Published in final edited form as:

J Urol. 2010 April ; 183(4): 1568–1572. doi:10.1016/j.juro.2009.12.042.

Variation Among Internet Based Calculators in Predicting Spontaneous Resolution of Vesicoureteral Reflux

Jonathan C. Routh^{*}, Edward M. Gong, Glenn M. Cannon Jr., Richard N. Yu, Patricio C. Gargollo, and Caleb P. Nelson[†]

Department of Urology (JCR, EMG, GMC, RNY, CPN) and Harvard Pediatric Health Services Research Fellowship Program (JCR), Children's Hospital Boston, Boston, Massachusetts, and Department of Urology, University of Texas–Southwestern Medical Center, Dallas, Texas (PCG)

Abstract

Purpose—An increasing number of parents and practitioners use the Internet for health related purposes, and an increasing number of models are available on the Internet for predicting spontaneous resolution rates for children with vesicoureteral reflux. We sought to determine whether currently available Internet based calculators for vesicoureteral reflux resolution produce systematically different results.

Materials and Methods—Following a systematic Internet search we identified 3 Internet based calculators of spontaneous resolution rates for children with vesicoureteral reflux, of which 2 were academic affiliated and 1 was industry affiliated. We generated a random cohort of 100 hypothetical patients with a wide range of clinical characteristics and entered the data on each patient into each calculator. We then compared the results from the calculators in terms of mean predicted resolution probability and number of cases deemed likely to resolve at various cutoff probabilities.

Results—Mean predicted resolution probabilities were 41% and 36% (range 31% to 41%) for the 2 academic affiliated calculators and 33% for the industry affiliated calculator ($p = 0.02$). For some patients the calculators produced markedly different probabilities of spontaneous resolution, in some instances ranging from 24% to 89% for the same patient. At thresholds greater than 5%, 10% and 25% probability of spontaneous resolution the calculators differed significantly regarding whether cases would resolve (all $p < 0.0001$).

Conclusions—Predicted probabilities of spontaneous resolution of vesicoureteral reflux differ significantly among Internet based calculators. For certain patients, particularly those with a lower probability of spontaneous resolution, these differences can significantly influence clinical decision making.

Keywords

forecasting; Internet; models; statistical; prognosis; vesico-ureteral reflux

In the quest to forecast more accurately relevant outcomes of diseases and their treatment clinical researchers have become increasingly sophisticated in the use of prediction models such as logistic regression based nomograms and risk scores, artificial neural networks and

[†]Correspondence: Department of Urology, Children's Hospital Boston, 300 Longwood Ave., HU-359, Boston, Massachusetts 02115 (telephone: 617-355-3776; FAX: 617-730-0474; caleb.nelson@childrens.harvard.edu).

^{*}Supported by Agency for Healthcare Research and Quality Grant T32-HS000063.

recursive decision trees. In pediatric urology these techniques have been applied to vesicoureteral reflux, specifically to predict the probability of its spontaneous resolution.^{1–4}

Several of these models have now been published as Internet based calculators, allowing providers and parents to enter pertinent clinical features of a patient and to determine the probability of spontaneous VUR resolution. The accuracy and consistency of such Internet calculators are of particular relevance to pediatricians and pediatric specialists, since in recent years there has been a dramatic increase in the use of the Internet to disseminate health information. Among adults using the Internet 75% to 85% report having gone online specifically to obtain health information.⁵ A recent poll revealed that 86% of adults who search for information on the Internet believe that the information they find there is reliable.^{5,6} As might be expected, teenagers and young Americans are even more dependent on the Internet as an information source than their elders, with approximately 90% of American teenagers accessing the Internet regularly, often for health related topics.^{6,7} Given the potential clinical ramifications of differing predictions, we determined whether currently available Internet based calculators significantly differ from one another in terms of predicted probabilities for spontaneous resolution of pediatric VUR.

PATIENTS AND METHODS

Identification of Internet Based Calculators

We identified relevant Internet based calculators by performing a systematic search modeled on previously described methods.⁸ Using the search engine Google, we queried the term “vesicoureteral reflux,” sequentially combining this term with each of the terms “spontaneous,” “resolution,” “spontaneous resolution,” “prediction,” “model,” “prediction model,” “calculator,” “Internet based” and “neural network.” We identified 4 Internet based calculators, including 1 produced by Children’s Hospital Boston (available at www.childrenshospital.org/vurcalculator), 2 produced by the University of Iowa (available at www.urocomp.net) and 1 produced by Q-Med Scandinavia (available at www.deflux.com). The 2 available calculators from Iowa differ primarily in that 1 incorporates nuclear renal scan data and the other does not.

Since many patients with newly diagnosed VUR will not undergo nuclear renal scan (at least not at initial diagnosis), and since the other 2 calculators do not incorporate these data, we chose to use only the Iowa calculator that does not require renal scan data. Because the Iowa calculator provides the probability of reflux resolution only at 2 years, we recorded only the 2-year probability for all 3 Internet based calculators. Similarly while the CHB and Q-Med calculators provide a resolution probability, the Iowa calculator provides the odds of resolution, from which we calculated the resolution probability. For patients with bilateral VUR grade and phase of onset were assumed to be equal on each side. Parameters considered by each calculator are listed in the Appendix.

Generation of Random Patient Sample and Statistical Analysis

We generated a random sample of 100 hypothetical patients with VUR using SAS®, version 9.1. For each patient we randomly selected several clinical variables, including age (integer values ranging from 0 to 11 years), VUR grade (I to V), gender, presentation (afebrile UTI, febrile UTI, prenatal hydronephrosis or sibling screen), laterality (unilateral or bilateral), cystogram demonstrated phase of VUR onset (bladder filling or voiding), ureteral duplication (present or absent), dysfunctional voiding (present or absent) and percent bladder volume (50% to 150% in 10% increments). We used a previously published definition of percent bladder volume, ie maximum bladder volume for a given patient

normalized for age expected bladder capacity.³ Overall, 9 clinical variables with up to 12 ordinal levels were randomly assigned for a total of 84,480 possible permutations.

To represent better a typical patient distribution, we over sampled patient age to simulate a Poisson distribution (median age 2 years), VUR grade to simulate a Gaussian distribution (median grade III with approximately a 1:2:4:2:1 ratio) and percent predicted bladder volume also to simulate a Gaussian distribution (median volume 100% of predicted). Additionally we programmed our randomization scheme to reflect a female-to-male gender ratio of 3:1 and a single-to-duplicated ureter ratio of 3:1. All parameter ranges were selected based on typical values for the patient population at our institution. Following randomization, the sample was screened to ensure that no variable combinations were duplicated and that each randomly generated patient was unique.

We then entered data for each patient into each of the 3 Internet based calculators and recorded the probability of spontaneous VUR resolution. For patients with unilateral VUR the left ureter was always assigned as the refluxing ureter. When data exceeded the maximum possible setting for a calculator the maximum possible setting was selected, eg if a patient was 11 years old but the calculator maximum age was 10, an age of 10 was entered.

We determined mean probabilities of spontaneous resolution across the cohort and 95% confidence intervals for each Internet based calculator. We compared the mean predicted probabilities of spontaneous resolution derived from each calculator using 1-way ANOVA. To determine further the clinical relevance of differences in predicted resolution probabilities, we compared the calculators in terms of ability to discriminate between cases that were likely to resolve spontaneously and those that were not. Specifically we compared the number of hypothetical patients whose predicted resolution probabilities were above or below a given threshold or cutoff value. Because no consensus exists regarding the most appropriate threshold value, we varied this threshold across a wide range (5%, 10%, 25%, 33%, 50% and 75% probability of resolution at 2 years). The numbers of cases above or below the threshold value were then compared between calculators using Fisher's exact test. All tests were 2-sided and p values of 0.05 or less were considered significant.

RESULTS

For all 100 hypothetical patients the mean predicted resolution probability was 40.7% (95% CI 35.7 to 45.7) for the CHB calculator, 36% (31.1 to 41) for the Iowa calculator and 32.5% (27.6 to 37.5) for the Q-Med calculator. These differences were statistically significant ($p < 0.02$).

In an attempt to define better whether these statistically significant differences were clinically meaningful we compared the predicted probabilities of spontaneous VUR resolution for several representative patients (table 1). For certain patients the calculators produced markedly different probabilities of spontaneous resolution, in some instances ranging from 24% to 89% for the same patient. Generally the CHB calculator produced resolution probabilities that were 8% higher (95% CI 1.1 to 15.2) than those produced by the Q-Med calculator and 5% higher (-2.4 to 11.7) than those produced by the Iowa calculator. Meanwhile, the probabilities provided by the Iowa calculator were 4% higher (95% CI -3.5 to 10.5) than those provided by the Q-Med calculator.

We then compared the calculators in terms of their discriminatory ability at various probability thresholds, ie number of patients whose probability of spontaneous resolution was below a particular cutoff value (table 2). Assuming a threshold of greater than 10% probability of spontaneous resolution at 2 years after diagnosis, the CHB calculator

predicted that all 100 patients would be above the 10% threshold, the Q-Med calculator predicted that 82 patients would be above the threshold and the Iowa calculator estimated that only 75 patients would be above the threshold ($p < 0.0001$). At a threshold of greater than 25% probability of resolution at 2 years the CHB calculator predicted that 84 patients would be above the threshold, while the Iowa and Q-Med calculators predicted that only 57 and 55 patients, respectively, would be above the threshold ($p < 0.0001$). In contrast at a cutoff value of 50% probability of spontaneous resolution the 3 calculators predicted similar numbers of patients above the threshold (78, 79 and 85, respectively, $p = 0.4$).

DISCUSSION

In this analysis we found that the probability of reflux resolution systematically differed among 3 Internet based calculators. For certain patients these calculators may produce predictions ranging from a low to a high probability of resolution despite identical patient characteristics. For example 1 patient was calculated to have a 2-year resolution probability ranging from 24% to 89%, while another had a probability of 7% to 48%. Perhaps more importantly the 3 calculators produced widely divergent discriminatory abilities when a variety of threshold cutoff values were used. For example if a 2-year resolution probability of 25% was determined to be a clinically meaningful threshold, 1 calculator would predict that 16 of 100 cases would be below this threshold or unlikely to resolve, while another predicted that 43 of these same cases would be unlikely to resolve. Interestingly these differences were present for each VUR grade as well.

While an experienced clinician who is well versed in the current VUR literature might be expected to evaluate these differences critically, someone less familiar with pediatric VUR may be unable to do so. In particular it seems doubtful that the average parent of a child with VUR could be expected to interpret reliably the differing predictions of the calculators we used. This issue is of particular relevance in terms of surgical decision making, given that parent preferences are widely cited as a motivating factor behind early surgical intervention.^{9,10} Parents who are presented with a 24% probability of resolution might choose a different (and possibly more aggressive) treatment algorithm than parents of a similar child who are given an 89% probability of resolution. Similarly a 6% resolution probability might lead a clinician to recommend a different management strategy than she/he would choose for a child with a 48% chance of spontaneous resolution. However, in both of these instances the child in question is exactly the same. The only difference is the methodology behind the calculator estimates. As such, it is clear that these calculators, which are freely accessible to parents, pediatricians and specialists, could reasonably be expected to exert an influence on management choices purely based on which calculator happens to be used.

Further complicating this issue is the high level of variability in patient understanding of risks and probabilities. Patients and families often have a poor understanding of quantitative risk information, although their understanding can be improved by incorporating uncertainty into the risk estimate or by the use of graphic presentation methods.^{11,12} Specifically it is well documented that patients commonly underestimate or overestimate medical risks, and multiple studies have shown that patient understanding of risk estimates can be greatly influenced by the specific wording used to frame a risk estimate.^{11–15}

Of particular relevance to probability calculators such as the ones we used is the finding that patients tend to be unrealistically optimistic in their interpretation of risk data.^{13,15} Although one might assume that the use of personalized risk estimates would be easier for patients to understand than generalized risk estimates, little data exist to support this idea. Similarly there is little evidence that patients receiving personalized risk data make better, or better informed, decisions. However, patients receiving risk information of any kind tend to have

higher satisfaction with their medical decisions.¹⁴ Also, multiple studies have revealed that risk estimates are a useful component of decision aids, which in turn can improve patient decision making and decision satisfaction. In this sense these calculators may serve an important and useful role for patients and families, despite their variability.

Predictive models have the potential to be useful adjuncts to clinical management. However, for clinicians to use them effectively a basic understanding of how these models function is critical. In this case we found significant variation between 3 predictive models using the identical set of randomly generated hypothetical patients. All modeling techniques have their own unique methodological advantages and disadvantages. The precision and accuracy of these models depend on the premise that the model technique being used is the most appropriate for that particular data set.

Interestingly each of the 3 calculators we tested uses a different predictive method. The CHB calculator uses a logistic regression model, the Iowa calculator uses a neural network model, and the Q-Med calculator uses a nomogram derived from a previous systematic review and meta-analysis.¹ It is important to realize that although each method has specific advantages compared to the others, none of these methods is inherently superior to any other for all situations. Generally use of model selection techniques is recommended to identify a best fitting model to make the best statistical inference possible for a given population.^{16–18} Without access to the original data from which each prediction model was derived it is impossible to make a post hoc determination regarding whether a given model is the best choice for a given population. Nevertheless, model choice and type can significantly impact how well a model functions for a particular population, as our results demonstrate.

Perhaps more importantly the generalizability of the model depends on the premise that the patients on whom the model is based are representative of the population to whom the model will be applied. For this reason it is difficult to overestimate the value of external validation for such models. It is reassuring that at least 2 of the 3 calculators we studied have undergone or are currently undergoing external validation (H. T. Nguyen, personal communication).¹⁹

Lastly it is noteworthy that this study is not intended to be, and should not be interpreted as, a methodological critique of 1 or more of the calculators we investigated, since all appear to use reasonable methods drawn from appropriately performed studies. Similarly this study was not intended to validate or otherwise judge the relative accuracy of 1 calculator vs another. Because we used a randomly generated cohort of hypothetical subjects instead of actual patients, it is impossible to assess the accuracy of any model. Rather, the goal of this study was simply to investigate the variation in results that would be obtained by a typical family or a typical clinician seeking to use a tool encountered during a cursory Internet search. While these data clearly show that different Internet based calculators will produce different probabilities of spontaneous resolution for a given child, these differences should be assumed to reflect methodological variation rather than quality variation. Unfortunately whatever the cause of these variations, their net effect is the same—for some patients using a particular calculator rather than another will result in clinically significant differences in the predicted chance of spontaneous resolution, which may in turn lead to a different treatment regimen than they might otherwise have chosen.

CONCLUSIONS

Predicted probabilities of spontaneous VUR resolution differ significantly among Internet based calculators. For certain patients, particularly those with a lower probability of

spontaneous resolution, these differences have the potential to influence clinical decision making significantly.

Abbreviations and Acronyms

CHB	Children's Hospital Boston
UTI	urinary tract infection
VUR	vesicoureteral reflux

References

1. Elder JS, Peters CA, Arant BS Jr, et al. Pediatric Vesicoureteral Reflux Guidelines Panel summary report on the management of primary vesicoureteral reflux in children. *J Urol* 1997;157:1846. [PubMed: 9112544]
2. Estrada CR Jr, Passerotti CC, Graham DA, et al. Nomograms for predicting annual resolution rate of primary vesicoureteral reflux: results from 2,462 children. *J Urol* 2009;182:1535. [PubMed: 19683762]
3. Knudson MJ, Austin JC, Wald M, et al. Computational model for predicting the chance of early resolution in children with vesicoureteral reflux. *J Urol* 2007;178:1824. [PubMed: 17707424]
4. Nepple KG, Knudson MJ, Austin JC, et al. Adding renal scan data improves the accuracy of a computational model to predict vesicoureteral reflux resolution. *J Urol*, part 2 2008;180:1648.
5. Fox, S. The engaged e-patient population. [Accessed July 8, 2009]. Available at <http://www.pewinternet.org/Reports/2008/The-Engaged-Epatient-Population.aspx>
6. Macgill, A. Parent and teen Internet use. [Accessed July 8, 2009]. Available at <http://www.pewinternet.org/Reports/2007/Parent-and-Teen-Internet-Use.aspx>
7. Skinner H, Biscope S, Poland B, et al. How adolescents use technology for health information: implications for health professionals from focus group studies. *J Med Internet Res* 2003;5:e32. [PubMed: 14713660]
8. Routh JC, Gong EM, Nelson CP. Pediatric urology and the Internet— does an uncommon topic decrease content quality? *J Urol* 2009;182:1569. [PubMed: 19683756]
9. Capozza N, Lais A, Matarazzo E, et al. Treatment of vesico-ureteric reflux: a new algorithm based on parental preference. *BJU Int* 2003;92:285. [PubMed: 12887485]
10. Ogan K, Pohl HG, Carlson D, et al. Parental preferences in the management of vesicoureteral reflux. *J Urol* 2001;166:240. [PubMed: 11435878]
11. Edwards A, Elwyn G, Mulley A. Explaining risks: turning numerical data into meaningful pictures. *BMJ* 2002;324:827. [PubMed: 11934777]
12. Johnson BB, Slovic P. Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. *Risk Anal* 1995;15:485. [PubMed: 7480948]
13. Edwards AG, Evans R, Dundon J, et al. Personalised risk communication for informed decision making about taking screening tests. *Cochrane Database Syst Rev* 2006;4:CD001865. [PubMed: 17054144]
14. Davis RE, Dolan G, Thomas S, et al. Exploring doctor and patient views about risk communication and shared decision-making in the consultation. *Health Expect* 2003;6:198. [PubMed: 12940793]
15. O'Connor AM, Stacey D, Rovner D, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev* 2001;3:CD001431. [PubMed: 11686990]
16. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361. [PubMed: 8668867]
17. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49:1225. [PubMed: 8892489]
18. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;130:515. [PubMed: 10075620]

19. Shiraishi K, Matsuyama H, Nepple KG, et al. Validation of a prognostic calculator for prediction of early vesicoureteral reflux resolution in children. *J Urol* 2009;182:687. [PubMed: 19539324]

Table 1

Predicted 2-year spontaneous resolution rates for 10 hypothetical patients

Pt Age—Gender	Presentation	Laterality	% Bladder Vol	VUR Grade	Reflux Onset	Ureteral Anatomy	Voiding Dysfunction	% CHB Calculator	% Q-Med Calculator	% University of Iowa Calculator
0—F	Febrile UTI	Bilat	110	II	Filling	Duplication	Absent	36	48	7
1—M	Febrile UTI	Bilat	100	I	Voiding	Duplication	Absent	47	63	50
1—F	Febrile UTI	Bilat	150	IV	Filling	Single	Absent	21	6	17
1—F	Afebrile UTI	Unilat	80	II	Voiding	Duplication	Absent	38	48	11
2—F	Afebrile UTI	Bilat	100	III	Filling	Single	Absent	32	24	89
3—F	Sibling	Bilat	90	III	Filling	Single	Present	37	14	25
4—M	Prenatal hydronephrosis	Bilat	110	II	Filling	Duplication	Present	45	48	7
6—F	Febrile UTI	Bilat	60	III	Voiding	Single	Absent	37	5	20
7—F	Febrile UTI	Unilat	90	IV	Filling	Duplication	Absent	19	30	2
7—F	Sibling	Unilat	140	III	Voiding	Single	Present	48	21	67

Table 2

Patients above and below 2-year resolution probability thresholds

2-Yr Resolution Probability Threshold	CHB Calculator		University of Iowa Calculator		Q-Med Calculator		p Value* (Fisher's exact test)
	% Above Threshold	% Below Threshold	% Above Threshold	% Below Threshold	% Above Threshold	% Below Threshold	
Greater than 5%	100	0	87	13	100	0	< 0.0001
Greater than 10%	100	0	75	25	82	18	< 0.0001
Greater than 25%	84	16	57	43	55	45	< 0.0001
Greater than 33%	61	39	56	44	45	55	0.07
Greater than 50%	22	78	21	79	15	85	0.4
Greater than 75%	0	100	6	94	0	100	0.004

APPENDIX

Parameters Considered by Each Internet Based Calculator

Children's Hospital Boston Calculator	University of Iowa Calculator	Q-Med Calculator
Age	Age	Age
VUR grade	VUR grade	VUR grade
Unilateral/bilateral	Unilateral/bilateral	Unilateral/bilateral
Gender	Gender	
Presentation	Presentation	
Single/duplicated ureter	Single/duplicated ureter	
	Bladder capacity	
	Reflux onset (filling/voiding)	
	Voiding dysfunction	