Digital Pattern Recognition for the Identification and Classification of Hypospadias Using Artificial Intelligence vs Experienced Pediatric Urologist

Nicolas Fernandez, Armando J. Lorenzo, Mandy Rickard, Michael Chua, Joao L Pippi-Salle, Jaime Perez, Luis H. Braga, and Clyde Matava

OBJECTIVE To improve hypospadias classification system, we hereby, show the use of machine learning/image recognition to increase objectivity of hypospadias recognition and classification. Hypospadias anatomical variables such as meatal location, quality of urethral plate, glans size, and ventral curvature have been identified as predictors for postoperative outcomes but there is still significant subjectivity between evaluators.

MATERIALS AND METHODS A hypospadias image database with 1169 anonymized images (837 distal and 332 proximal) was used. Images were standardized (ventral aspect of the penis including the glans, shaft, and scrotum) and classified into distal or proximal and uploaded for training with TensorFlow. Data from the training were outputted to TensorBoard, to assess for the loss function.

The model was then run on a set of 29 "Test" images randomly selected. Same set of images were distributed among expert clinicians in pediatric urology. Inter- and intrarater analyses were performed using Fleiss Kappa statistical analysis using the same 29 images shown to the algorithm.

RESULTS After training with 627 images, detection accuracy was 60%. With1169 images, accuracy increased to 90%. Inter-rater analysis among expert pediatric urologists was k= 0.86 and intrarater 0.74. Image recognition model emulates the almost perfect inter-rater agreement between experts.

CONCLUSION Our model emulates expert human classification of patients with distal/proximal hypospadias. Future applicability will be on standardizing the use of these technologies and their clinical applicability. The ability of using variables different than only anatomical will feed deep learning algorithms and possibly better assessments and predictions for surgical outcomes. UROLOGY 147: 264–269, 2021. © 2020 Elsevier Inc.

Hypospadias is the most common congenital anomaly that affects the penis with a global prevalence of 3.7 of 1000 newborns.[1] It is considered a multifactorial condition that is influenced by environmental and genetic factors.[2,3,4] Recent molecular studies have identified over 20 genes associated with isolated hypospadias, supporting the theory of hypospadias being a common phenotypic expression of different genotypes.[1,5] These findings may, in part, explain the variable surgical outcomes among experienced surgeons performing the same procedures.[6]

While most of the recently described surgical techniques have become more refined and report improved outcomes, there is yet to be a single procedure that consistently produces optimal outcomes with minimal complications.[7] Hypospadias characteristics such as glans size, urethral plate quality, meatal location, and degree of ventral curvature have been identified as possible predictors of surgical outcomes and complications[8,9] and form the basis of classification systems for this condition.[10] Merriman et al, proposed the GMS (glans, meatus, shaft) classification system, that includes not only the location of the meatus but also the characteristics of the glans including size; presence and appearance of glans groove; urethral plate quality; and the severity of ventral curvature.[10]

From the Division of Urology, Seattle Children's Hospital, University of Washington, Seattle, USA; the Department of Surgery, Division of Urology, Hospital for Sick Children, University of Toronto, Canada; the Division of Urology, Hospital Universitario San Ignacio, Pontificia Universidad Javeriana, Bogota, Colombia; the Department of Urology, Fundación Santa Fe de Bogota. Bogota, Colombia; the Division of Urology, McMaster Children's Hospital, McMaster University, Hamilton, Canada; the Division of Pediatric Urology, Sidra Medical and Research Center, Doha, Qatar; and the Department of Anesthesia, Hospital for Sick Children, University of Toronto, Canada

Address correspondence to: Clyde Matava, M.D., Department of Anesthesia, Hospital for SickKids, University of Toronto, 555 University Ave, Toronto, ON M5G 1 × 8, Canada. E-mail: clyde.matava@sickkids.ca

Submitted: June 9, 2020, accepted (with revisions): September 7, 2020

https://doi.org/10.1016/j.urology.2020.09.019
0090-4295

GMS score has performed well in predicting surgical outcomes in its initial implementations and has demonstrated good inter-rater correlation.[10] However, there is still a significant amount of subjectivity inherent with individual assessment of clinical variables.[11] This variability in assessment and classification leads to challenges in comparison of outcomes between centers and surgeons despite attempts at standardization.

There has been a recent movement in healthcare disciplines toward harnessing the machine learning and artificial intelligence technology for both predicting outcomes and automating interpretation of images.[12,13] Utilization of this technology improves reliability and removes subjectivity resulting in a standardized assessment of an object or condition. Herein, we describe the development and testing of a model trained in the recognition and assessment of hypospadias and its performance when compared to experts in pediatric urology and other clinicians in various healthcare settings. Our hypothesis is that our tool will accurately recognize a certain hypospadias phenotype and assign a defect type more reliably and with less variability than clinicians.

METHODOLOGY

After REB approval, we accessed and used images of hypospadias that had been collected and stored in an institutional clinical database. Consent for the capture of these images was obtained from the parents of hypospadias patients preoperatively as part of the standard of care to be used as a preoperative and postoperative clinical reference.

With the assistance of experts in machine learning and artificial intelligence, 2 experiments were performed utilizing the free and readily available TensorFlow platform. The first experiment was performed for the purpose of development, training and testing of an image recognition model to identify and classify hypospadias. The second experiment was performed for the comparison between the trained model and clinicians with different levels of experience in the classification of hypospadias.

Model Development, Training, and Testing

The hypospadias clinical image database was developed and maintained by acquiring images of hypospadiac penises from surgical patients who consented to the capturing of images for teaching and research purposes at the time of surgical consent. All photographs used for the present study were taken with the same standard methodology and contained images that depicted the entire ventral length of the penis with meatal location and scrotum. No images of megameatus were included in this study. A total of 2000 pictures were available to be used. A panel of 3 experts (authors) reviewed the images and classified them as either distal (meatus located above the mid shaft of the penis) or proximal (meatus below the midshaft of the penis). Conflicts were solved by consensus. Once classified, images were presented to the model for imaging recognition training. For this phase of the study, we used TensorFlow and developed the image recognition model.[14] Based on previous experience of our group, Inception V2 model was used to locate anatomical features. Object detection algorithm was run through a convolutional neural network (CNN) to determine if the images contained hypospadias. For image classification the entire image was run through a CNN which classified the image as a whole as distal or proximal.

During the first phase of the study the software was trained to recognize the presence of hypospadias. Image preprocessing was consistent and standardized for all images. Labeling boxes were part of the preprocessing phase. For all images, each box contained the glans as well as the urethral meatus, excluding the scrotum. Once the set of images was labeled and converted to TFRecord, it was used for training the object detection model. For object detection, a custom application was developed in our lab to allow users to load a set of images and draw and label bounding boxes around objects of interest.

The initial set of training images contained a total of 627 images, of which 187 were proximal and 440 were distal. For the initial tests, the models were trained for a small number of iterations (6000).

A second training was done with 1169 images, of which 837 were distal and 332 proximal. The number of iterations during training was also increased to 20707.

Once the model was trained, a set of random images from our database that were also classified by experts, not previously presented to the model were used to test the accuracy of the model.

Clinicians Classification and Testing

We developed an electronic survey using RedCap, which was distributed to clinicians with different levels of training in pediatric urology, urology, genetics, endocrinology, and general pediatrics. Respondents were asked to provide demographic information, including years of experience and level of training and indicate if they performed hypospadias surgeries and their volume of annual cases. We also gathered information about types of classification systems used by respondents when evaluating a patient with hypospadias. We then had survey respondents review and classify the same set of 31 hypospadias images used to test the model as distal or proximal. For intrarater analysis, we included different images of the same patient separately (ie, preoperative and intraoperative). Follow workflow in Figure 1.

Statistical Analysis and Sample size

Upon completion of the training phase, an arbitrary selection of 31 images was shown to experts (humans) and to the model. Inter- and intrarater analyses were carried out as well as model accuracy for detecting images. We compared results between the model and humans. Due to the number of raters and various clinical backgrounds, inter-rater agreement was calculated using Fleiss' Kappa by selecting a random sample of raters from each group. Sample size was estimated for a prespecified power of 90% while alpha was set at less than 0.05.[15] We compared kappa scores between pediatric urology clinicians to all other backgrounds, levels of training, those that operated vs those that did not and higher vs lower volumes of annual cases. Survey responses were tabulated and analyzed, and agreement kappa scores calculated using SPSS version 25. Responses were clustered according to the groups (Table 2). We used the majority answer from each group as the unified response for that clustered group. Accuracy was calculated by adding sensitivity plus specificity.

RESULTS

Initial Training and Testing

For the initial test, the model presented a desirable loss function for image classification (which represents a measure of error)

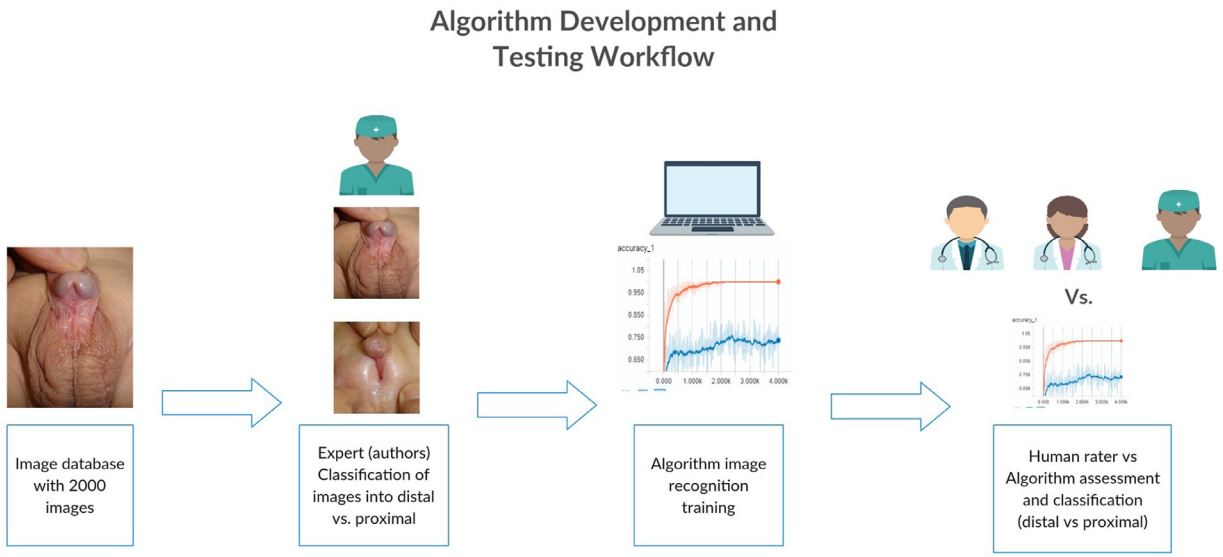

Figure 1. Algorithm creation and testing workflow. (Color version available online.)

that improved per iteration (Fig. 2A). When evaluating accuracy, we found the model had 75% accuracy at classifying hypospadias correctly.

Second Training and Testing

After increasing the number of images and iterations, the model improved the accuracy to 90% when tested. (Fig. 2B). This training phase confirmed that the overfitting was less than the initial training phase. As shown in Figure 3, we present an example of the output from the model showing a bounding box and confidence percentage. Due to the overfitting of the model beyond 20800 iterations, we confirmed that image classification algorithm was effective at memorizing images but could not apply what it had learned about new cases.

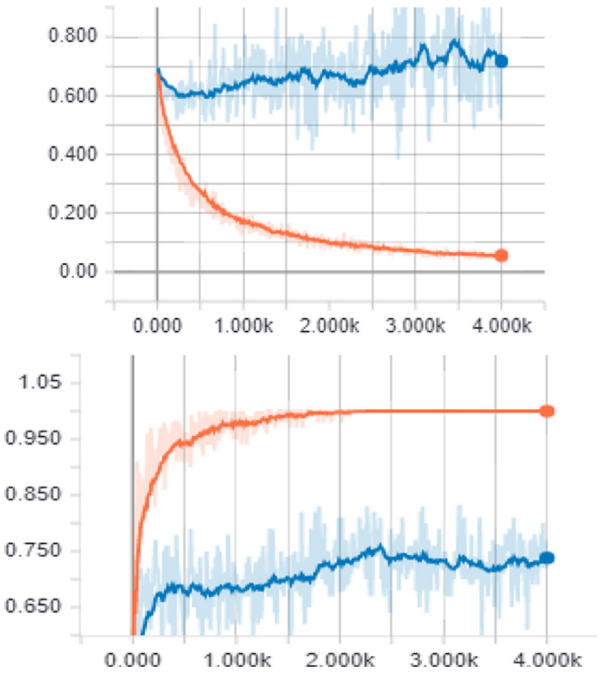

Figure 2. (A) Representation of values for each statistical measurement using first set of images (orange line) and validation process (blue line). (B) Accuracy reached after algorithm training. (Color version available online.)

Clinicians Demographics and Classification of Hypospadias

We received a total of 85 respondents to the survey and excluded 5 due to incomplete submissions for a total of 80 respondents included in the present study. Of those, 53 (69%) practiced pediatric urology followed by 12 (15%) who practiced general urology. Other represented specialties were general pediatrics 8 (9.3%), pediatric endocrinology 4 (4%) and genetics 3 (2.6%). Of those who practiced pediatric urology, 29 (53%) were staff physicians, 17 (27%) were nurse practitioners/nurses and 7(10%) were fellows. Years of experience of respondents and surgical volume of cases are presented in Table 1. Only 13%

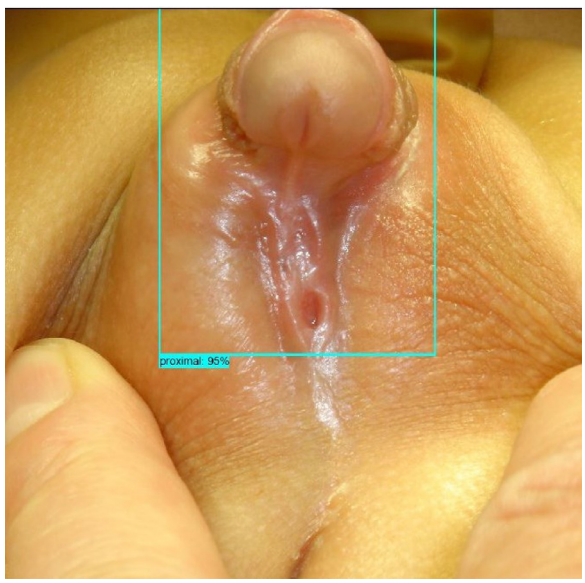

Figure 3. Output from the model showing a bounding box and confidence percentage. (Color version available online.)

Table 1. Rater demographics

Rater Demographics			Total
Pediatric urology	Staff	29	53
	Nurse practitioner	17	
	Fellow	7	
General urology			12
Pediatrics			8
Pediatric endocrinology			4
Genetics			3
Pediatric urology experience			
More than 10 years			15
Between 5 and 10 years			18
Between 2 and 5 years			13
For 1 year or (current pediatric urology fellow)			7
Number of hypospadias cases/year performed			
I do not operate			17
Between 20 to 50			19
Between 50 to 100			9
Less than 20			16
More than 100			4

of respondents reported using the GMS score as a mean to classify hypospadias with the remainder reporting a range of other systems including simple visual assessment "eye balling" and the HOSE classifying system.[16]

Overall inter-rater agreement was moderate (k = 0.64; CI 0.62-0.65), and very good for those specializing in pediatric urology (k = 0.84; CI 0.81-0.86) ($P < .05$). We noted no significant differences when comparing the group of clinicians specialized in pediatric urology, however there was a difference when we compared both general urology and pediatric urology to other specialties and general pediatricians. Kappa scores with associated confidence intervals can be found summarized in Table 2. Inter-rater agreement between the 5 pediatric urology experts was very good (k = 0.86; CI 0.75-0.97). Overall intrarater agreement for the 6 images from the same patient was poor (k = 0.28; CI 0.07-0.49), and good for the 5 experts in pediatric urology who rated the 6 images from the same patient (k = 0.74; CI 0.49-0.99).

Comparison between pediatric urology staff physicians and other pediatric urology professionals (residents, nurses or fellows), showed a similar performance with 0.80 (CI .75-.83) vs 0.81 (CI .77-.85) inter-rater agreements respectively ($P > .05$). A similar finding was identified when comparing raters with surgical and none surgical experience in hypospadias (0.81 [CI .78-.84] vs 0.83 [CI .78-.89]) respectively ($P > .05$).

A significant difference was identified when comparing pediatric urologist to none pediatric urology trained physicians who had 0.83 (CI .81-.85) vs 0.43 (CI .39-.48) inter-rater agreements respectively ($P < 0.05$). Other comparisons made are shown in Table 2.

Table 2. Accuracy and inter-rater agreement by groups

Group	k	CI	Accuracy
Staff physicians	0.80	0.75-0.83	96.9%
Other designations	0.81	0.77-0.85	96.9%
Surgeons	0.81	0.78-0.84	96.9%
Does not operate	0.83	0.78-0.89	96.9%
<50 cases per y	0.80	0.76-0.85	93.6%
>50 cases per y	0.84	0.78-0.90	96.9%
<5 y experience	0.85	0.80-.0.89	96.7%
>5 y experience	0.80	0.76-0.82	93.6%
Pediatric urology	0.83	0.81-0.85	100.0%
All other specialties	0.43	0.39-0.48	93.3%
Urology	0.81	0.78-0.82	96.7%
All other specialties	0.50	0.43-0.58	91%
Machine learning algorithm			90%

DISCUSSION

Hypospadias literature on surgical management and their outcomes has historically based all the results on purely anatomical variables as predictors. Identified predictive variables are urethral plate quality, meatus location, and ventral curvature.[17,18] Variability on surgeon`s perception makes this assessment very subjective and difficult for standardization.[8,18,19] In the case of more objective variables such as penile dimensions and angle measurements there is also a great amount of subjectivity.[6,20] The creation of an objective tool to classify hypospadias is of critical importance. Our results support the possibility of standardizing the way hypospadias are classified following experts' input. The use of novel technologies, such as machine learning algorithms and image recognition show that these artificial intelligence technologies are useful to emulate experts` experience in identifying and classifying hypospadias. If an algorithm such as the one we are presenting with the ability to capture anatomical variables as well as measurements is available, a more objective evaluation will be possible and will help improve patient care and data collection for research purposes. There is increasing interest in the pediatric urology community on a tool that standardizes the way hypospadias are classified and how literature is reported.[21]

Hypospadias surgery is a clear example of how surgeon's experience plays a direct role in surgical outcomes.[22] Our results show that there is no significant difference on classifying hypospadias based on the amount of cases treated per year. Nonetheless, expertise does play a role on intraoperative decision making. This is something that has never been objectively assessed and our results may support a potential use to evaluate this in the future. For example, a more experienced surgeon may consider a different surgical approach than a less-experienced surgeon for the same distal hypospadias patient and if we feed the algorithm with intraoperative images.[18] The algorithm's accuracy to correctly identify and classify hypospadias in our study was as high as 90% and correlates very well with expert's inter-rater kappa of 0.86. Our model supports the fact that we can reduce inter-rater variability and create a more standardize way to classify hypospadias. Real-time applicability of our algorithm can help guide surgeons perform a more standardized and reproducible procedure following experts support provided by the algorithm. Future applications of our proposed algorithm can include intraoperative imaging to create a more reproducible and objective decision-making algorithm for less experienced surgeons and guide them through-out the procedure and improve surgical outcomes.

Correlation between specialists with experience in hypospadias and nonexperts (ie, pediatricians) shows how important it is to have a tool that can be developed to guide nonurology experts in the assessments of hypospadias patients. Our results show how variable the appreciation and understanding of hypospadias is for nonurology experts. Since most cases of hypospadias are referred, it is of critical importance to improve patient detection and classification. Development of applications with this algorithm will guide clinicians on how to correctly classify the condition and standardize the management of these children and their families. Congenital anomalies surveillance systems rely most of their data collection on physical examination of newborns to detect for these anomalies.[23] Our results can be extrapolated with the potential benefit of being used at a large-scale surveillance system to improve detection and classification of congenital anomalies. It has been described that prompt detection and management of congenital anomalies reduces disability and burden to healthcare systems. For this reason, a tool that it is easily applicable and low cost can have a great impact on our communities.

For decades, surgeons have focused on anatomical variables looking to identify them as predictors of surgical outcomes. It may be possible that we have been limiting progress by looking at the wrong variables and excluding the possibility of nonanatomical ones that may play a role in surgical outcomes. The potential applicability of our algorithm may allow the possibility of adding novel variables such as histological data about tissues (Dartos, skin, etc), gene expression and interactions and molecular processes of wound healing.

The common hypospadias phenotype with a significant genetic variability has always been very difficult to correlate with surgical outcomes.[5,24] It has been described that patients with associated congenital anomalies and specific genotypes have worst outcomes that others.[25,26] Although our present results are at their earliest steps, the future of deep learning algorithms will allow the inclusion of genetic variants in predictive algorithms. For example, the possibility of genetic variants influencing surgical outcomes has not been explored in depth to date; this is the case for variants in the estrogen or androgen receptor genes and their possible interactions with wound healing processes. Clinical data about the effects of preoperative testosterone and surgical outcomes are still in debate and the use of a tool that can interpret previous anatomical variables including new genetic ones may allow for a better understanding of the postoperative natural history and surgical outcomes of patients with hypospadias.[27,28,29] The importance of identifying which are the patients that will benefit from hormonal therapy or which are the ones that will have poor healing after hormonal therapies may be based on genotyping them.

We acknowledge the limitation of only classifying hypospadias in 2 types, knowing that historically after Browne's classification system, all scientific papers have followed this anatomical concept.[30] We decided to focus on a binary classification (distal vs proximal) given the fact that surgical outcomes vary if it is a proximal vs distal case.[31] Also, we chose not to include variables such as quality of urethral plate because there is no current agreement about how to best classify it. Lastly, we did not include measurements of different penile dimensions because our available anonymous database did not include this information. Future studies will focus on including pixel interpretation as a way to estimate penile dimensions as a new algorithm to automatize penile curvature estimation based on Gaussian curvature evaluation (current ongoing study). Our results prove the concept that the algorithm does emulate an expert experience at classifying hypospadias correctly.

CONCLUSION

Image recognition model after established training has an accuracy detection rate of 90% which emulates the almost perfect inter-rater agreement between experts. Future applications of this technology may be used as a predictive tool for surgical outcomes and to identify image properties to better define difficult variables such as the quality of the urethral plate.

References

1. Van Der ZLFM, Van RIALM, Feitz WFJ, Franke B, Knoers NVAM, Roeleveld N. Aetiology of hypospadias: a systematic review of genes and environment. *Human Reprod Update*. 2012;18:260–283.
2. Kojima Y, Kohri K, Hayashi Y. Genetic pathway of external genitalia formation and molecular etiology of hypospadias. *J Pediatr Urol*. 2010;6:346–354.
3. George M, Schneuer FJ, Jamieson SE, Holland AJ. Genetic and environmental factors in the aetiology of hypospadias. *Pediatr Surg Int*. 2000;31:519–527.
4. Fernandez N, Henao-Mejía J, Monterrey P, Pérez J, Zarante I. Association between maternal prenatal vitamin use and congenital abnormalities of the genitourinary tract in a developing country. *J Pediatr Urol*. 2012;8:121–126.
5. Fernández N, Pérez J, Zarante I. Is hypospadias a spectrum of different diseases? MAMLD1 gen: a new candidate gene for hypospadias. *Urol Colomb*. 2015;24:155–160.
6. Gong EM, Cheng EY. Current challenges with proximal hypospadias: we have a long way to go. *J Pediatr Urol*. 2017;13:457–467. Available from: https://doi.org/10.1016/j.jpurol.2017.03.024.
7. Aslam R, Campbell K, Wharton S, Bracka A. Medium to long term results following single stage Snodgrass hypospadias repair. *Br J Plast Surg*. 2013;66:1591–1595. Available from: http://dx.doi.org/10.1016/j.bjps.2013.06.041.
8. Bush NC, Villanueva C, Snodgrass WT. Glans size is an independent risk factor for urethroplasty complications after hypospadias repair. *J Pediatr Urol*. 2015;11.355.1e-5.
9. Spinoit AF, Poelaert F, Groen LA, Van Laecke E, Hoebeke P. Hypospadias repair at a tertiary care center: long-term follow-up is mandatory to determine the real complication rate. *J Urol*. 2013;189:2276–2281. Available from: http://dx.doi.org/10.1016/j.juro.2012.12.100.
10. Merriman LS, Arlen AM, Broecker BH, Smith EA, Kirsch AJ, Elmore JM. The GMS hypospadias score: assessment of inter-observer reliability and correlation with post-operative complications. *J Pediatr Urol*. 2013;9:707–712.
11. Orkiszewski M. A standardized classification of hypospadias. *J Pediatr Urol*. 2012;8:410–414.

12. Lorenzo AJ, Rickard M, Braga L, Guo Y, Oliveria J-P. Predictive analytics and modeling employing machine learning technology: the next step in data sharing, analysis and individualized counseling explored with a large, prospective prenatal hydronephrosis database. *Urology*. 2018;123:204–209.
13. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115–118.
14. Google. TensorFlow. 2018 [cited 2018 Aug 2]. Available from: https://github.com/tensorflow/models/tree/master/research/object_detection. Accessed September 26, 2020.
15. Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Cohen's Kappa. *Epidemiol Biostat Public Heal*. 2017;14:e12267-1–e12267-10.
16. Holland AJA, Smith GHH, Ross FI, Cass DT. HOSE: an objective scoring system for evaluating the results of hypospadias surgery. *BJU Int*. 2001;88:255–258.
17. Snodgrass WT, Bush N, Cost N. Tubularized incised plate hypospadias repair for distal hypospadias. *J Pediatr Urol*. 2010;6:408–413.
18. Pfistermuller KLM, McArdle AJ, Cuckow PM. Meta-analysis of complication rates of the tubularized incised plate (TIP) repair. *J Pediatr Urol*. 2015;11:54–59.
19. El-Hout Y, Braga L, Pippi Salle JL, Moore K, Bägli DJ, Lorenzo AJ. Assessment of urethral plate appearance through digital photography: do pediatric urologists agree in their visual impressions of the urethral plate in children with hypospadias? *J Pediatr Urol*. 2010;6:294–300.
20. Villanueva CA. Goniometer not better than unaided visual inspection at estimating ventral penile curvature on plastic models. *J Pediatr Urol*. 2019;15:628–633. [cited 2020 Jan 7]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31680019.
21. Spinoit AF, Poelaert F, Van Praet C, Groen LA, Van Laecke E, Hoebeke P. Grade of hypospadias is the only factor predicting for re-intervention after primary hypospadias repair: a multivariate analysis from a cohort of 474 patients. *J Pediatr Urol*. 2015;11:70.e1–70.e6.
22. David W, Green P, Beglinger S, et al. Hypospadias surgery in England: higher volume centres have lower complication rates. *J Pediatr Urol*. 2017;13:481.e1–481.e6.
23. Poletta FA, Gili JA, Castilla EE. Latin American collaborative study of congenital malformations (ECLAMC): a model for health collaborative studies. *Public Health Genomics*. 2014;17:61–67.
24. Huang G, Shan W, Zeng L, Huang L. Androgen receptor gene CAG repeat polymorphism and risk of isolated hypospadias : results from a meta-analysis. *Genet Mol Res*. 2015;14:1580–1588.
25. Johnson EK, Jacobson DL, Finlayson C, et al. Proximal hypospadias: isolated genital condition or marker of more? *J Urol*. 2020;204:345–352.
26. Al-Juraibah F, Lucas-Herald A, Nixon R, et al. Association between extra-genital congenital anomalies and hypospadias outcome. *Sex Dev*. 2019;13:67–73.
27. Chua M, Gnech M, Ming J, et al. Preoperative hormonal stimulation effect on hypospadias repair complications: meta-analysis of observational versus randomized controlled studies. *J Pediatr Urol*. 2017;13:470–480.
28. Gorduza DB, Gay CL, De Mattos E, et al. Does androgen stimulation prior to hypospadias surgery increase the rate of healing complications? A preliminary report. *J Pediatr Urol*. 2011;7:158–161.
29. Luo CC, Lin JN, Chiu CH, Lo FS. Use of parenteral testosterone prior to hypospadias surgery. *Pediatr Surg Int*. 2003;19:82–84.
30. Browne D. An operation for hypospadias. *Proc R Soc Med*. 1949;42:466–468.
31. Pippi Salle J, Sayed S, Salle A, et al. Proximal hypospadias: a persistent challenge. Single institution outcome analysis of 3 surgical techniques over a 10-year period. *J Pediatr Urol*. 2016;12:28.e1–28.e7.