Development and use of machine learning models for prediction of male sling success

A proof-of-concept institutional evaluation

Jin K. Kim[1,2*], Kurt A. McCammon[3*], Kellie J. Kim[4], Mandy Rickard[2], Armando J. Lorenzo[1,2], Michael E. Chua[2,5]

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; [2]Division of Urology, Department of Surgery, The Hospital for Sick Children, Toronto, ON, Canada; [3]Department of Urology, Eastern Virginia Medical School, Norfolk, VA; Devine-Jordan Center for Reconstructive Surgery and Pelvic Health, Urology of Virginia, Virginia Beach, VA, United States; [4]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada; [5]Institute of Urology, St. Luke's Medical Center, Quezon City, Philippines

*Equal contributors

Cite as: Kim JK, McCammon KA, Kim KJ, et al. Development and use of machine learning models for prediction of male sling success: A proof-of-concept institutional evaluation. *Can Urol Assoc J* 2023;17(10):E309-14. http://dx.doi.org/10.5489/cuaj.8265

Published online July 21, 2023

Appendix avaialble at *cuaj.ca*

ABSTRACT

INTRODUCTION: For mild to moderate male stress urinary incontinence (SUI), transobturator male slings remain an effective option for management. We aimed to use a machine learning (ML)-based model to predict those who will have a long-term success in managing SUI with male sling.

METHODS: All transobturator male sling cases from August 2006 to June 2012 by a single surgeon were reviewed. Outcome of interest was defined as 'cure': complete dryness with 0 pads used, without the need for additional procedures. Clinical variables included in ML models were: number of pads used daily, age, height, weight, race, incontinence type, etiology of incontinence, history of radiation, smoking, bladder neck contracture, and prostatectomy. Model performance was assessed using area under receiver operating characteristic curve (AUROC), area under precision-recall curve (AUPRC), and F1-score.

RESULTS: A total of 181 patients were included in the model. The mean followup was 56.4 months (standard deviation [SD] 41.6). Slightly more than half (53.6%, 97/181) of patients had procedural success. Logistic regression, K-nearest neighbor (KNN), naive Bayes, decision tree, and random forest models were developed using ML. KNN model had the best performance, with AUROC of 0.759, AUPRC of 0.916, and F1-score of 0.833. Following ensemble learning with bagging and calibration, KNN model was further improved, with AUROC of 0.821, AUPRC of 0.921, and F-1 score of 0.848.

CONCLUSIONS: ML-based prediction of long-term transobturator male sling is feasible. The low numbers of patients used to develop the model prompt further validation and development of the model but may serve as a decision-making aid for practitioners in the future.

INTRODUCTION

Stress urinary incontinence (SUI) in males is a bothersome condition that usually results from sphincter deficiency secondary to prostate surgeries or insults to the pelvic floor or sphincter.[1] While conservative management, including pelvic floor therapy, can be helpful, those who have refractory symptoms are considered for surgical management, usually in the form of artificial urinary sphincter (AUS) and more recently, transobturator male slings.[2-4] AUS remains a preferred option for moderate to severe SUI following radical prostatectomy; male slings can be considered for mild to moderate SUI.

The success rates of transobturator male slings have been reported as >70%. Clinical characteristics that were identified as predictors of success based on traditional regression included concomitant urge incontinence symptoms and preoperative SUI severity.[5] Machine learning (ML) has been used in clinical medicine, especially in the area of personalized medicine, using data to predict and classify patients based on diagnosis or likelihood of treatment success.[6] Despite its potential benefits, ML has not been applied in creating a model to predict male SUI patients who may benefit from transobturator male slings. Hence, this investigation aimed to develop an ML algorithm to predict male SUI patients who will have success following transobturator male sling insertion.

METHODS

Following approval by the institutional research ethics board (REB# 18-10-WC-0236), a retrospective assessment of prospectively maintained database on all male patients who underwent transobtruator male sling (AdVance Boston Scientific, Minnetonka, MN, U.S.) insertion between August 2006 and June 2012 by a single surgeon was conducted. All redo cases were excluded from the analysis. Patients with missing data were excluded from the analysis. The operative technique has been previously described.[7] While patients with severe SUI are counselled to pursue AUS, those who prefer male sling are still offered the procedure.[5] The collected data was internally validated by a random counter-verification of 15% of total extracted data.

The ML classifier models assessed included logistic regression, K-nearest neighbor (KNN), naive Bayes, decision tree, and random forest. The preoperative clinical variables included in the database were: age, smoking status, diagnosis of diabetes, race, height, weight, prior prostatectomy, prior pelvic radiation, prior SUI management, bladder neck contractures, type of incontinence, severity of urgency incontinence (if present), potential need for concomitant procedures, number of preoperative pads, and etiology of incontinence. Plot densities of these variables were created to determine whether each variable affected outcomes (Supplementary Figure 1; available at *cuaj.ca*). The closer the overlapping lines on these graphs, the less discrimination for 'cure' based on the assessed variable. Based on these, diagnosis of diabetes and concomitant procedures were less likely to contribute significantly to the model and were removed for feature selection.

Python 3.8.13 (Python Software Foundation, *http://python.org*) was used for model development. Following confirmation of lack of significant outliers, the continuous variables were standardized and scaled to each feature, avoiding biases due to variables being measured at different scales and contributing unequally to models. The models were built using an 85:15 train-test split (85% of data used for model training, 15% used for model performance evaluation, split at random). Grid search was performed to optimize and tune the hyperparameters of KNN and random forest models. Based on the highest performing algorithm, further ensemble learning method was applied. For this study, bagging method was chosen, as it allows a random sample data in a training set to be selected with replacement to allow individual data to be used more than once. These new samples are trained independently and parallel to each other. In the end, based on these independently trained models, the majority prediction is taken and used to produce a more accurate outcome.

The model performance was assessed using sensitivity, specificity, area under receiver operating characteristic curve (AUROC), area under precision-recall curve (AUPRC), and F1-score. Validation curve of the ensemble model was assessed using validation curve. The model was subsequently calibrated using sigmoid method with three cross-validations to improve validation. We further interpreted the explainability of the KNN model by identifying the most important features in the final calibrated ensemble model using permutation importance.

RESULTS

A total of 181 patients were included in our analysis. Following train-test split at random, 153 patients were included in the training set and 28 patients were included in the testing set. Overall, the mean followup was 56.4 months (standard deviation [SD] 41.6), with at least 24-month followup data available for all patients. Slightly, more than half (53.6%, 97/181) of patients had procedural success. The baseline characteristics of patients in each training and testing set are summarized in Table 1.

The five classifier models were developed using our data. Grid search showed that the best hyperparameters for random forest classifier was n-estimator of 57 and best hyperparameters for KNN was n-neighbor of 23 with uniform weight. The performances for each model are summarized in Table 2.

The AUROC and AUPRC curves were developed for all models (Figure 1). KNN model had the highest balanced performance among the five models, with AUROC of 0.759, AUPRC of 0.916, and F-1 score of 0.833. The ensemble KNN model was developed using the bagging method. The bagging KNN model had AUROC of 0.791, AUPRC of 0.919, and F-1 score of 0.812 (Table 2).

A validation curve was built for the bagging CNN model to assess for model reliability. The initial curve had a poor predictability when the predicted probability of cure was low. It was also more likely to over-forecast when predicted and true probability were <0.5, and under-forecast when predicted and true probability were >0.5. The bagged CNN model was then calibrated using sigmoid method. Following calibration, validation curve showed a curve closer to perfect calibration. It was much less likely to over-forecast with lower predictabilities. While it is still likely to under-forecast the likelihood of cure, it may serve as a more

conservative tool to prevent false-positives (Figure 2). The calibrated model had the best performance among all models, with AUROC of 0.821, AUPRC of 0.921, and F-1 score of 0.848 (Figure 3). The calibrated model hyperparameters are detailed in Appendix A (available at *cuaj.ca*).

Using permutation importance, the top features contributing to the model were identified. Preoperative number of pads used was the most important feature in predicting success of male sling. In addition, weight and height (likely interacting to behave like body mass index), as well as severity of incontinence and type of incontinence had relatively greater importance compared to other features (Supplementary Figure 2; available at *cuaj.ca*).

Table 1. Summary of baseline characteristics for training and testing groups

	Training (n=153)		Testing (n=28)	
	Median, n	IQR, %	Median, n	IQR, %
Age at sling surgery (years)	67.5	63.6–72.0	67.6	60.8–74.4
Height at sling surgery (cm)	177.8	172.7–182.9	177.8	172.9–182.9
Weight at sling surgery (kg)	89.6	81.4–98.9	90.2	84.0–94.2
Preoperative number of pads	3.5	2.0–5	3.5	2.0–4.6
Race				
White	105	68.6%	19	67.9%
Black	39	25.5%	7	25.0%
Other	9	5.9%	2	7.1%
Incontinence type				
Stress urinary incontinence	109	71.2%	20	71.4%
Mixed urinary incontinence	44	28.8%	8	28.6%
Incontinence etiology				
Prostatectomy	140	91.5%	25	89.3%
Other prostate therapy	7	4.6%	1	3.6%
Radiation	3	2.0%	2	7.1%
Neurogenic bladder/spinal cord injury	3	2.0%	0	0.0%
Concomitant procedure required				
No	136	88.9%	24	85.7%
Yes	17	11.1%	4	14.3%
Pelvic radiation				
No	121	79.1%	21	75.0%
Yes	32	20.9%	7	25.0%
Prostatectomy				
No	15	9.8%	4	14.3%
Yes	138	90.2%	24	85.7%
Smoking history				
No	72	47.1%	12	42.9%
Prior smoker	69	45.1%	14	50.0%
Current smoker	12	7.8%	2	7.1%
Bladder neck contracture				
No	118	77.1%	24	85.7%
Yes	35	22.9%	4	14.3%

IQR: interquartile range.

DISCUSSION

Artificial intelligence and ML allows creation of models that predict outcomes by performing challenging tasks as humans would do.[8] Experts in their field often use personalized medicine with a plethora of learned experience and prior knowledge; however, for those with less experience, being able to have a model that will provide predictions based on a large patient dataset will allow validation of one's own prediction and aid clinical decision-making for healthcare practitioners and informed decision-making for patients.

With increasing computing power available, there is growing interest in using ML to aid medical decision-making and predictions. In urology, there have been efforts to incorporate ML to predict surgical outcomes in benign prostatic hyperplasia, as well as oncologic outcomes, such as biochemical recurrence following radical prostatectomy.[9,10] ML has also been used for endourologic procedures and urolithiasis.[11] Despite numerous ongoing efforts to use ML in urologic practice, there has not yet been a study that assesses its utility in predicting male urinary incontinence surgery outcomes. This study sought to assess the potential clinical utility of ML in prediction of surgical outcome following transobturator male sling surgery. Using our institutional single-surgeon database, we were able to train a ML model to achieve high predictive potential, with AUROC and AUPRC of 0.821 and 0.921, respectively.

While personalized medicine is becoming increasingly popular in fields such as oncology, where there is heterogeneity across disease and genes that may be strongly associated with outcomes, it can be possible across all fields of urology, including male functional urology.[12,13] Mourmouris et al suggested this is possible as they described prediction of clinical outcomes of

Table 2. Summary of model performance in achieving 'cure' status from male sling

Model	Sensitivity	Specificity	AUROC	AUPRC	F-1 score
Logistic regression	0.764	0.636	0.701	0.765	0.765
K-nearest neighbor	0.882	0.636	0.759	0.916	0.833
Naive Bayes	0.764	0.727	0.746	0.891	0.788
Decision tree	0.647	0.727	0.687	0.842	0.750
Random forest	0.823	0.727	0.775	0.801	0.750
Bagging K-nearest neighbor	0.764	0.818	0.791	0.919	0.812
Calibrated bagging K-nearest neighbor	0.823	0.818	0.821	0.921	0.848

urinary flow following benign prostatic hyperplasia surgery using a random forest classifier model.[9] Using the same database used in this study, Chua et al reported that long-term outcomes of the preoperative moderate to severe SUI was the only independent predictor for failure to achieve cure in long-term followup.[5] To supplement this knowledge, our ML-based model, using the same patient data, allows patients to understand their individualized risk of failing to achieve cure based on additional clinical characteristics that are incorporated into the algorithm training.

Limitations

There are several limitations to this study. The first is the sample size; there were limited number of patients involved in developing this model, which may make our model noisy and with high degree of variance. Nonetheless, as KNN performed the best among the five models that were initially assessed, we attempted to decrease the amount of noise and variance affecting the prediction and were able to do so with ensemble learning. This is likely due to the nature of KNN, which performs relatively well in small datasets. As the dataset grows, KNN may become financially and computationally inefficient, as it requires more memory and data storage compared to other models. It is true that there are concerns with overfitting (model too closely resembles the training dataset and may not perform well to external data) with KNN; however, we attempted to minimize this effect by using hyperparameter tuning to attain a k of 23, meaning prediction is based on 23 patients with the most similar characteristics as the one being assessed (i.e., if ≥12, the majority of the neighbors, had achieved 'cure' among the 23 neighbors, the model will predict 'cure' for the assessed patient; Supplementary Figure 3; available at *cuaj.ca*). Generally, the higher number of k leads to less overfitting, as it averages the value over a greater neighbourhood.[14] Moroever, using plot densities, we attempted to reduce the number of features to reduce the noise by removing variables that did not seem to significantly contribute to the model.

We also constructed validation curves to assess for model reliability and were able to further calibrate our bagged KNN model to create the model with best prediction. The validation curves allow us to assess how

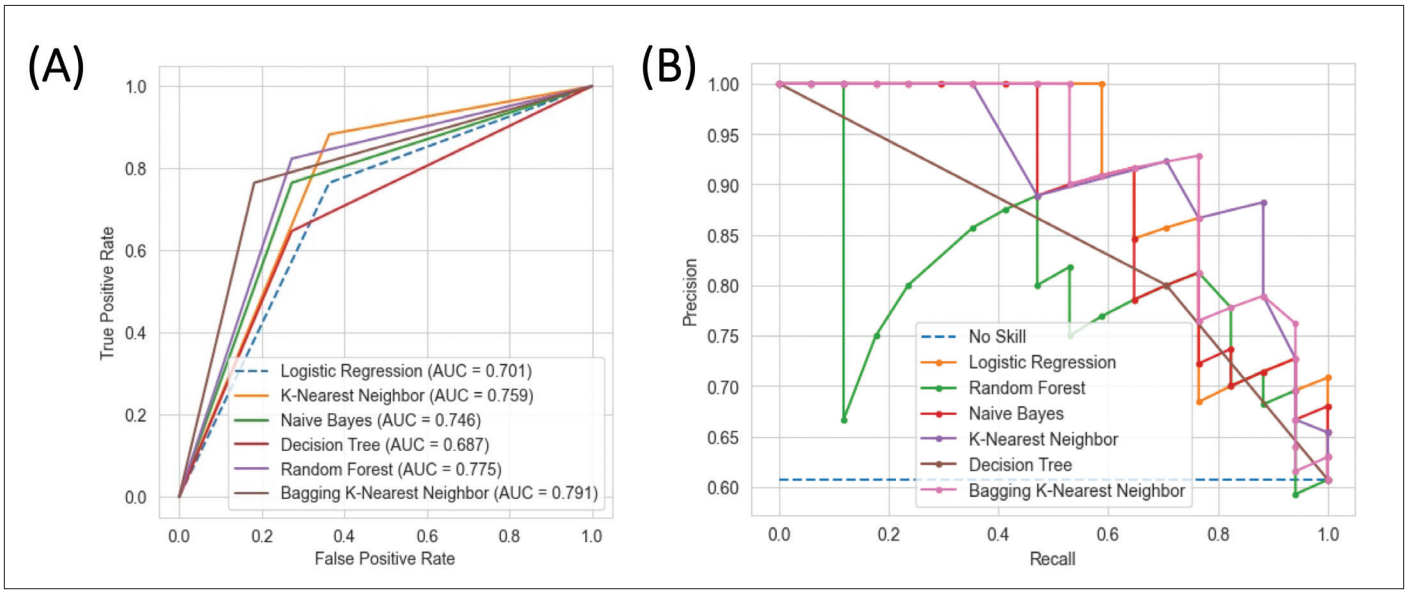

Figure 1. (A) Receiver operating curves for evaluated models. (B) Precision recall curves for evaluated models.

well-calibrated the model is in predicting its outcome. Our bagging KNN with calibration showed that while it may have a slightly higher likelihood of overpredicting success for some, there is a good concordance with a theoretical perfectly calibrated model (Figure 2). Moreover, the more significantly contributing variables in the model (height, weight, severity of incontinence, and type of incontinence) appear to be variables that have been shown in the past to be predictive of male sling success or SUI outcomes.[15-17] While the high predictive potential from our current model suggests there may be a role for use of ML models in clinical practice of male SUI, its immediate use and generalization is limited by both a small study cohort and single-surgeon series of patients. As the model was developed using retrospectively collected data, there may also be a predisposition to selection bias. Moreover, the outcome measures did not use validated questions, such as ICIQ-SF, to assess patient-reported outcomes; however, as our outcome measure was complete dryness or zero pads used, there may be less subjectivity in reporting.

Despite the limitations, this is the first study using ML models to predict outcomes in male SUI. Our model shows promise for use in clinical practice. Via increasing the amount of patient data by prospectively maintaining our database and involving other institutions, we hope to continue to evolve this model to improve its predictive performance and generalizability.

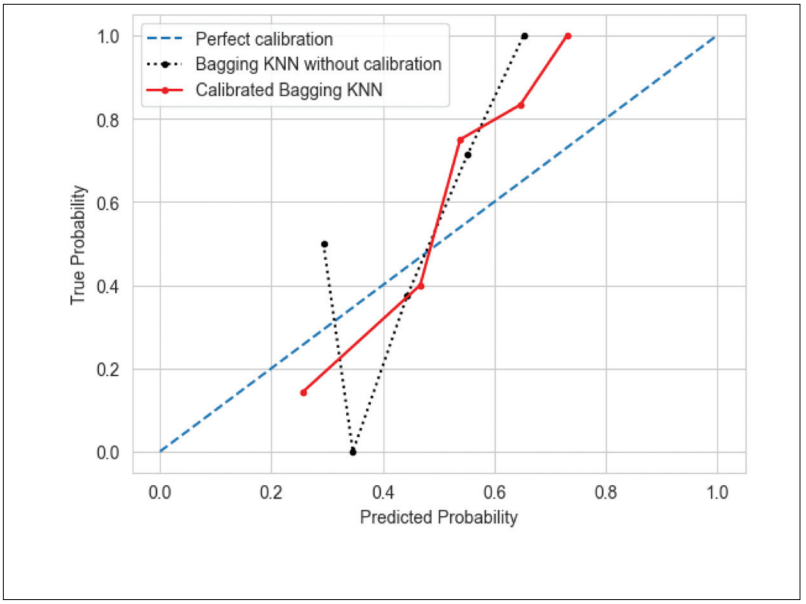

Figure 2. Validation curves for bagging K-nearest neighbor (KNN) algorithm pre- and post-calibration.

CONCLUSIONS

ML-based prediction of long-term transobturator male sling is feasible. The low numbers of patients used to develop the model prompt further validation and development of the model but may serve as a decision-making aid for practitioners in the future.

COMPETING INTERESTS: The authors do not report any competing personal or financial interests related to this work.

This paper has been peer-reviewed.

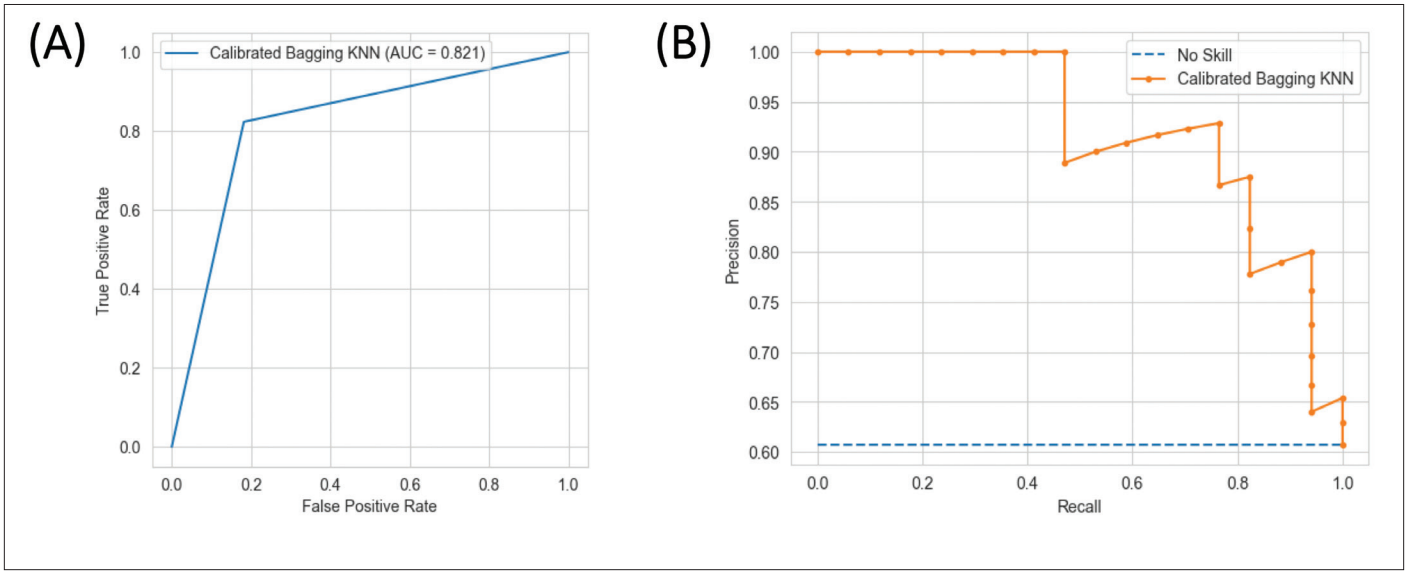

Figure 3. (A) Received operating characteristics curve for calibrated bagged K-nearest neighbor (KNN). (B) Precisions recall curve for calibrated bagged KNN.

REFERENCES

1. Chung E, Katz DJ, Love C. Adult male stress and urge urinary incontinence: A review of pathophysiology and treatment strategies for voiding dysfunction in men. *Aust Fam Physician* 2017;46:661-6.
2. Rehder P, Haab F, Cornu JN, et al. Treatment of postprostatectomy male urinary incontinence with the transobturator retroluminal repositioning sling suspension: 3-year followup. *Eur Urol* 2012;62:140-5. https://doi.org/10.1016/j.eururo.2012.02.038
3. Kowalik CG, DeLong JM, Mourtzinos AP. The advance transobturator male sling for post-prostatectomy incontinence: Subjective and objective outcomes with 3 years followup. *Neurourol Urodyn* 2015;34:251-4. https://doi.org/10.1002/nau.22539
4. Bauer RM, Grabbert MT, Klehr B, et al. 36-month data for the AdVance XP® male sling: Results of a prospective, multicenter study. *BJU Int* 2017;119:626-30. https://doi.org/10.1111/bju.13704
5. Chua ME, Zuckerman J, Mason JB, et al. Long-term success durability of transobturator male sling. *Urology* 2019;133:222-8. https://doi.org/10.1016/j.urology.2019.07.032
6. Schork NJ. Artificial intelligence and personalized medicine. *Cancer Treat Res* 2019;178:265-83. https://doi.org/10.1007/978-3-030-16391-4_11
7. Zuckerman JM, Edwards B, Henderson K, et al. Extended outcomes in the treatment of male stress urinary incontinence with a transobturator sling. *Urology* 2014;83:939-45. https://doi.org/10.1016/j.urology.2013.10.065
8. Maadi M, Akbarzadeh Khorshidi H, et al. A review on human-AI interaction in machine learning and insights for medical applications. *Int J Environ Res Public Health* 2021;18:2121. https://doi.org/10.3390/ijerph18042121
9. Mourmouris P, Tzelves L, Feretzakis G, et al. The use and applicability of machine learning algorithms in predicting the surgical outcome for patients with benign prostatic enlargement. Which model to use? *Arch Ital Urol Androl* 2021;93:418-24. https://doi.org/10.4081/aiua.2021.4.418
10. Ekşi M, Evren İ, Akkaş F, et al. Machine learning algorithms can more efficiently predict biochemical recurrence after robot-assisted radical prostatectomy. *Prostate* 2021;81:913-20. https://doi.org/10.1002/pros.24188
11. Suarez-Ibarrola R, Hein S, Reis G, et al. Current and future applications of machine and deep learning in urology: A review of the literature on urolithiasis, renal cell carcinoma, and bladder and prostate cancer. *World J Urol* 2020;38:2329-47. https://doi.org/10.1007/s00345-019-03000-5
12. Zhang S, Bamakan SMH, Qu Q, et al. Learning for personalized medicine: A comprehensive review from a deep learning perspective. *IEEE Rev Biomed Eng* 2019;12:194-208. https://doi.org/10.1109/RBME.2018.2864254
13. MacEachern SJ, Forkert ND. Machine learning for precision medicine. *Genome* 2021;64:416-25. https://doi.org/10.1139/gen-2020-0131
14. Sun K, Du W, Shi N. A survey of kNN algorithm. *Inf Eng Appl Comput* 2018;1. https://doi.org/10.18063/ieac.v1i1.770
15. Monn MF, Jarvis HV, Gardner TA, et al. Impact of obesity on male urethral sling outcomes. *Ther Adv in Urol* 2020;12. https://doi.org/10.1177/1756287220927997
16. Collado SA, Resel FL, Domínguez-Escrig JL, et al. AdVance/AdVance XP transobturator male slings: Preoperative degree of incontinence as predictor of surgical outcome. *Urology* 2013;81:1034-9. https://doi.org/10.1016/j.urology.2013.01.007
17. Cornu JN, Sèbe P, Ciofu C, et al. Mid-term evaluation of the transobturator male sling for post-prostatectomy incontinence: Focus on prognostic factors. *BJU Int* 2011;108:236-40. https://doi.org/10.1111/j.1464-410X.2010.09765.x

CORRESPONDENCE: Dr. Jin K. Kim, Division of Urology, Hospital for Sick Children, Toronto, ON, Canada; jjk.kim@mail.utoronto.ca