Deep Learning-based Recalibration of the CUETO and EORTC Prediction Tools for Recurrence and Progression of Non–muscle-invasive Bladder Cancer

Mateusz Jobczyk a,b,1, Konrad Stawski c,i,*, Marcin Kaszkowiak c, Francesco Soria e,f, Waldemar Różański a, Shahrokh F. Shariat e,g,h,i,j,k, Wojciech Fendler c,i

a Department of Urology, Copernicus Memorial Hospital, Medical University of Lodz, Lodz, Poland; b Department of Urology, The Hospital Ministry of the Interior and Administration, Lodz, Poland; c Department of Biostatistics and Translational Medicine, Medical University of Lodz, Lodz, Poland; d Department of Urology, Medical University of Silesia, Zabrze, Poland; e Department of Urology, Medical University of Vienna, Vienna, Austria; f Division of Urology, Department of Surgical Sciences, Torino School of Medicine, Turin, Italy; g Karl Landsteiner Institute of Urology and Andrology, Vienna, Austria; h Department of Urology, Weill Cornell Medical College, New York, NY, USA; i Department of Urology, University of Texas Southwestern, Dallas, TX, USA; j Department of Urology, Second Faculty of Medicine, Charles University, Prague, Czech Republic; k Institute for Urology and Reproductive Health, Sechenov University, Moscow, Russia; l Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA

Abstract

Despite being standard tools for decision-making, the European Organisation for Research and Treatment of Cancer (EORTC), European Association of Urology (EAU), and Club Urologico Espanol de Tratamiento Oncologico (CUETO) risk groups provide moderate performance in predicting recurrence-free survival (RFS) and progression-free survival (PFS) in non–muscle-invasive bladder cancer (NMIBC). In this retrospective combined-cohort data-mining study, the training group consisted of 3570 patients with de novo diagnosed NMIBC. Predictors included gender, age, T stage, histopathological grading, tumor burden and diameter, EORTC and CUETO scores, and type of intravesical treatment. The models developed were externally validated using an independent cohort of 322 patients. Models were trained using Cox proportional-hazards deep neural networks (deep learning: DeepSurv) with a proprietary grid search of hyperparameters. For patients treated with surgery and bacillus Calmette-Guérin-treated patients, the models achieved a c-index of 0.650 (95% confidence interval [CI] 0.649–0.650) for RFS and 0.878 (95% CI 0.873–0.874) for PFS in the training group. In the validation group, the c-index was 0.651 (95% CI 0.648–0.654) for RFS and 0.881 (95% CI 0.878–0.885) for PFS. After inclusion of patients treated with mitomycin C, the c-index for RFS models was 0.6415 (95% CI 0.6412–0.6417) for the training group and 0.660 (95% CI 0.657–0.664) for the validation group. Models for PFS achieved a c-index of 0.885 (95% CI 0.885–0.885) for the training set and 0.876 (95% CI 0.873–0.880) for the validation set. Our tool outperformed standard-of-care risk stratification tools and showed no evidence of overfitting. The application is open source and available at https://biostat.umed.pl/deepNMIBC/.

1 These authors contributed equally to this work.
* Corresponding author. Department of Biostatistics and Translational Medicine, Medical University of Lodz, Mazowiecka 15, 92-215 Lodz, Poland. Tel: +48 42 272 53 85. E-mail address: konrad@konsta.com.pl (K. Stawski).

https://doi.org/10.1016/j.euo.2021.05.006
2588-9311/© 2021 European Association of Urology. Published by Elsevier B.V. All rights reserved.
Prediction of recurrence and progression for non–muscle-invasive bladder cancer (NMIBC) remains pivotal for optimal patient care. Many predictive models can help in identifying patients at higher risk of recurrence and progression, but their usefulness has been questioned in recent years [1,2]. In previous work [3] we evaluated the stratification tools most frequently used as developed by the European Organisation for Research and Treatment of Cancer (EORTC), Club Urologico Espanol de Tratamiento Oncologico (CUETO), and European Association of Urology (EAU). Although the EORTC tool provided the best prediction of recurrence and progression for a mixed population of patients, all the tools demonstrated moderate performance in predicting recurrence and progression [3].

Since the publication of the original CUETO [4] and EORTC scales [5], there have been numerous advances in data analysis and artificial intelligence [6]. The aim of the present study was to merge the EORTC and CUETO risk stratification scales for primary NMIBC using state-of-the-art Cox proportional-hazards (CPH) deep neural networks. This multicenter retrospective cohort analysis included patients with de novo diagnosed NMIBC treated with transurethral resection of bladder tumor (TURBT). Using clinical factors from the EORTC and CUETO scales (gender, age, T stage, histopathological grade, number of tumors, tumor diameter, and bacillus Calmette–Guérin [BCG] administration), we applied a modern CPH deep neural network (DeepSurv [7]) for personalized prediction of recurrence or progression to muscle-invasive bladder cancer. Models were developed using our proprietary grid search of hyperparameters (Supplementary Table 1) on a previously published training group (3570 patients) [8] and validated in an independent validation group (322 patients) [3] from our previous single-institution retrospective analysis. The maximum follow-up period was 10 years. Patients in the validation group were treated with or without BCG maintenance, while mitomycin C (MMC) was also administered in selected cases in the training group. To keep the groups comparable, we excluded patients treated with MMC from the first part of the analysis (classical models), leaving 2557 patients for further research. In the second part of the analysis (extended models), model development was repeated after inclusion of MMC-treated patients in the training group. This allowed us to extend the predictive ability of the tool to MMC-treated patients. Owing to the low number of patients with concomitant carcinoma in situ, these were considered jointly with T1 cases.

Harrell’s c index was used to evaluate model performance. Models with the highest sum of c indices for the training and validation sets were selected and compared with reference performance for the validation set provided in our previous publication [3]. The c indices were compared using a two-sample z test for equality of proportions with continuity correction. Detailed inclusion and exclusion criteria, the full methodology, and additional results are provided in the Supplementary material.

The cohort characteristics are provided in Supplementary Table 2. Progression and recurrence events were more frequent in the validation group; however, all EORTC and CUETO recurrence and progression scores were significantly higher in the training group. No difference in progression-free survival (PFS) was noted between the training and validation groups (p = 0.34), but there was a statistically significant difference in recurrence-free survival (RFS; p = 0.0009; Fig. 1) Univariable and multivariable analysis results are provided in Supplementary Table 3. In brief, multivariable Cox regression revealed that older age (hazard ratio [HR] 1.01; p = 0.0216), higher tumor grade (HR 3.67; p < 0.0001), multiplicity of tumors (HR 1.46; p = 0.0036), and large tumor size (HR 1.88; p < 0.0001) have a significant impact on PFS. Older age (HR 1.02; p < 0.0001), T1 stage (HR 0.71; p = 0.0004), higher tumor grade (HR 1.26; p = 0.0003), higher number of tumors (HR 1.40; p < 0.0001), and larger tumor diameter (HR 1.63; p < 0.0001) were significantly associated with RFS.

The c indices for the best classical models were 0.6500 (95% confidence interval [CI] 0.6496–0.6504) for RFS and 0.8738 for PFS (95% CI 0.8733–0.8742). In the validation group, these models achieved a c index of 0.6508 (95% CI 0.6477–0.6538) for RFS and 0.8814 for PFS (95% CI 0.8779–0.8848). Integrated Brier scores are provided in Figure 2. PFS models showed good predictive properties, which, as expected, slowly declined with increasing follow-up. Conversely, for RFS, we observed a sharp decline in the predictive accuracy of the model over early time points, with further gradual improvement.

Similar c indices for RFS prediction for both cohorts suggests there is no overfitting (p = 0.6324). Surprisingly, a significant increase in concordance was noted between the training and validation cohorts in scoring the final model for PFS prediction (p < 0.0001). All of these metrics show improvement in predictions compared to the previously published validation of the original EORTC and CUETO tools for the same validation group [3].

MMC-treated patients were characterized by significantly different survival and clinical features (Supplementary Table 4). After inclusion of MMC-treated patients in model development, the final extended models for PFS achieved a c index of 0.8851 (95% CI 0.8848–0.8854) for the extended training group and 0.8764 (95% CI 0.8729–0.8799) for the validation group. For RFS, the corresponding indices were 0.6579 (95% CI 0.6576–0.6581) and 0.6644 (95% CI 0.6613–0.6676).

Patient summary: We created and validated a new tool to predict recurrence and progression of early-stage bladder cancer. The application uses advanced artificial intelligence to combine state-of-the-art scales, outperforms these scales for prediction, and is freely available online.

© 2021 European Association of Urology. Published by Elsevier B.V. All rights reserved.

Few patients treated with surgery alone have experienced progression before treatment continuation, probably leading to selection bias associated with treatment selection and loss to follow-up. This suggests that the models inflate the survival metrics for high-risk patients initially treated without BCG or MMC, and should be interpreted with caution.

The best-performing models were embedded into a freely accessible web application allowing PFS and RFS estimation for individual patients. The application data, source code, and models were published at https://git.btm.umed.pl/konrad/jobczyk2020-app.

The main advantage of our model is the potential to recognize intricate patterns with a data-mining modeling approach using deep learning. In recent years the same technique has revolutionized almost every industry area, including medicine [9]. New EAU risk groups were recently developed by Sylvester et al [10] using the same predictors. In the case of PFS, the authors reported a Harrell’s bias-corrected (with bootstrapping) c index of 0.80 at 5 years and 0.6674).

Fig. 1 – Kaplan-Meier curves of (A) progression-free survival and (B) recurrence-free survival for the training and validation data sets show significant differences in survival between the groups. An additional group of patients treated with mitomycin C (MMC) was included in this comparison.

Fig. 2 – Prediction error curves showing the relationship between Brier score (y-axis) and time at which it was assessed (x-axis). The Brier score is used to evaluate the accuracy of a predicted survival function at given time points. It is calculated as the squared mean distance between the observed survival status and the expected survival probability. Prediction of progression-free survival in (A) the training group and (B) the validation group and of recurrence-free survival in (C) the training group and (D) the validation group. Integrated Brier scores (IBS) describing the overall model performance are provided for each plot in the header along with the maximum time (as t). The threshold Brier score of 0.25, indicating lack of predictive performance, is marked as a dashed line. The lower the Brier score, the better is the accuracy of the prediction.
0.79 at 10 years of follow-up. As our model achieved better concordance (0.88) for the whole observation period on external validation, we consider our solution to provide a significant improvement.

By embedding our predictive models in a web-based application, we have provided clinicians and researchers with a tool that can estimate RFS and PFS from particular predictors. In addition, inclusion of BCG and MMC receipt in the models means that they can help in assessing the benefit or harm associated with these therapies for patients with given clinical characteristics. Our tool supersedes the predictive ability of current scales and is easily applicable in daily clinical practice. The tool is freely available at https://biostat.umed.pl/deepNMIBC/.

Author contributions: Konrad Stawiski had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Stawiski, Jobczyk.
Acquisition of data: Jobczyk, Stawiski, Soria, Shariat.
Analysis and interpretation of data: Stawiski, Kaszkowiak.
Drafting of the manuscript: Stawiski, Jobczyk.
Critical revision of the manuscript for important intellectual content: Fendler, Shariat, Rajwa, Soria, Różański.
Statistical analysis: Stawiski, Kaszkowiak.
Obtaining funding: Stawiski.
Administrative, technical, or material support: Fendler.
Supervision: Fendler.
Other: None.

Financial disclosures: Konrad Stawiski certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: The study was partly funded by the National Science Centre in Poland (research grant no. 2018/29/N/NZ5/02422 to K.S.) and the Polpharma Scientific Foundation. The sponsors played a role in manuscript preparation.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.euo.2021.05.006.

References