ANNALS OF FAMILY MEDICINE

Online Supplementary Material

Knottnerus BJ, Geerlings SE, Moll van Charante EP, ter Riet G. Towards a simple diagnostic index for acute uncomplicated urinary tract infections. *Ann Fam Med*. 2013;11(5):442-451.

http://www.annfammed.org/content/11/5/442

Supplemental Appendix. Detailed Description of Variable Selection Method

All analyses were performed in Stata/SE 10.1 (StataCorp LP).

Multiple Imputation

Missing values were imputed using multiple imputation by chained equations[1,2] (using Stata's *ice* command), creating 45 complete data sets. For multiple imputation, 58 variables were used, including possible interactions (see below).

After multiple imputation, patients with missing values for the dependent variable were dropped.[3]

Variable Selection

To avoid overfitting by using too many candidate predictors, the 22 most relevant variables (based on literature and clinical usefulness) were used for the analysis, including 2 interaction variables (see below).

Logistic regression with bootstrapped backward elimination was used to derive a parsimonious diagnostic index. The result of the urine culture was the binary dependent variable, $\geq 10^3$ colony-forming units (CFUs) of a single uropathogen per milliliter (mL) being defined as a positive culture according to international guidelines.[4] For nominal predictors, categorical coding was used to create binary dummy variables. For ordinal predictors, ordinal coding was used. For the only continuous variable "age," fractional polynomials were used to select the best fitting function, but only the linear function was selected.[5]

We selected 5 different sets of variables, based on clinical practice:

1. History = only patient history questions
2. History + dipstick = variables selected in the model above plus + urine dipstick variables
3. History + dipstick + sediment = variables selected in 2 models above + urinary sediment variables
4. History + dipstick + dipslide = variables selected in first 2 models + dipslide variables
5. History + dipstick + sediment + dipslide = all selected variables

For each of the 5 sets of variables, the following procedure was performed:

From each of the 45 imputed data sets, 600 bootstrap samples were generated. In each bootstrap sample, variables were eliminated backwards at a significance level of .05 using Stata's *mfpboot* command.

For each variable, a bootstrap inclusion fraction (BIF) was obtained in each imputed data set. Next, the mean BIF across all 45 imputed data sets was calculated for each variable. Variables with a mean BIF ≥66.67% were retained in the final models.

After having obtained the final set of variables, parameter-wise shrinkage of their regression coefficients was performed in each imputed data set to correct for possible overoptimism,[6] using Stata's *cval* command with parameter-wise option and 10-fold cross-validation. The resulting 45 shrunk regression coefficients and their standard errors were averaged. The averaged regression coefficients were used to compose risk scores as described by Sullivan et al.[7]

Variables

The 58 variables below were used for multiple imputation. The 22 variables in italics were used for the logistic regression analysis.

History

Name of health centre / GP
Age in years
Ethnicity
General health according to patient
Marital status
Duration of symptoms in days
Lower abdominal pain
Lower abdominal pain when?
Back pain
Back pain when?
Blood in urine according to patient
Burning sensation during micturition
Not able to empty bladder completely
False urge to urinate
More frequent micturition than usually
Urine incontinence: frequency
Urine incontinence: type
Micturition of smaller amounts than usually
Pain during micturition
Bad smell of urine
Urge to urinate hard to control
Vaginal discharge
Vaginal irritation or itching
Currently menstruating?
Sexual activity past week (n)
Voiding directly after sex
Patient thinks she has a UTI?
Bother at social activities
Bother at work/school
Days willing to delay antibiotic
Last menstruation >1 year ago?
Diabetes mellitus (according to patient)
Any first-grade relative >2x/year UTI
UTIs in past year according to patient (n)
≥1 *UTI ever diagnosed according to patient*
Periods of UTI symptoms without consulting physician
≥4 hours no urinating until collection of urine sample

Urine collection

Midstream urine
Minutes from urine collection until placing into fridge

Dipstick

Blood
Glucose
Leucocyte esterase
Nitrite positive
Protein

Sediment

Bacteria / HPF
Leucocytes / HPF
Squamous epithelial cells / HPF

Dipslide

Cystine Lactose Electrolyte Deficient (CLED) medium
MacConkey medium

Culture

≥10^3 CFU of a single uropathogen per milliliter (mL)[a]

Interactions

Patient thinks she has a UTI? × UTIs in past year according to patient (n)
Patient thinks she has a UTI? × ≥1 UTI ever diagnosed according to patient
Patient thinks she has a UTI? × Periods of UTI symptoms without consulting physician
Patient thinks she has a UTI? × ≥10^3 CFU/mL of uropathogen in culture
UTIs in past year according to patient (n)× ≥10^3 CFU/mL of uropathogen in culture
≥1 UTI ever diagnosed according to patient × ≥10^3 CFU/mL of uropathogen in culture
Periods of UTI symptoms without consulting physician × ≥10^3 CFU/mL of uropathogen in culture

CFU = colony-forming units; HPF = high-power field; UTI = urinary tract infection.
[a] Dependent variable.

References

1. van Buuren S: Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.* 2007;16(3):219-242.
2. White IR, Royston P, Wood AM: Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med.* 2011;30(4):377-399.
3. von Hippel PT: Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociol Methodol.* 2007;37:83-117.
4. European urinalysis guidelines. *Scand J Clin Lab Invest Suppl.* 2000;231(Suppl 231):1-86.
5. Royston P, Sauerbrei W. *Multivariable Model-Building.* New York NY: Wiley; 2008.
6. Verweij PJ, van Houwelingen HC: Cross-validation in survival analysis. *Stat Med.* 1993;12(24):2305-2314.
7. Sullivan LM, Massaro JM, D'Agostino RB Sr.: Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Stat Med.* 2004;23(10):1631-1660.