what extent the hypospadias surgery improves the penile appearance. Furthermore, assessment of the separate items can identify aspect(s) of the cosmetic appearance that need critical attention. Consequently, surgeons can try to improve the cosmetic result by refining or adapting their operative techniques. Furthermore, the HOPE-score can help to compare cosmetic outcomes from different techniques and/or specify the reasons for the different degrees of outcome.

The unbiased ability to modify operative technique in order to achieve desirable cosmetic appearance is paramount to the evolution of hypospadias surgery [27]. Therefore we will use these favorable measurement properties of the HOPE-score in our national multicenter study. To date (November 2012) 940 patients have been enrolled. As soon as 1000 patients are evaluable, we will analyze which surgical techniques consistently result in the best cosmetic outcome(s). The surgical procedure(s) resulting in optimal cosmesis will then be recorded anonymously and distributed to the group members on DVD enabling all participating hypospadias surgeons to adopt techniques that objectively optimize cosmetic appearance. We anticipate that in this way our HOPE-score outcomes will improve in time.
In our national registry we also document the parental/patient opinion of the cosmetic appearance using the same HOPE-score items, as well as satisfaction with general cosmetic appearance. So, in the future, the HOPE-score by a professional panel can determine if optimal or suboptimal cosmesis is associated with patient-measured outcomes such as satisfaction with penile appearance, self-esteem, body image or inhibition in sexual contacts.

**Conclusion**

The HOPE-score incorporates all relevant surgically-correctable items in hypospadias and uses important elements of objectivity. The HOPE-score demonstrates a good reliability and validity, supporting its use as an objective measure of cosmetic appearance after hypospadias surgery.

The HOPE-score can help to identify aspects of the operative procedure that need critical attention, enabling surgeons to adopt techniques associated with optimal cosmetic outcomes.

**Funding source**

No funding for this study.

**Conflicts of interest**

No conflict of interest.

**Ethical approval**

Study was approved by Erasmus Medical Center ethics committee.

**Acknowledgments**

Pieter Dik, pediatric urologist, made the digital drawings of the HOPE reference pictures Ko Hagoort, author’s editor, edited the manuscript draft.

**References**


RE: Introducing the HOPE (Hypospadias Objective Penile Evaluation)-score: A validation study of an objective scoring system for evaluating cosmetic appearance in hypospadias patients

To the Editor

The authors report their work of validating another score to assess the cosmetic outcome after hypospadias repair. They claim that this is the first validated instrument to do so. This is obviously not correct, as in 2009, the Pediatric Penile Perception Score (PPPS) was published after validation for prepubertal boys by independent urologists [1]. In 2013, after validation of this same instrument for postpubertal men, the name was modified to Penile Perception Score (PPS) [2].

In the article on HOPE the authors conclude that their instrument could be used as an objective outcome measure after hypospadias repair. However, the instrument was validated for prepubertal genitalia only and, therefore, the HOPE-score cannot be recommended for adult genitalia or long time follow-up beyond puberty.

In contrast to other instruments, such as the score developed by Ververidis 2005 [3] or Weber 2013 [2], the HOPE-score does not include the item "general appearance", despite its clinical relevance and its good internal consistency with other items of appearance in the PPS. Instead, the HOPE score includes one more item regarding the appearance of the meatus and the rotation of the penis, which is again assessed by the orientation of the meatus. Therefore, three out of five items depend on the appearance of the meatus, resulting in a poorly balanced instrument for cosmesis of an operated penis.

Published instruments to assess the cosmesis of the penis after hypospadias repair are far from being perfect. The authors of the HOPE-score introduced reference pictures to facilitate the assessment and improve interrater reliability. This may indeed be a progress. However, unfortunately, the reference photographs are neither standardized nor of good quality. This may explain why the interrater reliability of the HOPE-score is no better than that of the PPPS. Therefore, the present study hardly allows taking advantage of the potential benefit of reference photographs.

Validating instruments is painstaking work that requires following meticulously the highest methodological and technical standards. HOPE-fully, other research projects will go through this process to further improve instruments. The use of reference pictures or drawings may indeed allow development of new, better instruments with a higher interrater reliability and stability. In the meantime, outcome studies should probably rely on the existing instruments that were validated for all age groups, such as the PPS [2].

References


© 2013 Journal of Pediatric Urology Company Published by Elsevier Ltd. All rights reserved.