Applications of Machine Learning in Pediatric Urology

FROM THE GUEST EDITORS

Michael Ernst¹ & Christopher Jaeger²

¹Stony Brook University Hospital, Stony Brook, NY
²Boston Children’s Hospital, Boston, MA

Personalized medicine has gained popularity due to its ability to tailor disease screening, workup, and treatment to the individual patient. Ideally, this maximizes desired clinical outcomes while minimizing morbidity and cost. Machine learning algorithms built from volumes of clinical data are a powerful tool in the modern clinical environment. They can be used to offer evidence-based, personalized medicine.

Contributing Authors

Leonid Aksenov, MD
Nicolas Fernandez, MD, PhD
Katherine Fischer, MD
Adree Khondker, BS
Armando Lorenzo, MD, MSc
Adam Maxwell, PhD
Robyn Reed, MD, PhD
Mandy Rickard, NP
Jonathan Routh, MD, MPH
Margaret Shnorhavorian, MD, MPH
Gregory Tasiaan, MD, MSc, MSCE
Hsin-Hsiao Scott Wang, MD, MPH, MBA
Jack Weaver, MD, MSTR
Rena Xu, MD, MBA
Justin Ziemba, MD, MSEd

FROM THE EDITOR

Elizabeth B. Yerkes, MD

The Fellow Representatives always choose intriguing and instructive topics for their DPU editions!

I do have to disclose that I am skeptical of machine learning and artificial anything. Based upon my daily meeting of the minds with the machines (EPIC and his/her/their friends), I have a general distrust that machines could do my job better, helping us see more clearly to provide better counseling for our patients. How could they possibly know? How are they taught what they are learning and disseminating? Smarter mammograms? Excellent! Big datasets? Do those reflect the patients or the decisions we made together to derive their outcomes? In the case of Urodynamics (not included in this Edition), if the tracings are inconsistent and we cannot agree, how is the machine to be trusted? You are going to have to sell that one to me. What about the ‘art’ of medicine? That was my favorite part.

The authors present work that informs and begins to quell my skepticism. They share their progress, optimism and cautions about the field and wrap up with an honest debate. Perhaps I can get behind the personalized aspect of what the machines have to offer. Please sit with me in the row of seats where negative pre-conceptions are cast aside and enjoy this Edition of Dialogues in Pediatric Urology! Thank you to Michael Ernst and Chris Jaeger for their hard work pulling this Edition together while finishing up fellowship and starting clinical year!

With this Edition, I bid farewell to this role with Dialogues in Pediatric Urology and look forward to seeing where our new Editor in Chief Dr. Doug Storm will place DPU within the modern era of communication within our specialty!
From the Guest Editors (continued from page one)

Machine learning considers individual patient characteristics to predict specific outcomes. The applications of machine learning within pediatric urology are rapidly growing. This growth poses substantial implications for clinical care and ongoing research efforts within the specialty. This edition of the Dialogues in Pediatric Urology aims to explore the emergence of machine learning in Pediatric Urology.

This edition will first explain the science of machine learning, including how machine learning differs from deep learning and neural networks. It will then highlight various applications of machine learning in Pediatric Urology using databases derived from clinical data, imaging studies, photographs, and research databases to predict relevant outcomes. Lastly, the edition will emphasize the benefits as well as the potential limitations and drawbacks of machine learning in a debate-style format.

We thank all of our invited contributors for sharing their expertise with the broader Pediatric Urology community in this edition of the Dialogues in Pediatric Urology.

Introduction to the Science of Machine Learning

Jack Weaver¹ & Gregory Tasian²

¹ University Hospitals - Rainbow Babies and Children’s Hospital, Department of Urology, Case Western Reserve University School of Medicine, Cleveland, OH
² Children’s Hospital of Philadelphia, Division of Pediatric Urology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA

In 1959 Arthur Samuel created The Samuel Checkers-playing Program which was among the world’s first successful self-learning programs.¹ This program could successfully compete with amateur checker players and learn from prior experience. Following the development of this program Samuel, a graduate of MIT and developer for IBM, coined the term “machine learning.”¹²

Today, machine learning is a rapidly growing branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention. These programs can improve automatically through experience and the use of data, and ultimately perform tasks that they were not explicitly programmed to do.

Machine learning is currently used in medicine in a variety of ways. In fact, there are fields of medical study that are completely centered around the application of machine learning. Radiomics, for example, is a field in which a large number of features are extracted from medical images using data-characterization machine learning algorithms.¹ Radiomics is fundamentally based on highly complex pattern recognition. Medical images are scanned for suspicious structures with the hopes of identifying abnormalities that may not be evident to a human reviewer. The algorithms play a supporting role while a radiologist or other physician is ultimately responsible for the final interpretation of the medical image. The goal of some radiomics systems is to detect the earliest signs of abnormality on medical images that human observers cannot. In 1998, the first commercial computer aided diagnosis system for mammography, the ImageChecker system, was approved by the US Food and Drug Administration (FDA).⁴ In the following years several commercial systems for analyzing mammography, breast MRI, medical imaging of lung, colon, and heart also received FDA approvals.⁴

To date, extensive radiomics research has been performed by pediatric urologists with a particular emphasis on analyzing renal ultrasounds to predict renal outcomes.⁵,⁶ Specifically, our group has demonstrated that imaging features ascertained from ultrasound data using deep convolutional neural networks improved the classification of children with congenital anomalies of the kidney and urinary tract (CAKUT) and controls.⁸ In our current NIH funded project, deep learning imaging features extracted from the initial postnatal ultrasound of posterior urethral valve patients are being used to predict chronic kidney disease progression.

Another field of medical study that relies heavily upon machine learning is bioinformatics. The field of bioinformatics seeks to increase the understanding of biological processes through the application of computationally intensive techniques such as pattern recognition, data mining, and machine learning algorithms. Bioinformaticians frequently use unsupervised machine learning algorithms to cluster genes with related patterns into gene families, assign genotypes, and infer population structures.

While the application of machine learning in medical research is intriguing, there are notable deficiencies and barriers that exist. For instance, in the field of radiomics there is a lack of standardized assessment measures for computer aided diagnostic systems which can cause difficulty obtaining FDA approval for their commercial use. Studies that have attempted to address this issue by proposing standardized evaluations have not been widely adopted.⁷,⁸ Skepticism and reluctance of healthcare providers to adopt new computer aided diagnostic systems in clinical practice creates an additional challenge. Furthermore, while artificial intelligence has the potential to improve human health by providing new ways to analyze complex data and predict clinical

(continued on next page)
Introduction to Machine Learning (continued from previous page)

outcomes, the current model for artificial intelligence in medical research is a non-recursive, waterfall procedure that has little collaboration between stakeholders and does not consider how humans make decisions through cycles of data acquisition, interpretation, and integration in different real-world environments. As a result, artificial intelligence models are brittle and unresponsive to real-world clinical care. Developing a model that is overfit and lacks generalizability is a potential pitfall of machine learning models created using small, single institution data sets. To avoid this pitfall, it is critical to validate models in separate larger datasets while ultimately calibrating them to accurately navigate through the complex decision-making process of clinical care to achieve the outcome of interest.

The importance of clinicians partnering with data scientists in the conception and development of machine learning models cannot be overstated. As machine learning research becomes more prevalent in pediatric urology it is more important than ever for urologists to effectively collaborate with data scientists. Pediatric urologists with no experience in performing machine learning research can still serve as critical members of the team as they can impart clinical wisdom that will help guide study design and ensure that the model is evaluating clinically relevant questions that will potentially address a known clinical knowledge gap. While machine learning is a great tool, if not harnessed properly, it can result in models with very little clinical relevance.

References:


Nicholas Fernandez1, Adam Maxwell2, Robyn Reed3, Margarett Shnorhavorian1
1 Division of Pediatric Urology, Seattle Children’s Hospital, University of Washington, Seattle, WA
2 Applied Physics Laboratory and Department of Urology, University of Washington, Seattle, WA
3 Department of Pathology, Seattle Children’s Hospital, University of Washington, Seattle, WA

A key contributor in the evolution of medical practice has been the concept of evidence-based medicine.1 Database creations and predictive nomograms have demonstrated their role in advancing the field. A good example has been the evolution of multivariate nomograms for the prediction in prostate cancer. These nomograms have facilitated decision-making and selection of different therapeutic options and their probable outcomes.2-4

Machine learning (ML) technologies, as novel tools, have demonstrated their role in enhancing clinician’s medical practice and analyzing large datasets. Many algorithms have been created to emulate or outperform humans using image and large-scale dataset analysis with semi-autonomous or supervised algorithms.5,6 Most commonly, datasets have been structured with variables that are already known to have a clinical and or predictive value. For example, ML has been applied for the detection, classification, and staging of prostate cancer (continued on next page)
Enhancement of Hypospadias (continued from previous page)

using already known predictive nomograms based on PSA, TNM and Gleason score.\textsuperscript{5-7}

Nonetheless, the main limitation of this and other approaches has been the conception and creation of the algorithm around a known predictive nomogram with arbitrary selection of variables to build the dataset for analysis. For this reason, the future of artificial intelligence algorithms using large scale comprehensive databases that include clinical, demographic, genetic and even medical performance variables may support identification of novel predictive variables and enhance medical practice as we know it now.

Hypospadias is a good example where phenotypification is subjective, lacking good predictive value with poor correlation to the genotype. The use of ML technologies may enable detection of novel phenotype predictive variables that quantitatively describe tissue quality and might improve subjectivity, aid pre and intraoperative decision making, and predict surgical outcomes. For decades, hypospadias surgical management has centered around anthropometric assessments based on the location of the urethral meatus, severity of penile curvature, and urethral plate appearance.\textsuperscript{8} With the aim to reduce subjectivity in phenotyping patients with hypospadias, a computational statisti-
cal algorithm commonly employed in ML using digital images of patients with hypospadias is demonstrated, for the first time, how it was possible to emulate experienced surgeons’ phenotype assessment and classification.\textsuperscript{4} However, this approach lacked a key intangible component in the assessment of the phenotype which is what most surgeons call: “tissue quality”. In real-time, surgeons, purely by subjective visual assessment and almost as an unconscious “artistic process”, select and handle tissues to reconstruct the congenital defect. The success of hypospadias surgical reconstruction is difficult to reproduce even by the same experienced surgeon.

The impact of the genotype and “tissue quality” on the reconstruction have not been thoroughly explored to date. With this concept, ongoing research is being focused on evaluating “tissue quality” using digital image pixel cluster analysis and its correlation to histological architecture and the genotype to predict their role in surgical outcomes. Pilot data analysis on surgery-naïve patients has demonstrated that colorimetric k-means cluster analysis can differentiate urethral plates with and without chronic inflammation. A total of 22 patients underwent hypospadias surgery for the first time. Urethral meatus was distal shaft in 7 patients, 6 coronal, 4 glanular, 3 midshaft, 2 penoscrotal. Average

(continued on next page)
Enhancement of Hypospadias (continued from previous page)

GMS score was 7.14 (± 1.58). Twenty-one patients had complete histological analysis by a blinded pathologist. Of those, 11 (52.3%) had an abnormal pathology report. Six had presence chronic inflammation at the urethral plate. The second most common finding was hyperkeratosis. Digital image segmentation and k-means pixel analysis created categorical color grouping and differentiation for each region of interest including urethral plate (Figure 1). Patients with urethral plate inflammation displayed a mean k1 cluster distance of 64.2 vs 53.1 for those without (p=0.022). Similar for k2, a mean cluster distance of 31.1 with inflammation vs 19 without was found (p=0.032) (Figure 2), indicating the algorithm can identify measurable differences from these digital images. Future directions of our group aim to develop of semi-supervised convolutional neural network algorithm using this novel comprehensive approach that includes, pixel cluster analysis, anthropometric assessments, histological architecture, and genetic data to help us understand and predict how we see and treat hypospadias and improve our assessment of the tissue quality used in the reconstruction. Our approach will also serve to merge and incorporate next generation genomic sequencing data to explore the role of variable expressivity in tissue quality and surgical outcomes to enhance surgeon’s assessment of patient’s phenotype.

References:

Moving Towards Objective Grading of Vesicoureteral Reflux with Machine Learning

Adree Khondker, Mandy Rickard, Armando Lorenzo
Division of Urology, Department of Surgery, The Hospital of Sick Children, Toronto, ON

Vesicoureteral Reflux Grading

Vesicoureteral reflux (VUR) is often a benign condition that may downgrade or resolve without intervention. However, in the event of recurrent urinary tract infections (UTIs) surgical management may be indicated in order to prevent UTIs and renal scarring.1,2 The method of VUR management is dependent on the grade, which ranges from I-V in accordance with the international grading system.3 However, as with other grading systems, such as the SFU grading system for hydronephrosis, assigning VUR grades is subjective and provider dependent. The current reported inter-rater reliability for VUR grading is 60%, resulting in one-third of voiding cystourethograms (VCUG) having discordant grades between clinicians.4,5

When assigning VUR grades, ureter dilatation, tortuosity, and blunting of renal calyces is considered. For example, grade III is defined as “mild” dilatation and tortuosity of the ureter, whereas grade IV VUR has “moderate” dilatation and significant tortuosity. Defining grades based on these subjective thresholds leads to bias and disagreement between clinicians. Artificial intelligence (AI) can improve current grading inter-clinician reliability, understanding, and strengthen relationships between clinical findings with disease severity.6

qVUR: a quantitative tool to determine VUR grade

AI is built upon machine learning (ML) which uses correlations in images and metrics to build analytical models and make predictions.7 We developed qVUR, a ML-based model which uses four features from a VCUG to determine VUR grade: proximal ureter width, distal ureter width, maximum ureter width, and ureter tortuosity (Figure 1). In our most recent work,8 these features predicted low (I-III) versus high (IV-V) VUR grade with an AUC of 0.90. Since then, we have extended our model to predict individual VUR grade, by training it on VCUGs from more than 1000 children with VUR at multiple institutions, with an AUC of 0.83. Our model reliably grades the same VCUG with the same parameters, rather than a clinician who will apply subjective cri-

(continued on next page)
**Objective Grading** (continued from previous page)

...va...
Automated Segmentation of the Kidneys, Ureters and Bladder Using a Convolutional Neural Network Model to create a “Urinary Tract Atlas”

Background and Objectives

Urolithiasis is a chronic condition characterized by episodic and often debilitating symptoms, including pain, at the time of stone passage. When patients initially present with a symptomatic ureteral stone there are two main treatment approaches: conservative treatment with a trial of stone passage or early surgical intervention with either stent placement or definitive stone removal. Though prior studies have attempted to identify factors that accurately predict the likelihood of spontaneous stone passage, such as stone size and location, it remains difficult to predict which patients will pass their stones if given the opportunity versus those who will ultimately require surgical intervention.1

Observation with a trial of passage is ideal for those patients who will eventually pass their stones because it avoids the unnecessary risks of surgery and anesthesia as well as the potential pain and morbidity known to be associated with stone procedures. However, observation can also lead to additional emergency room visits, missed days of school and work and an extended period of severe symptoms with the patient ultimately requiring surgery anyways. In these situations, the trial of passage is less advantageous and may be unnecessarily burdensome to the patient.

The goal of our group, the Children’s Hospital of Philadelphia and University of Pennsylvania Center for Machine Learning in Urology, is to harness the power of deep learning to improve diagnosis, management, and clinical outcomes of common benign urologic conditions, including stone disease, in pediatric and adult patients. Prior work by our group has shown that the automated measurement and characterization of kidney stones and renal anatomy using a deep learning model is feasible, accurate and more efficient when compared to manual measurements. Our long-term aim is to create a deep learning model that incorporates both patient clinical characteristics and CT imaging characteristics to accurately predict the likelihood of ureteral stone passage for an individual patient’s stone episode leading to improved, individualized patient care.

Figure 1. Original images, ground truth based on manual segmentation and predicted segmentations based on current deep learning model for kidneys (red), ureters (green) and bladder (blue).
As an interim step in pursuit of this goal, we have been working to develop a novel machine learning algorithm that segments the kidneys, ureters and bladder on CT scan in an automated fashion, which we have termed a “Urinary Tract Atlas.” The creation of this initial model is necessary because of the difficulty in identifying and visualizing the ureters on non-contrast CT scan, which is the most common imaging modality used to evaluate ureteral stones in clinical practice. The ureters can be difficult to accurately trace on non-contrast imaging even for experienced radiologists and urologists. Therefore, in order to train a machine learning model that incorporates imaging characteristics of ureteral stones, the model first must be taught to accurately localize the ureter and any stones within it.

Using CT urograms (CTUs) obtained from adult patients undergoing imaging for the diagnoses of either hematuria or nephrolithiasis we have begun to train and test this model. Each CTU is first manually segmented by labeling the kidneys, ureters and bladder. Manual segmentation involves a urologist tracing and outlining the relevant renal and ureteral anatomy on each slice of the CT scan. These manual labels serve as the ground truth for model training as well as for assessment of accuracy of the model. Figure 1 shows examples of kidney, ureter and bladder labels generated using our current atlas model along with the original images and manual segmentations that served as ground truth. Similarly, Figure 2 shows a 3D rendering of the urinary tract as segmented using our model compared with manual segmentation.

**Future Directions**

We have developed a novel machine learning model that can label the kidneys, ureters and bladder in an automated fashion. This will provide a foundation for our team to build upon and create a machine learning model to predict the likelihood of ureteral stone passage using features extracted from CT images as well as clinical data. Next steps include training the current Atlas model to segment the kidneys, ureter and bladder on non-contrast CT scans as opposed to CTUs since non-contrast CT scans are the type of imaging most commonly used to evaluate ureteral stones in clinical practice. Once the model is able to accurately recognize the ureters and ureteral stones, our hope is that it can be used to extract imaging features that may aid in the more accurate predication of stone passage.

Deep learning has the advantage of being able to identify and utilize imaging characteristics not traditionally recognized or measured by clinicians and we believe this will enable us to create a more accurate model to predict stone passage than clinical characteristics or manual measurements alone. Ultimately, this model will allow the delivery of better individualized patient care by identifying those patients most likely to benefit from early surgical intervention for urolithiasis versus those likely to have a successful trial of passage. The Urinary Tract Atlas model described here may also be useful in developing machine learning applications for a variety of urologic conditions, both malignant and benign, affecting the ureter as automated segmentation of the ureter on CT scan has not been previously described.

**References:**


**Acknowledgement:**

The authors would like to acknowledge the entire Penn Center for Machine Learning in Urology team for their work on this project including Yuemeng Li, Yong Fan, Benjamin Schurhamer, Joanie Garrat, Joey Logan, Ester Choi, Dawud Hamden, Alice Xiang, Iqra Nadeem and Roby Daniel.

**Funding:**

NIDDK P20 CHOP/ Penn Center for Machine Learning in Urology AUA Care Foundation and SPU Sushil Lacy Research Scholar Award (KMF)
Power of Modern Analytics in Personalized Management for Children with UTI/VUR

Classic biostatistics methods such as regressions have generated significant insights for clinicians in terms of description of care patterns and factors associated with clinical outcomes. However, due to the statistical modeling restrictions, the predictive performance for such methods is usually poor with high sample size requirements. To enable personalized care, accurate predictions are necessary for prediction model implementation. Modern analytical methods such as machine learning (ML) algorithms provide unprecedented advantage for this task. It does not require restrictive assumptions about data-generating processes. This allows models to empirically identify higher-order interactions and non-linear relationships between variables and use a broader range of functions to fit patterns. This will not only allow for testing known hypotheses but also allow for identifying new, testable predictive relationships. This property also reduces the burden of curating data which is often noisy due to human input error. Additionally, there are well-described methods to avoid overfitting, a critical property when working with high-dimensional data such as that for our cohort, generalize predictions or inferences made on a training dataset to an unknown test dataset, which essentially replicates their use in clinical settings. Furthermore, since ML algorithms are not pre-specified, predictive performance is expected to continually improve post-deployment as the model retrains on new data.

In clinical decision making, we as pediatric urologists have been offering somewhat personalized care to our patients based on our training, experience, and knowledge. This often leads to significant practice variation. It is difficult to compare, review, and truly learn from those decisions in a scalable and evidence-driven fashion.

One of the hallmark examples is the controversy in choosing the timing of voiding cystourethrogram (VCUG) for children who presented with urinary tract infection (UTI). Given VCUG’s invasive nature with catheter placement, iatrogenic infection risks, and radiation, there is a desire for more judicious VCUG use. American Academy of Pediatrics 2011 guidelines recommended deferring VCUG until after the second UTI which is not in line with urologic communities who preferred a lower VCUG threshold. The guidelines are controversial, and groups have expressed concern regarding the potential consequences of a delayed diagnosis of vesicoureteral reflux (VUR). This controversy has led to significant practice variability in VCUG timing and suggested that our current clinical tools are inadequate to guide management in targeted fashion.

Another example is the choice of continuous antibiotic prophylaxis (CAP) in children with VUR. The Randomized Intervention for Children with Vesico-Ureteral Reflux (RIVUR) trial and meta-analyses have shown that continuous antimicrobial prophylaxis is effective in reducing recurrent UTI in children with VUR. However, significant controversies remain regarding the role of CAP in VUR management. For example, the trial was underpowered to define CAP impact on most significant longer-term outcomes like renal scarring. Additionally, some have expressed concerns with potential CAP side effects on somatic growth and antibiotic resistance. Although CAP generally appears safe and the suspected long-term effects are debatable, there is little doubt that more selective use of CAP would be preferable to a one-size-fits-all approach. As the hypothesis of individualized drug responses flourishes, individualized treatment has become a hot research area. In reality, many clinicians have already adapted their practice based on this assumption. Instead of all-or-none, CAP prescription is often “personalized” by patient characteristics such as index UTI severity and VUR grade; such decisions are typically based on clinician experience, and lack transparency with respect to the factors that go into decision-making.

(continued on next page)
By joining forces between Boston Children’s Hospital and Massachusetts Institute of Technology, we developed novel ML models based on large clinical trials data to allow accurate risk prediction.\textsuperscript{11,12} We also developed user-friendly user interface and offered these model for free to reduce the barrier for adaptation (PredictVUR as free app in Apple store as shown in Figure). If we can accurately predict the individual outcomes ahead of time based on robust data in these challenging scenarios, we can achieve optimized personalized care based on data-driven risk estimates instead of arbitrary preference between often conflicting guidelines. These high-performance ML model serve as promising basis to prove the feasibility to develop models and reliably predict the outcome of interest. With these actionable insights, we can individualize management and counsel family in truly data-driven fashion. Only those who need the management (VCUG and CAP as our examples) will get it, whereas those who don’t need these interventions would be spared for the invasiveness, adverse effect, and likely wasted resource. Proper implementation of such prediction models will empower the clinicians to be confident in providing best individualized management to reach the true value-based care for our patients.

References:

Machine learning, a field of artificial intelligence that utilizes computer algorithms to learn from datasets, recognize patterns, and guide decision-making, is rapidly gaining momentum in healthcare. This is the direct result of increasing computational power and increasingly available healthcare data. Machine learning is actively being applied to adult urology and is gaining a foothold in pediatric urology. The value of machine learning lies in its ability to analyze large datasets, identify patterns that may be overlooked by standard statistical methods, and improve efficiency in healthcare. While the potential risks of any new technology must be recognized, we argue that these risks are outweighed by the value of thoughtful, deliberate, and transparent implementation of machine learning in pediatric urology.

Technological advances in healthcare have generated an overabundance of data. It is estimated that a typical neonatal intensive care unit generates 40 terabytes of data per year. While some might say data is power, we argue that too much data can actually handicap our ability to gain knowledge – in the modern era, there is often too much data for us to process! In order to derive clinically meaningful knowledge from clinical data, we must utilize machine learning to make sense of big data and draw conclusions. In adult urology, machine learning algorithms have demonstrated better predictive efficiency compared to conventional regression analysis when predicting post-prostatectomy biochemical recurrence and graft survival in renal transplant recipients. The hidden patterns in datasets found via machine learning may reveal previously unrecognized truths – such as the identification of thousands of previously unrecognized virus genomes. In pediatric urology, a machine learning algorithm created for the prediction of vesicoureteral reflux (VUR), identified previously underappreciated clinical associations between recurrent urinary tract infection (rUTI) associated VUR and symptomatic non-febrile rUTI. Likewise, machine learning techniques have been applied to other vexing issues in pediatric urology such as management of hydronephrosis.

Machine learning can make healthcare more productive and efficient; this allows providers to focus more on patients. This analytical technique is ideally suited for the synthesis of abundant and shallow data which could augment the work providers already do by helping establish a diagnosis, planning treatment, and predicting outcomes. Decreases in workload would allow providers to focus on the holistic aspects of healthcare which cannot be replicated by algorithms or machines. To be clear, the nuanced choice of whether or not to pursue a specific treatment can only be made within the context of the provider/patient dyad – we are not yet at the point where algorithms can accurately model human decision-making (particularly for preference-sensitive conditions).

While machine learning holds promise in pediatric urology, we must acknowledge and consider liabilities that come with this tool. An algorithm is only as good as the dataset it is trained on – data inconsistencies and biases could be cemented by an algorithm and incorporated into models. Machine learning is also vulnerable to black box effect, spurious associations, and dataset shift. In order to mitigate these liabilities, it is imperative that transparent algorithms are developed from carefully curated datasets through the multidisciplinary collaboration of experts from urology and computer science – with a strong emphasis on keeping results relevant to patient care and a keen focus on ethics.

Machine learning has incredible potential to improve the quality-of-care delivery and decision-making in pediatric urology. It will allow us to make sense of the massive amount of data that is produced, identify patterns that we cannot identify using conventional means, and increase efficiency in medicine.

References:
4. Wong NC, Lam C, Patterson Le et al.: Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy. BJU Int, 123; 51, 2019
7. MIT ORC Personalized Medicine Group and Advanced Analytics Group of Pediatric Urology: Targeted workup after initial febrile urinary tract infection: Using a novel machine learning model to identify children most likely to benefit from voiding vystourethrogram. J Urol, 202; 144, 2019
Con: The Unfulfilled Potential of Machine Learning in Clinical Practice

Machine learning (ML), a form of artificial intelligence that uses algorithms to learn from data, has garnered increasing attention in recent years for its promise to enrich clinical decision-making. In pediatric urology, use of ML has been proposed to predict detrusor overactivity, obstruction after pyeloplasty, and the need for antibiotic prophylaxis in children with vesicoureteral reflux. While the potential benefits of ML are tremendous, certain limitations must be addressed before it can progress from academic exercise to clinical asset.

First, greater transparency is needed into how ML predictions are formulated. This matters not just for model validation purposes but fundamentally for understanding what a model contributes: does it offer new insight or merely reflect clinicians’ pre-existing inclinations? The distinction has important implications for generalizability and clinical relevance. While physiologic patterns are likely to transcend data sets, clinician behavior frequently varies across institutions and geographies. ML models that base predictions on clinician-influenced training data may not generalize to data sets from other practice settings – a prerequisite of any clinical tool using ML output. A study by Beaulieu-Jones et al. used 43 million inpatient hospitalizations to assess a risk-stratification model built on “clinician-initiated” data – billing data for clinician activities from the first day of a patient’s hospitalization – as compared to a benchmark model built on detailed electronic health record (EHR) data. The model using clinician-initiated data achieved performance nearly equivalent to that of the EHR benchmark for predicting mortality, readmissions, and length of stay. This suggests the benchmark and similar models, which were supposedly trained on complete data sets, may simply be predicting clinician judgments rather than generating independent insights from more objective data elements. Without realizing this, clinicians might assume the model perceives something they do not and follow its recommendations inappropriately.

Second, while not everyone must understand how ML algorithms work, a typical clinician should feel comfortable using the products of ML in their practice. Just as the ability to interpret traditional statistical methods allows clinicians to evaluate new clinical evidence, basic skills for assessing ML products will enable clinicians to incorporate ML-driven insights confidently into their decision-making rather than exclude or defer blindly to them. Additionally, medicolegal standards for following or overriding ML-based recommendations must be codified so clinician decisions are not swayed by fear of legal consequence. Even seemingly benign clinical decision support tools can adversely influence provider behavior. In a multi-center randomized controlled trial by Perry et al., for instance, the use of an EHR alert flagging patients at risk for acute kidney injury was associated with significantly increased mortality at certain hospitals. System-wide education, training, and guidelines for the use of ML-based tools will be essential to avoid harm.

Addressing these limitations of ML represents a necessary and feasible first step toward achieving clinical utility. Like any tool, ML will only have a positive impact if wielded correctly. With the right preparation and safeguards, it can transform medical practice; without them, it may prove dangerous or simply unworkable, a powerful tool that cannot leave the shed.

References: