Supplementary material

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic			Checklist Item	
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	✓
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	✓
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	✓
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	✓
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	✓
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	✓
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	✓
	5b	D;V	Describe eligibility criteria for participants.	✓
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	✓
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	✓
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	✓
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	✓
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	✓
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	✓
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	✓
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	✓
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	✓
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	✓
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	✓
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	✓
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	✓
	15b	D	Explain how to the use the prediction model.	✓
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	✓
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as non-representative sample, few events per predictor, missing data).	✓
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	✓
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	✓
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	✓
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	✓
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	✓

Missing data

We are aware of the potential bias introduced by complete case analysis. In that vein, two options were explored in an attempt to impute missing data.

1) We explored multiple imputation by chained equations (MICE). This is a common method of multiple imputation. In Stata 16, multilevel logistic regression is available using MICE (command used was `mi estimate: meqrlogit`), however the computing time for one step of the backward elimination process with only 3 imputed datasets using mixed effect multivariable regression, took over 30 minutes. Ideally to reduce bias with imputation we would use at least 10 imputed datasets, which would require considerably longer. With a minimum of 5 steps in backward elimination, the computing time will be at least fivefold. Furthermore, the bootstrap validation process requires the same model development method as the original model. So, with every model elimination step required in addition to 200 repetitions of each model, we would re-run the multilevel logistic regression at least 1000 times. We felt the computing time for this with MICE was unjustifiable.

2) As we have a large dataset, we favoured full information maximum likelihood (FIML), based on maximum likelihood function used in logistic regression. In Stata 16 this is incorporated into the structural equation modelling (SEM) option. However, FIML is only available using SEM, and multilevel logistic regression requires use of generalised SEM, which does not support FIML.

Due to our large dataset and the low proportion of missing data (**total of 7.96% for final model**), although it is regrettable to lose any data in analysis, we felt that complete case analysis would be sufficient to arrive at a predictive model representative of the target population.

Statistical analysis and model development

We chose not to split the data into a development and a validation cohort firstly due to the low number of events in the rarer cancers which would limit the number of candidate predictors in the secondary models, and secondly as we intend to perform a separate study to externally validate the prediction model as recommended by the TRIPOD statement.

Interaction terms
Two-way interaction terms were tested for all fixed effect variables within the full model (all candidate predictors) to check for significant interactions (p value less than 0.05). We decided on this approach as there were novel predictors for which we were unsure of their interactions. We used clinical reasoning to select the most appropriate significant interaction terms and included these in the backward stepwise selection for development of the secondary predictive models. Interaction terms were dropped in the backward stepwise selection if they failed to reach significance or the variable in the interaction term had been dropped.

Model development
We decided on fitting the full multivariable and performing backward stepwise elimination as we were exploring new candidate predictors. Crucially though we judged the clinical importance of each predictor before eliminating it and kept it in the model if we felt it was too important to drop. Below is a list of the clinical importance and practical use that we attached to each variable before the selection process. Clinical importance was judged using evidence from the literature and clinical knowledge. Practical use was determined by how easily collected the information is to the clinician, and its availability within patient records. Predictors were used at patient level only and we did not include any predictors at centre or country level. The least significant variable was

dropped one at a time in the backward selection until all variables in the model reached significance with a p value less than 0.05.

Table to summarise the clinical importance and practical use of each candidate predictor

Predictors for all cancers	**Clinical (predictive) importance**	**Practical use**
Type of haematuria	Important	High
Age	Important	High
Sex	Important	High
Smoking history	Important	High
Family history of cancer	Important	Low
Use of catheter	Less important	High
High risk occupation	Less important	Low
High risk travel	Less important	Low
High risk medications	Less important	Low
Pelvic radiotherapy history	Less important	Medium
Flank pain	Less important	Medium
Urinary tract infection	Unknown	High
Ethnicity	Unknown	High
Anticoagulation	Unknown	High
Dysuria or suprapubic pain	Unknown	Medium
Any lower urinary tract symptoms	Unknown	Medium

SUPPLEMENTARY CONTENTS:

Supplementary Table 1: Cancer classification

Case definition category		Type of cancer			
		Bladder	**Renal**	**Upper tract urothelial cancer**	**Prostate**
Cancer positive	1. Histological evidence for cancer	Histologically confirmed bladder cancer (as defined by the WHO classification of tumours) attained by biopsy, TURBT or cystectomy	Histologically confirmed RCC attained by biopsy or nephrectomy (partial or radical)	Histologically (biopsy or nephroureterectomy) or cytologically (ureteroscopic brushings or CTU + positive urine cytology) confirmed UTUC	Histologically confirmed prostate cancer (biopsy or prostatectomy)
	2. Clinical evidence for cancer	Individual case review in the absence of histology: Biopsy/histology not performed but after individual multidisciplinary case review it was felt highly likely that flexible/rigid cystoscopy findings and/or imaging represented a positive finding of bladder cancer	Contrast CT (any IV contrast, urogram or not) or MRI confirmed renal carcinoma. Bosniak 3 and above cysts will be considered positive in the absence of histology	(i) CT urogram (excretory urogram phase) positive for UTUC unless followed by a negative confirmatory test (after individual multidisciplinary case review) *or* (ii) In the absence of histology visual inspection of tumour in ureter from cystoscopy or ureteroscopy (after individual multidisciplinary case review)	Patient considered to have a clinical diagnosis of prostate cancer (after multidisciplinary case review)
Cancer negative	3. Clinical evidence not sufficient to be determined cancer	An abnormality on flexible cystoscopy or imaging with a lower likelihood of being cancer than in category 2 and would not meet a clinical threshold for a diagnosis of cancer as per multidisciplinary case review. Confirmatory tests are not planned.	Bosniak 2f cysts or other equivocal renal parenchymal masses (e.g. fat poor AML vs RCC with repeat imaging planned)	An abnormality on CT where biopsy/washings were attempted but histological result was equivocal, and the final treatment plan is to monitor or for no intervention	Not relevant for prostate cancer
	4. Negative investigations for cancer	Cystoscopy, imaging, cytology and any histology are negative for bladder cancer	Imaging and any histology are negative for RCC	Imaging, cytology and any histology are negative for UTUC	No prostate cancer detected on histology.

Supplementary Table 2: Variables and number of patients included in each analysis of the backward stepwise elimination for bladder cancer model development

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of patients in model/Total of 10282 (%)	9034 (87.9%)	9102 (88.5%)	9334 (90.8%)	9400 (91.4%)	9407 (91.5%)	9464 (92.0%)
Number of bladder cancers in model (%)	1626 (18.0%)	1637 (18.0%)	1663 (17.8%)	1669 (17.8%)	1669 (17.7%)	1679 (17.7%)
Fixed Variables	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria
	Male	Male	Male	Male	Male	Male
	Age	Age	Age	Age	Age	Age
	Ex-smoker	Ex-smoker	Ex-smoker	Ex-smoker	Ex-smoker	Ex-smoker
	Smoker	Smoker	Smoker	Smoker	Smoker	Smoker
	Family history of urothelial cancer	Family history of urothelial cancer	Family history of urothelial cancer	Family history of urothelial cancer	Family history of urothelial cancer	Family history of urothelial cancer
	Previous investigation for haematuria	Previous investigation for haematuria	Previous investigation for haematuria	Previous investigation for haematuria	Previous investigation for haematuria	Previous investigation for haematuria
	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI
	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI
	Catheter use	Catheter use	Catheter use	Catheter use	Catheter use	Catheter use
	Pelvic radiotherapy	Pelvic radiotherapy	Pelvic radiotherapy	Pelvic radiotherapy	Pelvic radiotherapy	Pelvic radiotherapy
	Anticoagulation	Anticoagulation	Anticoagulation	Anticoagulation	Anticoagulation	Anticoagulation
	Dysuria	Dysuria	Dysuria	Dysuria	Dysuria	Dysuria
	High risk occupation	High risk occupation	High risk occupation	High risk occupation	High risk occupation	
	Storage LUTs	Storage LUTs	Storage LUTs	Storage LUTs	Storage LUTs	
	Mixed LUTs	Mixed LUTs	Mixed LUTs	Mixed LUTs		
	High risk medication	High risk medication	High risk medication	High risk medication		
	Ethnicity	Ethnicity	Ethnicity			
	High risk travel	High risk travel				
	Voiding LUTs					
Interaction terms	Visible haematuria and Male	Visible haematuria and Male	Visible haematuria and Male	Visible haematuria and Male	Visible haematuria and Male	Visible haematuria and Male
	Visible haematuria and age	Visible haematuria and age	Visible haematuria and age	Visible haematuria and age	Visible haematuria and age	Visible haematuria and age
	Anticoagulation and age	Anticoagulation and age	Anticoagulation and age	Anticoagulation and age	Anticoagulation and age	Anticoagulation and age
	Storage LUTs and dysuria	Storage LUTs and dysuria	Storage LUTs and dysuria	Storage LUTs and dysuria	Storage LUTs and dysuria	
	Mixed LUTs and Male	Mixed LUTs and Male	Mixed LUTs and Male	Mixed LUTs and Male		
	Voiding LUTs & Male					
Random variables	Country	Country	Country	Country	Country	Country
	Centre	Centre	Centre	Centre	Centre	Centre

Supplementary Table 3: Prediction model for bladder cancer using mixed effects multivariable logistic regression

Variable	Coefficient	Odds ratio	95% Confidence interval	P value
Non-visible haematuria		1.00		
Visible haematuria	1.97	7.19	5.03 – 10.3	<0.001
Female		1.00		
Male	0.74	2.10	1.43 – 3.09	<0.001
Age (years)	0.07	1.07	1.06 – 1.09	<0.001
Age per five-year difference	0.17	1.19	1.12–1.26	<0.001
Never smoked		1.00		
Ex-smoker	0.77	2.15	1.85 – 2.52	<0.001
Current smoker	1.11	3.05	2.56 – 3.63	<0.001
Family history of urothelial cancer				
No		1.00		
Yes	0.62	1.86	1.25 – 2.78	0.001
Previous haematuria investigation				
No		1.00		
Yes	-0.80	0.45	0.35 – 0.58	<0.001
Dysuria/suprapubic pain				
No		1.00		
Yes	-0.24	0.78	0.66–0.93	0.006
Anticoagulation				
No		1.00		
Yes	-0.21	0.81	0.67–0.98	0.031
UTI history				
None		1.00		
Single	-0.68	0.51	0.40 – 0.65	<0.001
Recurrent	-0.70	0.50	0.37 – 0.66	<0.001
Catheter use				
No		1.00		
Yes	-1.53	0.22	0.14 – 0.33	<0.001
Pelvic radiotherapy history				
No		1.00		
Yes	--0.56	0.57	0.35 – 0.92	0.022
Interaction terms				
Visible haematuria & Male	-0.84	0.43	0.28 – 0.65	<0.001
Visible haematuria & Age	-0.02	0.98	0.96 – 0.99	0.010
Age & Anticoagulation	-0.01	0.99	0.97 – 1.00	0.037
Intercept	-3.04			
Inter-country variance	0.83		0.39 – 1.76	
Inter-centre variance	0.35		0.23 – 0.55	
Intraclass correlation for country	0.18		0.10 – 0.33	
Intraclass correlation for centre	0.26		0.17 – 0.38	

Number of observations in model = 9464. Number of country groups = 26 with a mean of 364 observations per group (min=40, max=4294). Number of centre groups = 110 with a mean of 85.3 observations per group (min=36, max=611). Age has been centred about its mean. Performance of model to predict bladder cancer: AUC = 0.86 (95% CI 0.85 – 0.87). Performance of model to predict all urinary tract cancers: AUC = 0.86 (95% CI 0.85 – 0.87)
AUC = Area Under the Curve for Receiver Operating Characteristics. UTI = Urinary tract infection.

Supplementary Table 4: Variables and number of patients included in each analysis of the backward stepwise elimination for Upper Tract Urothelial Cancer (UTUC) model development

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of patients in model/Total of 10282 (%)	9493 (92.3%)	9808 (95.4%)	10004 (97.3%)	10025 (97.5%)	10031 (97.6%)	10052 (97.8%)
Number of UTUC cancers in model	118 (1.24%)	120 (1.22%)	122 (1.22%)	122 (1.22%)	122 (1.22%)	122 (1.21%)
Fixed Variables	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria
	Age	Age	Age	Age	Age	Age
	Flank pain	Flank pain	Flank pain	Flank pain	Flank pain	Flank pain
	Smoker	Smoker	Smoker	Smoker	Smoker	Smoker
	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	
	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	Recurrent episodes of UTI	
	Ex-smoker	Ex-smoker	Ex-smoker	Ex-smoker	Ex-smoker	
	Male	Male	Male	Male		
	Previous haematuria investigation	Previous haematuria investigation	Previous haematuria investigation			
	Family history of urothelial cancer	Family history of urothelial cancer				
	Anticoagulation					
Interaction terms	Visible haematuria and male	Visible haematuria and male	Visible haematuria and male	Visible haematuria and male		
	Visible haematuria and age	Visible haematuria and age				
	Anticoagulation and age					
Random variables	Country	Country	Country	Country	Country	
	Centre	Centre	Centre	Centre	Centre	

Supplementary Table 5: Prediction model for Upper Tract Urothelial Cancer (UTUC) using mixed effects multivariable logistic regression

Variable	Coefficient	Odds ratio	95% Confidence interval	P value
Non-visible haematuria		1.00		
Visible haematuria	1.49	4.46	2.23 – 8.92	<0.001
Age (years)	0.05	1.05	1.03 – 1.06	<0.001
Age per five-year difference	0.23	1.26	1.17 – 1.37	<0.001
Flank pain				
No		1.00		
Yes	1.31	3.73	2.32 – 5.99	<0.001
Current smoker				
No		1.00		
Yes	0.85	2.34	1.53 – 3.57	<0.001
Intercept	-9.50			
Inter-country variance	0.00		-	
Inter-centre variance	0.40		0.15 – 1.07	

Number of observations in model = 10,052. Number of country groups = 26 with a mean of 386.6 observations per group (min=40, max=4630). Number of centre groups = 110 with a mean of 89.8 observations per group (min=1, max=610).
Performance of model 8 to predict UTUC cancer: AUC = 0.82 (95% CI 0.79 – 0.86)
Performance of model 8 to predict all urinary tract cancers: AUC = 0.74 (95% CI 0.73 – 0.75)
AUC = Area Under the Curve for Receiver Operating Characteristics

Supplementary Table 6: Variables and number of patients included in each analysis of the backward stepwise elimination for renal cancer model development

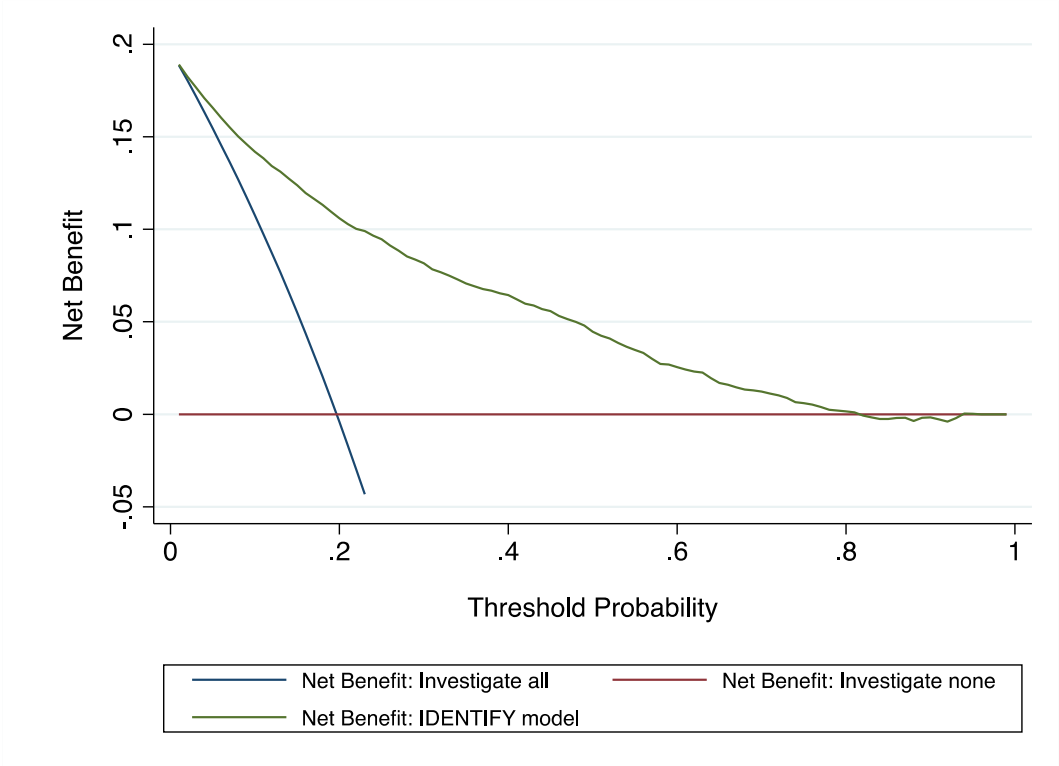
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Number of patients in model/Total of 10282 (%)	9493 (92.3%)	9656 (93.9%)	10005 (97.3%)	10026 (97.5%)	10032 (97.6%)	10053 (97.8%)
Number of renal cancers in model	100 (1.05%)	102 (1.06%)	103 (1.03%)	103 (1.03%)	103 (1.03%)	103 (1.02%)
	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria	Visible haematuria
	Flank pain	Flank pain	Flank pain	Flank pain	Flank pain	Flank pain
Fixed Variables	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	Single episode of UTI	
	Recurrent episodes of UTIs	Recurrent episodes of UTIs	Recurrent episodes of UTIs	Recurrent episodes of UTIs	Recurrent episodes of UTIs	
	Male	Male	Male	Male		
	Previous haematuria investigation	Previous haematuria investigation	Previous haematuria investigation			
	Age	Age	Age			
	Anticoagulation	Anticoagulation				
	Smoker	Smoker				
	Ex-smoker	Ex-smoker				
	Family history of renal cancer					
Interaction terms	Anticoagulation and age	Anticoagulation and age	Anticoagulation and age			
	Visible haematuria and age	Visible haematuria and age	Visible haematuria and age			
	Visible haematuria and male	Visible haematuria and male				
Random variables	Country	Country	Country	Country	Country	Country
	Centre	Centre	Centre	Centre	Centre	Centre

Supplementary Table 7: Prediction model for renal cancer using mixed effects multivariable logistic regression

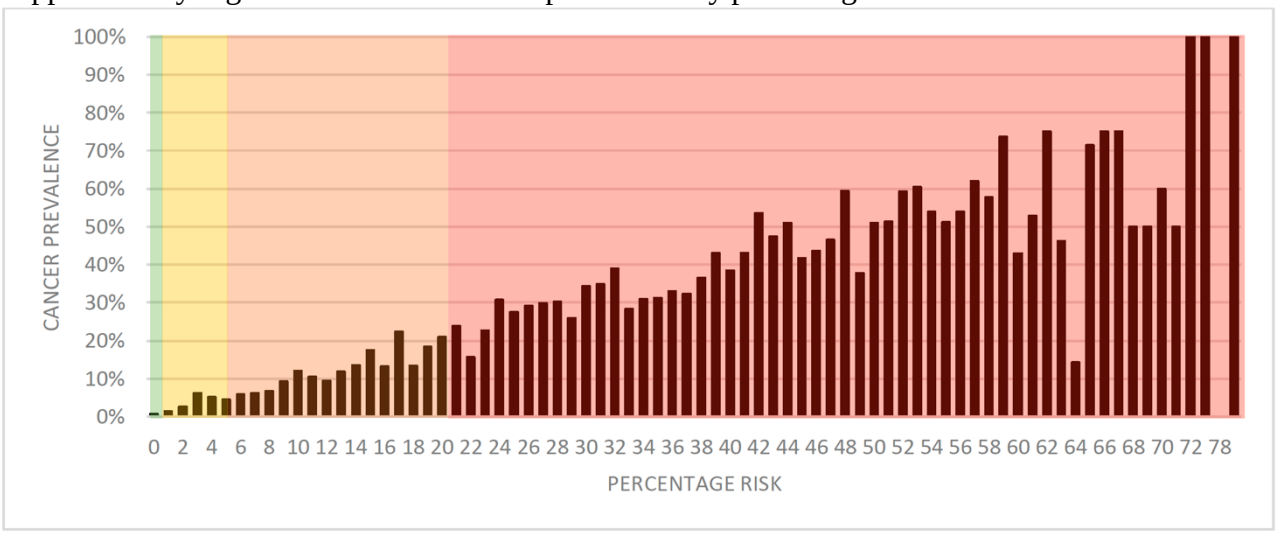
Variable	Coefficient	Odds ratio	95% Confidence interval	P value
Non-visible haematuria		1.00		
Visible haematuria	1.13	3.10	1.72 – 5.59	0.001
Flank pain				
No		1.00		
Yes	0.66	1.93	1.12 – 3.32	0.01
Intercept	-5.62			
Inter-country variance	0.00		-	
Inter-centre variance	0.20		0.03 – 1.24	

Number of observations in model = 10,053. Number of country groups = 26 with a mean of 386.7 observations per group (min=40, max=4631). Number of centre groups = 110 with a mean of 89.8 observations per group (min=1, max=618).
Performance of model 8 to predict renal cancer: AUC = 0.76 (95% CI 0.72 – 0.81)
Performance of model 8 to predict all urinary tract cancers: AUC = 0.65 (95% CI 0.64 – 0.66)
AUC = Area Under the Curve for Receiver Operating Characteristics

Supplementary Figure 1: Decision curve analysis comparing the net benefit of using the IDENTIFY model over investigating all or no patients with haematuria



Supplementary Figure 2: Observed cancer prevalence by percentage risk



Green = Very-low-risk; Yellow = Low-risk; Orange = Intermediate-risk; Red = High-risk